Probabilistische Netze

Patrick Rammelt

Inhaltsverzeichnis

1 Einleitung

In vielen Bereichen können Computermodelle bei der Prädiktion der Auwirkungen bestimmter Faktoren, bzw bei der Diagnose der Ursachen eingetretener Zustände - seien sie gut oder schlecht - helfen. Dabei besteht eine Art Dreiecksbeziehung zwischen dem was beobachtbar ist, dem was unbekannt und aufgrund der beobachteten Grössen zu ermitteln ist und Einflussgrössen die in dieses Verhältnis hineinspielen die aber z.T. nicht einmal klar zu benennen sind und deren Einfluss als zufällige Schwankung ("Rauschen") erscheint. Könnte man die letzte Gruppe beliebig zugunsten der ersten verkleinern, so müssten sich potentiell bessere, ja (fast[1]) perfekte Vorhersagen machen lassen. Jedoch ergibt sich dadurch auch ein entsprechend komplexes System, in dem ja nicht nur die Einflussgrössen als solche zu benennen sind sondern auch ihre genauen Wechselwirkungen untereinader beschrieben werden müssen. Es ist in der Praxis weder möglich alle Einflussfaktoren und deren Wechselwirkungen zu kennen, noch wäre ein solches System von seiner Komplexität her handhabbar. Daher werden Methoden benötigt, um mit den resultierenden Unsicherheiten vernünftig umgehen zu können. Eine Möglichkeit hierzu bieten Probabilistische Methoden, die solche Unsicherheiten in der Form von (bayes'schen) Wahrscheinlichkeiten modellieren. Dieses soll angefangen bei einem einfachen Beispiel erläutert werden. Schritt für Schritt wird dabei eine Methode herausgearbeitet die für die speziellen Anforderungen dieser Arbeit als besonders geeignet erscheint.

2 Ein einfaches Modell

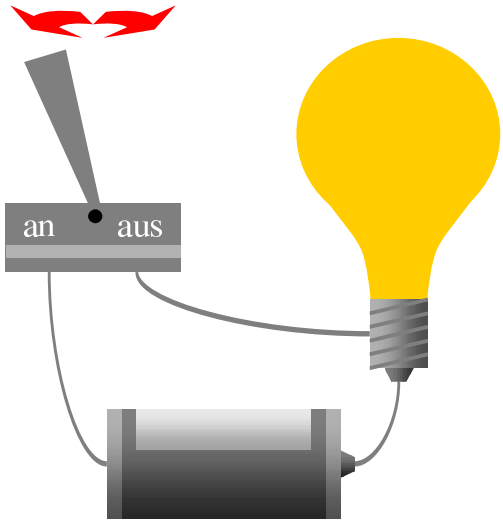


Abbildung 1: *Schalter-Lampe-System*

Erkenntnisse über die Welt oder einzelne Teilbereiche (Teilsysteme) resultieren oft aus der Beobachtung dieser Systeme. Beobachtet werden dabei einzelne *Ereignisse*, die einer Menge (theoretisch) möglicher Ereignisse, dem Ereignisraum Ω entspringen.

[1] Zumindest in der Quanten-Theorie ist der Zufall grundsätzlich nie ganz zu beseitigen

Ein einfaches Beispiel ist ein System bestehend aus einem Schalter, einer Stromquelle, einer Lampe und den verbindenden Kabeln (siehe Abbildung 2). Die zu beobachtenden Ereignisse können mithilfe bestimmter *Merkmale* erfasst werden. So kann im vorliegenden Beispiel z.B. beobachtet werden:

- die Stellung des Schalters
- ob die Lampe leuchtet oder nicht leuchtet

Damit wird bereits eine Auswahl der (vermeintlich) relevanten Merkmale getroffen, denn genausogut hätten auch noch weitere oder ganz andere Merkmale z.B. bezüglich der Stromquelle oder der Kabel erhoben werden können. Ein Merkmal teilt die Ereignisse in bestimmte Zustände aus einer merkmalsspezifischen Zustandsmenge ein. Das Merkmal Schalterstellung (kurz: Schalter) z.B. hat die Zustandsmenge {"an","aus"}, das Merkmal Lampe besitzt die Zustandsmenge {"leuchtet","leuchtet nicht"}. Diese Einteilung der Zustände erscheint zwar recht naheliegend, dennoch ist auch sie im allgemeinen Ergebnis einer Festlegung, die auch anders hätte vorgenommen werden können. Zweckmäßiger Weise sind die Zustände eines Merkmals so gewählt, dass sie *umfassend* sind, d.h. das es für jedes Ereignis auch einen Zustand aus der Zustandsmenge dieses Merkmals gibt. Zudem sollten die Zustände eines Merkmals *disjunkt* gewählt werden. Ist die Zustandsmenge eines Merkmals sowohl umfassend als auch disjunkt, so tritt zu einem bestimmten Zeitpunkt immer genau ein Zustand ein. Die beiden hier genannten Zustandsmengen sind ausserdem *diskret*, d.h. es werden bestimmte *Kategorien* wie "an", "aus" / "leuchtet", "leuchet nicht" gebildet. Es sind aber auch Merkmale mit kontinierlichen Zustandsmengen (kurz *kontinuierliche Merkmale*) denkbar. So hätte z.B. ein Merkmal Helligkeit gemessen in Lux erhoben werden können, das Zustände aus der Menge der positiven reellen Zahlen ($\mathbb{R}^+$) annehmen kann. Durch eine *Diskretisierung* kann ein solches Merkmal in ein *diskretes Merkmal* überführt werden. So hätte das Merkmal Lampe aus dem Merkmal Helligkeit durch Festlegung einer Helligkeitsgrenze zur Differenzierung der Zustände "leuchet" und "leuchtet nicht" gewonnen weden können. Im Folgenden werden vorerst nur diskrete Merkmale behandelt.

Bezüglich unterschiedlicher Merkmale kann theoretische jede Kombination von Zuständen eintreten. Das wären im hier angeführten Beispiel:

- Schalter an **und** Lampe leuchtet
- Schalter aus **und** Lampe leuchtet nicht
- Schalter an **und** Lampe leuchtet nicht (Fehlerfall, z.B. Lampe, Schalter, Stromquelle oder Kabel defekt)
- Schalter aus **und** Lampe leuchtet (Fehlerfall, z.B. durch Kurzschluss im Schalter)

Eine solche Kombination von Zuständen mehrerer Merkmale wird im weiteren als *Zustandskonfiguration* bezeichnet. Ohne die beiden Fehlerfälle bestünde ein *deterministischer* Zusammenhang zwischen den beiden Merkmalen Schalter und Lampe: Immer wenn der Schalter auf "an" steht leuchtet die Lampe, dagegen leuchtet sie nie, wenn er auf "aus" steht. Wenn die genannten Fehlerquellen die einzigen Gründe für eine Abweichung von dieser "Regel" darstellen, könnte prinzipiell ein deterministischer Zusammenhang angenommen werden, wenn weitere Merkmale zur Beschreibung der Funktionstüchtigkeit aller Komponenten gebildet werden: Die Lampe leuchtet wenn der Schalter auf "an" steht und alle Komponenten funktionieren oder der Schalter auf "aus" steht, ein Kurzschluss im Schalter vorliegt und alle anderen Komponenten funktionieren. Das Problem hierbei ist hauptsächlich die Beobachtbarkeit der Merkmale, die sich auf die Funktionstüchtigkeit der Komponenten beziehen.

Bei praktisch allen "Real-Life-Problemen" ist es unmöglich alle Einflussfaktoren (Merkmale) zu beobachten oder auch nur prinzipiell zu wissen welche Einflussfaktoren überhaupt eine Rolle spielen. Das gilt natürlich insbesondere wenn menschliches Verhalten das System beeinflusst.

Damit bleibt eine Unsicherheit bezüglich des Zusammenhangs von Merkmalen. Wenn der Schalter z.B. auf "an" steht leuchtet die Lampe meistens, machmal jedoch auch nicht. Um "meistens" und "manchmal" genauer zu spezifizieren bietet sich die Verwendung von Wahrscheinlichkeiten an.

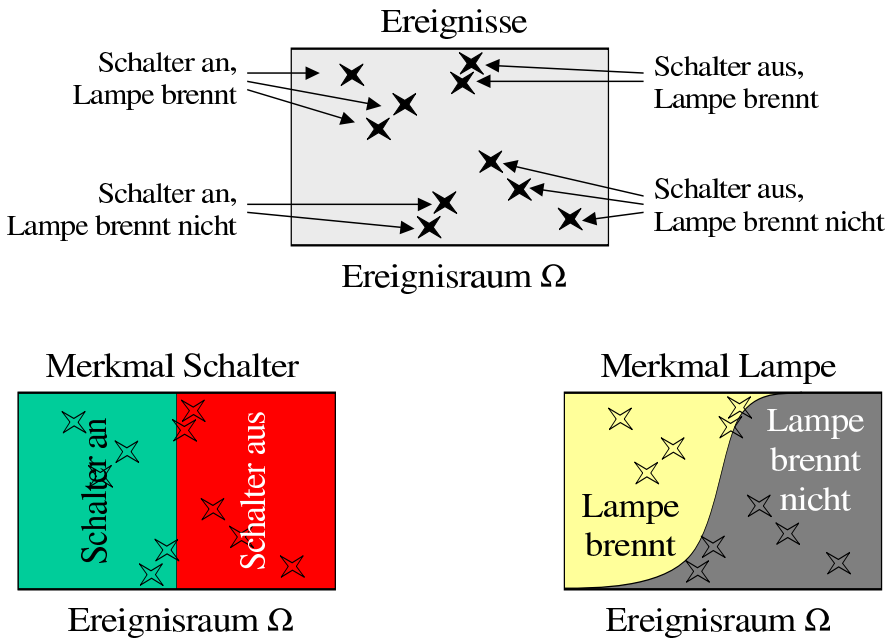


Abbildung 2: *Merkmale teilen den Ereignsiraum auf*

Die Wahrscheinlichkeitsrechnung basiert auf den *Axiomen von Kolmogorov*:

Axiom 1 *Die Wahrscheinlichkeitsfunktion $P(.)$ ordnet jedem Ereignis a des Ereignisraums Ω eine nicht reelle Zahl zwischen 0 und 1 zu*

$$0 \leq P(a) \leq 1$$

Axiom 2 *Der Ereignisraum Ω umfasst* **alle** *möglichen Ereignisse. Es gilt:*

$$P(\Omega) \quad = \quad 1$$

Axiom 3 *Für disjunkte Ereignisse a und b ($a, b \subseteq \Omega$) gilt:*

$$P(a \vee b) \quad = \quad P(a) + P(b)$$

Aus den Axiomen folgt u.a., dass die Summe der Wahrscheinlichkeiten der einzelnen Zustände eines Merkmals "1" ergibt. Z.B. sind $P(Schalter = an) + P(Schalter = aus) = 1$ und $P(Lampe = leuchtet) + P(Lampe = leuchtet\ nicht) = 1$. Eine recht intuitive Veranschaulichung des Konzepts (diskreter) Merkmale und der beschreibenden Wahrscheinlichkeiten bietet die Übersetzung

von Wahrscheinlichkeiten in Flächen innerhalb eines zweidimensional dargestellten Ereignisraums Ω. Dieses ist in Abbildung 2 gezeigt. Der gesamte Ereignisraum Ω hat den Flächeninhalt 1 (entsprechend $P(\Omega) = 1$). Die Wahrscheinlichkeiten für die vier Zustandskombinationen der Merkmale sind der Abbildung zu entnehmen, ebenso wie die Wahrscheinlichkeiten für die nur nach einem Merkmal unterschiedenen Ereignisse, die am Rand aus der Addition von jeweils zwei Wahrscheinlichkeiten (Flächen) entstanden sind. Z.B. ist die Wahrscheinlichkeit, dass die Lampe brennt und der Schalter auf an steht $P(Schalter = an, Lampe = leuchtet) = 0.49$ und die Wahrscheinlichkeit, dass die Lampe leuchtet (unabhängig von der Schalterstellung) liegt bei $P(Lampe = leuchtet) = 0.4905$ $(= 0.49 + 0.0005)$.

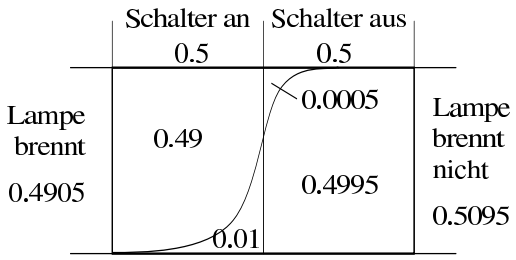


Abbildung 3: *Wahrscheinlichkeiten als Flächen*

Man kann sich die Entstehung eines Ereignisses, also die Festlegung des konkreten Zustands so vorstellen, dass jemand einen Pfeil auf die Ω-Fläche wirft (Einige "Treffer"/Ereignisse sind in Abbildung 2 zu sehen). Jeder Punkt der Fläche hat dabei die gleiche Chance getroffen zu werden[2] - der Pfeil trifft jedoch mit Sicherheit genau einen Punkt in Ω. Nun entspricht die Wahrscheinlichkeit eines bestimmten Bereichs getroffen zu werden seiner Fläche (Fläche / Gesamtfläche). Angenommen es soll eine Wette darauf abgeschlossen werden dass das nächste zu beobachtende Ereignis Schalter="an" und (aber) Lampe leuchtet nicht ist. In Kenntnis der wahren Wahrscheinlichkeit $0.01 = 1\%$ für dieses Ereignis sollte man die Wette nur abschliessen, wenn der Gewinn mehr als das 100-fache des Einsatzes beträgt. Genau durch dieses Beispiel ist der Nutzen von Wahrscheinlichkeitswerten in der Vorhersage von Ereignissen begründet. Der hier verwendete Bayes'sche Wahrscheinlichkeitsbegriff erlaubt genau diese Interpretation. In der klassischen frequentistischen Sicht hat ein einzelnes Ereignis dagegen keine Wahrscheinlichkeit. Wahrscheinlichkeiten beziehen sich hier immer auf Auftretenshäufigkeiten (engl. "frequencies") in einer Grundgesamtheit. Obwohl hier der Bayes'sche Wahrscheinlichkeitsbegriff Anwendung findet soll in diesem Sinne im Folgenden als "wahre" Wahrscheinlichkeit die relative Auftretenshäufigkeit von bestimmten Zuständen bei unendlich vielen Versuchen (Ereignissen) verstanden werden. Da unendlich viele Ereignisse nicht beobachtet werden können sind die wahren Wahrscheinlichkeiten nie genau bekannt und können nur geschätzt werden. Dazu gibt es u.a. die Möglichkeit einer subjektiven Schätzung aufgrund von Erfahrung, Wissen z.B. über Naturgesetze etc.. Auch die Formulierung eines rein subjetiven "Glaubens" als Wahrscheinlichkeit ist in der bayes'schen Wahrscheinlichkeitslehre gestattet. Eine weitere Möglichkeit Wahrscheinlichkeiten zu schätzen besteht aber natürlich in der Auswertung von Daten, also einer Sammlung von Beobachtungen. Hierbei gehen die klassische frequentistische und bayes'sche Wahrscheinlichkeitslehre konform. Zudem verdeutlicht sich

[2]ein Paradoxon der Wahrscheinlichkeitsrechnung mit kontinuierlichen Größen besteht darin, dass die Wahrscheinlichkeit für einen einzelnen "Punkt" eigentlich Null ist. Genauer wäre daher hier: Gleich große Flächen haben die gleiche Chanche getroffen zu werden

durch den Bezug auf Daten die innere Konsistenz der Berechnungen besonders gut.

Zur Vereinfachung der Schreibweise sei im Weiteren:

$$\begin{array}{rcll} \text{Lampe} & = & \text{leuchtet} & \equiv \; l_1 \\ \text{Lampe} & = & \text{leuchtet nicht} & \equiv \; l_2 \\ \text{Schalter} & = & \text{an} & \equiv \; s_1 \\ \text{Schalter} & = & \text{aus} & \equiv \; s_2 \end{array}$$

Desweiteren wird z.B. die Schreibweise Lampe=leuchtet bzw. Lampe=l_1 mit l_1 abgekürzt. Die Merkmale Lampe und Schlater selbst werden mit L bzw. S bezeichnet. L und S stehen also als Platzhalter für die möglichen Zustände l_1 und l_2 bzw. s_1 und s_2.

Realistischer als die eben beschriebene Wette währe wohl z.B. eine Wette darauf, dass die Lampe nicht angeht, **wenn** das nächste Mal mithilfe des Schalters von "aus" auf "an" geschaltet wird. Die Schalterstellung "an" wird also zur Vorraussetzung gemacht. Diese Wahrscheinlichkeit sei mit $P(l_2|s_1)$ bezeichnet - es ist die *bedingte Wahrscheinlichkeit* für Lampe "leuchtet nicht" *gegeben* der Schalter steht auf "an". Der Mögliche Ereignisraum wird sozusagen auf den Teilbereich verengt, der mit der Vorraussetzung (Schalter = "an") übereinstimmt. Damit gilt:

$$P(l_2|s_1) \quad = \quad \frac{P(l_2, s_1)}{P(s_1)} = \frac{0.01}{0.5} = 0.02$$

Definition 1 *Für alle Wahrscheinlichkeitsrerechnungen gilt:* $0/0 := 0$

Definition 2 $P(b) = 0 \Rightarrow P(a|b) = 0$

Satz 1 Satz von der bedingten Wahrscheinlichkeit:

$$P(a, b) = P(a|b) \cdot P(b) \tag{1}$$

Mit den Definitionen 1 und 2 folgt auch:

$$P(a|b) = \frac{P(a, b)}{P(b)} \tag{2}$$

Angenommen es wurden 2000 Ereignisse bezüglich der Merkmale Lampe und Schalter beobachtet. Die Anzahlen der verschiedenen Zustände / Zustandskombinationen dieses *Datensatzes* $D(L, S)$ sei mit folgende Tabelle gegeben:

D(L,S)	Lampe= leuchtet (l_1)	Lampe= leuchtet nicht (l_2)	$\sum$
Schalter=an (s_1)	980	20	1000
Schalter=aus (s_2)	1	999	1000
$\sum$	981	1019	2000

Tabelle 2: *Beobachtungshäufigkeiten (Datensatz $D(L, S)$)*

Der Einfachheit halber sind diese Daten so gewählt, dass sie die hier vorgegebenen wahren Wahrscheinlichkeiten perfekt wiederspiegeln (jeder Wert durch die Gesamtanzahl 2000 geteilt ergibt wieder die Werte aus Abbildung 2):

$P(L,S)$	l_1	l_2	$\sum_L$
s_1	$980/2000 = 0.49$	$20/2000 = 0.01$	0.5
s_2	$1/2000 = 0.0005$	$999/2000 = 0.4995$	0.5
$\sum_S$	$981/2000 = 0.4905$	$1019/2000 = 0.5095$	1

Tabelle 2: *Verbundwahrscheinlichkeitstafel* $P(L,S)$

Es sei jedoch angemerkt, dass es höchst unwahrscheinlich ist bei 2000 Versuchen und gegebenen wahren Wahrscheinlichkeiten aus Abbildung 2 genau diese Anzahlen zu erhalten[3]. Genau darin besteht die Schwierigkeit der Schätzung der wahren Wahrscheinlichkeiten aus Daten.

Aus der Tabelle 2 lassen sich auch die *bedingten Wahrscheinlichkeiten* ableiten. Zunächst für Lampe bei gegebener Schalterstellung:

$P(L\|S)$	l_1	l_2	$\sum_L$
s_1	$980/1000 = 0.98$	$20/1000 = 0.02$	1
s_2	$1/1000 = 0.001$	$999/1000 = 0.999$	1

Tabelle 2: *bedingte Wahrscheinlichkeitstafel* $P(L|S)$

Und entsprechend für Schalter bei gegebenem Lampenzustand:

$P(S\|L)$	l_1	l_2	
s_1	$980/981 \approx 0.99898$	$20/1019 \approx 0.01963$	
s_2	$1/981 \approx 0.00102$	$999/1019 \approx 0.98037$	
$\sum_S$	1	1	

Tabelle 2: *bedingte Wahrscheinlichkeitstafel* $P(S|L)$

Auch für die bedingten Wahrscheinlichkeiten gilt, dass sich die Wahrscheinlichkeiten für die Zustände eines Merkmals unter gleichen Vorbedingungen (gegebene Merkmale) zu eins addieren.

Die Tabellen 2, 2 und 2 zeigen jeweils auch *Randwahrscheinlichkeiten*, die durch das Aufsummieren (*Marginalisieren*) der Werte für die Zustände eines Merkmals entstehen. So gilt z.B. $P(L) = \sum_S P(L,S)$, $P(S) = \sum_L P(L,S)$ und $\sum_{L,S} P(L,S) = P(\Omega) = 1$. Dabei sind $P(L)$ und $P(S)$ die sogenannten *A-Priori-Wahrscheinlichkeiten* ("unbedingten" Wahrscheinlichkeiten) der Merkmale L

[3]Man stelle sich vor, dass noch eine weitere Beobachtung hinzukommt, dann ist es hier schon aus numerischen Gründen unmöglich die wahren Wahrscheinlichkeiten perfekt abzubilden

und S.

Definition 3 Marginalisierung *über das Merkmal B*

$$P(A) = \sum_B P(A,B) \quad \equiv \quad P(a_i) = \sum_j P(a_i, b_j), \qquad \forall_i$$

Satz 2 Marginalisierungsregel:

$$P(A) \quad = \quad \sum_B P(A,B)$$

Satz 3 Mehrfach-Marginalisierungsregel:

$$P(A) \quad = \quad \sum_C \sum_B P(A,B,C) = \sum_B \sum_C P(A,B,C) = \sum_{B,C} P(A,B,C)$$

Die Tafel $P(L,S)$ beinhaltet alle Informationen über den Datensatz $D(L,S)$ bis auf die Anzahl der Beobachtungen N (der Datensatz $D(L,S)$ kann aus $P(L,S)$ erzeugt werden indem alle Werte mit $N = 2000$ multipliziert werden). Die Tafeln $P(L|S)$, $P(S|L)$, $P(L)$ und $P(S)$ beinhalten jeweils nur einen Teil der Informationen, die nötig wären um $D(S|L)$ zu rekonstruieren. Nach Satz 1 gilt aber $P(L,S) = P(L|S) \bullet P(S)$ und $P(L,S) = P(S|L) \bullet P(L)$.

Die Multiplkation von Tafeln "$\bullet$" ist wie folgt definiert:

Definition 4 Multiplikation von Wahrscheinlichkeits-Tafeln:

$$P(A,B) = P(A|B) \bullet P(B) \quad \equiv \quad P(a_i, b_j) = P(a_i|b_j)P(b_j), \qquad \forall_{i,j}$$

Entsprechend gilt für die die Division:

Definition 5 Division von Wahrscheinlichkeits-Tafeln:

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \equiv \quad P(a_i|b_j) = \frac{P(a_i,b_j)}{P(b_j)}, \qquad \forall_{i,j}, \qquad \frac{0}{0} := 0$$

Insbesondere die Rekonstruktion von $P(L,S)$ aus den Tafeln $P(S)$ und $P(L|S)$ kommt der menschlichen Intuition entgegen, da die Schalterstellung als ursächlich (kausal) für den Lampenzustand angesehen werden kann. So enthält $P(L|S)$ "Regeln" der Form: "Wenn der Schalter sich in Position s_i befindet, dann ist die Lampe mit einer Wahrscheinlichkeit von $P(l_j|s_i)$ im Zustand l_j". Problematischer ist allerdings oft die Angabe der A-Priori-Wahrscheinlichkeiten (hier für für $P(S)$). Auch kann nicht in allen Fällen eine Kausalität in der Abhängigkeit zwischen Merkmalen vorausgesetzt werden. Diese Probleme sollen zunächst aber noch vernachlässigt werden.

Damit ist mit $P(S)$ und $P(L|S)$ ein Wahrscheinlichkeitsmodell für das "Schalter-Lampe-System" aufgestellt. Es gibt verschiedene Informationen die dieses Modell zu liefern vermag: Neben den A-Priori-Wahrscheinlichkeiten $P(S)$ (direkt gegeben) und $P(L)$ $(= \sum_S P(S, L))$ ist es von besonderem Interesse die *A-Posteriori-Wahrscheinlichkeitsverteilung* eines Merkmals zu ermitteln, also die Wahrscheinlichkeitsverteilung die sich ergibt nachdem Informationen (Beobachtungen) über die Zustände vom anderen Merkmalen vorliegen. Solche Informationen werden im Folgenden als *Evidenz* bezeichnet; die Berechnung der A-Posteriori-Wahrscheinlichkeitsverteilungen als *Inferenz*. Die Wahrscheinlichkeiten dafür, dass die Lampe leuchtet bzw. nicht leuchtet wenn der Schalter auf "an" bzw. "aus" steht ist ja mit $P(L|S)$ direkt gegeben. Nicht sofort ablesen kann man dagegen z.B. die Wahrscheinlichkeit dass der Schalter wirklich auf "aus" steht wenn nur bekannt ist, dass die Lampe nicht leuchtet $(P(s_2|l_2))$. Aus den beobachteten Daten ist ersichtlich dass diese Wahrscheinlichkeit bei $999/(999 + 20) \approx 0.9804$ liegt. Es fehlt aber noch eine Berechnungsvorschrift für $P(S|L)$ aus $P(S)$ und $P(L|S)$.

Satz 4 *Benutzt man die Sätze 1 und 2 so erhält man die* **Bayes'sche Formel***:*

$$P(S|L) = \frac{P(L|S)P(S)}{P(L)}, \tag{3}$$

wobei gilt:

$$P(L) = \sum_S P(L, S) = \sum_S P(L|S) \bullet P(S). \tag{4}$$

Ein besonders Intuitive veranschaulichung der Abhängigkeits-Beziehungen zwischen Merkmalen bietet die Darstellung als Graph. Dabei bilden die Merkmale die Knoten, während eine kausale Abhängigkeit durch eine gerichtete Kante von der Ursache auf die Wirkung dargestellt wird. Der in Abbildung 2 dargestellte Graph zusammen mit den Wahrscheinlichkeitstafeln $P(S)$ und $P(L|S)$ bilden bereits ein sehr einfaches *Probabilistisches Netz* (PN).

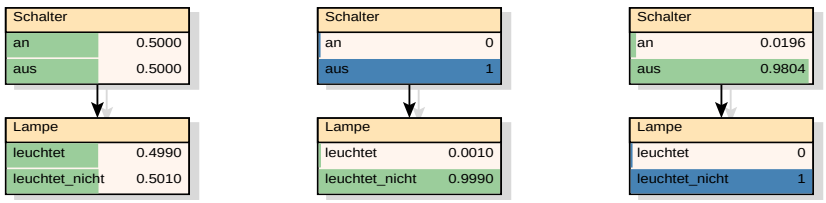


Abbildung 4: *Schalter-Lampe-PN; links: A-Priori-Wahrscheinlichkeiten, mitte: A-Posteriori-Wahrscheinlichkeiten für Schalter=aus, rechts: A-Posteriori-Wahrscheinlichkeiten für Lampe=leuchtet nicht*

3 Probabilistische Netze

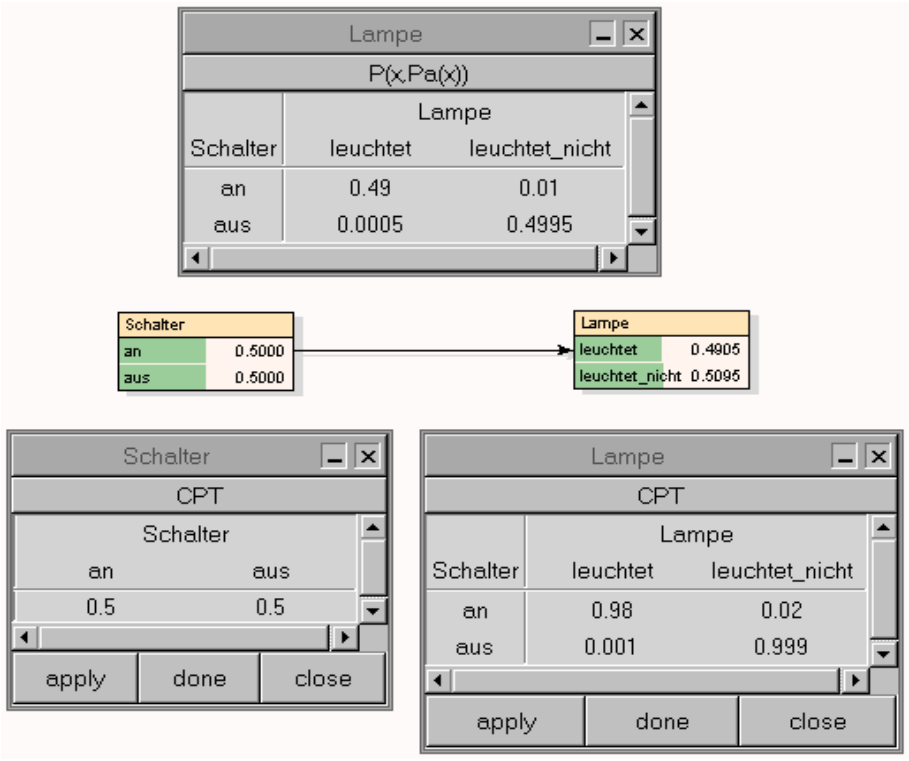


Abbildung 5: *Schalter-Lampe-PN; CPT's für die Knoten Schalter und Lampe (unten), sowie die Verbundwahrscheinlichkeitstafel beider Knoten (oben)*

Der am einfachen "Schalter-Lampe"-Beispiel veranschaulichte Ansatz lässt sich nun auch auf komplexere Zusammenhänge mit mehr Merkmalen ausweiten. Hierzu müssen aber einige Begriffe allge-

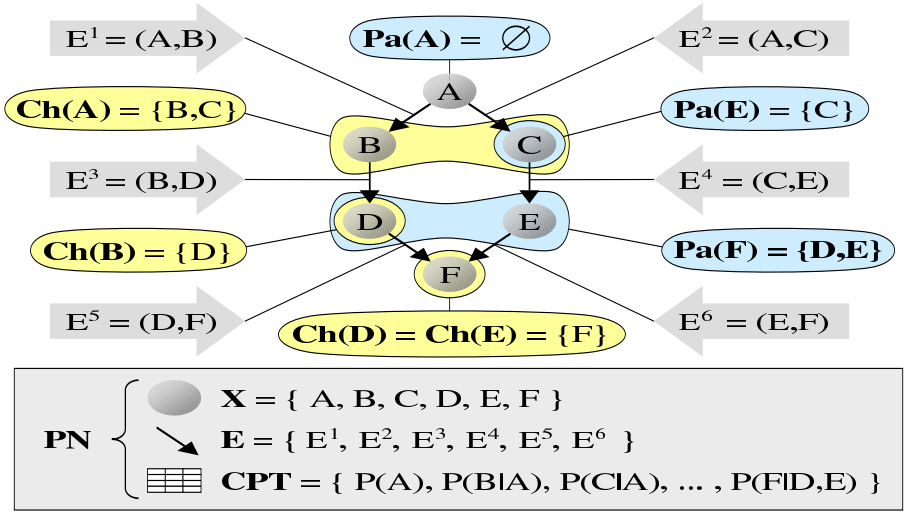


Abbildung 6: *Ein PN ist definiert durch seine Knoten, gerichteten Kanten und CPT's*

meingültiger und formaler definiert werden als das im letzten Abschnitt der Fall war.

Definition 6 *Ein PN beinhaltet eine Anzahl von* $N > 0$ **Merkmalen**. *Die Gesamtheit aller Merkmale sei mit* $\mathbf{X}$ *(Fettdruck für Mengen) bezeichnet. Die einzelnen Merkmale werden mit* $X^1, \ldots, X^N$ *bezeichnet. Vereinfachend werden aber auch weiterhin (dünngedruckte) Grossbuchstaben wie A, B, C, ... zur Bezeichnung einzelner Merkmale in* $\mathbf{X}$ *verwendet.*

$$\mathbf{X} = \{X^1, X^2, \ldots, X^N\} \tag{5}$$

Die Merkmale werden als **Knoten** *in einem Graphen dargestellt. Die Begriffe Merkmal und Knoten werden im Folgenden synonym verwendet.*

Definition 7 *Der* k*-te Zustand des* i*-ten Merkmals* X^i *sei in Fortführung der bisherigen Notation mit* x^i_k *bezeichnet.*

Definition 8 *Eine* **direkte kausale Abhängigkeit** *zwischen zwei Merkmalen* U *und* W *wird durch eine* **gerichtete Kante**, *die von der Ursache* U *auf die Wirkung* W *zeigt dargestellt.*

Definition 9 *Ein* **Pfad** *von A nach B ist eine Serie von Kanten, die von Knoten A (über Zwischenknoten) zu Knoten B führt. Kantenrichtungen bleiben dabei unberücksichtigt.*

Definition 10 *Ein* **gerichteter Pfad** *ist ein Pfad bei dem nur gerichtete Kanten entsprechend ihrer Richtungen "beschritten" werden.*

Definition 11 *Die* **Länge eines (gerichteten) Pfads** *ist die Anzahl der Knoten die besucht werden.*

Definition 12 *Ein Knoten A heisst* **Elternknoten** *von Knoten B, wenn es eine gerichtete Kante von A nach B gibt. Die* **Menge alle Elternknoten** *eines Knotens B wird mit* $\mathbf{Pa}(\mathbf{B})$ *bezeichnet. Umgekehrt heisst B* **Kindknoten** *von A. Die* **Menge alle Kindknoten** *eines Knotens A wird mit* $\mathbf{Ch}(\mathbf{A})$ *bezeichnet.*

Definition 13 *Ein Knoten A heisst* **Vorgänger** *von Knoten B, wenn es einen gerichteten Pfad von A nach B gibt. Die* **Menge aller Vorgänger** *eines Knotens B wird mit* $\mathbf{An}(\mathbf{B})$ *bezeichnet. Umgekehrt heisst B* **Nachfolger** *von A. Die* **Menge aller Nachfolger** *eines Knotens A wird mit* $\mathbf{De}(\mathbf{A})$ *bezeichnet.*

Definition 14 *Ein* **(gerichteter) Zyklus** *ist ein (gerichteter) Pfad der Länge* $L > 2$ *von Knoten A zurück zum Knoten A, wobei jeder Knoten des Zyklus' genau einmal besucht wird - bis beim Schließen des Zyklus' A zum zweiten Mal erreicht wird (Für die* **Länge eines zyklischen Pfads** *zählt der Anfangs-/Endknoten nur einmal).*

Definition 15 *Ein* **Probabilistisches Netz** *wird definiert durch: 1. einen gerichteten azyklischen Graphen (DAG) G (dieser darf Zyklen jedoch keine gerichteten Zyklen aufweisen) und 2. den dazu passenden bedingten Wahrscheinlichkeitstafeln* $P(A|\mathbf{P}(\mathbf{A}))$ *für jeden Knoten A gegeben seine Eltern* $\mathbf{Pa}(\mathbf{A})$ *in G. Im Falle eines elternlosen Knotens (*$\mathbf{Pa}(\mathbf{A}) = \emptyset$*) ist die A-Priori-Wahrscheinlichkeitsverteilung* $P(A)$ *anzugeben. Vereinheitlichend werden aber beide Typen als CPT's ("Conditional Probability Tables") bezeichnet.*

4 Abhängigkeitsstruktur

4.1 Realer / virtueller Datensatz

Im Abschnitt 2 wurden die CPT's aufgrund eines Datensatzes mit 2000 Beobachtungen empirisch bestimmt. Das einfache dort vorgestellte Schalter-Lampe-Modell erlaubt es die Beobachtungen aus dem PN rückzurechnen, indem die Verbundwahrscheinlichkeitstafel für alle Knoten mit der Beobachtungsanzahl $N = 2000$ multipliziert wird.

$$\begin{aligned} & 2000 \bullet P(L|S) \bullet P(S) \\ = \; & 2000 \bullet P(L, S) \end{aligned}$$

$$= \quad 2000 \bullet \begin{array}{c|cc} & l_1 & l_2 \\ \hline s_1 & 0.49 & 0.01 \\ s_2 & 0.0005 & 0.4995 \end{array} = \begin{array}{c|cc} & l_1 & l_2 \\ \hline s_1 & 980 & 20 \\ s_2 & 1 & 999 \end{array}$$

Satz 5 *Die Verbundwahrscheinlichkeitstafel aller Knoten kann über die* **Kettenregel** *gewonnen werden:*

$$P(\mathbf{X}) \quad = \quad \prod_i P(X^i | \mathbf{Pa}(\mathbf{X^i}))$$

Lemma 1 *Durch Vorgabe einer Beobachtungsanzahl N kann daraus ein* **virtueller Datensatz** *D^v mit N Beobachtungen erzeugt werden.*

$$D^v = P(\mathbf{X}) \bullet N$$

Die Tatsache, dass sich hier abhängig von der Gesamtanzahl der virtuellen Beobachtungen N eventuell nicht ganzzahlige "Beobachtungsanzahlen" für einzelne Zustandskombinationen ergeben ist für die weitere Betrachtung unerheblich. Inwieweit D^v mit einem gleichgroßen Datensatz D übereinstimmt aus dem die bedingten Wahrscheinlichkeiten empirisch gewonnen wurden hängt davon ab wie gut es dem PN gelingt die in D vorhandenen Abhängigkeiten abzubilden. Welche Abhängigkeitsbeziehungen darstellbar sind wird durch die Netzstruktur bestimmt, von ihr hängt bei fester Merkmalsmenge die Komplexität des Modells ab.

Würde man aus dem Schalter-Lampe-Modell die (einzige) Kante Schalter → Lampe entfernen so würde sich bei gleicher Datenlage die CPT des Knotens Lampe zu $P'(L) = (0.4905, 0.5095)$ verändern. Dann aber würde sich auch eine veränderte Verbundwahrscheinlichkeit $P'(L, S)$ ergeben (nach der Kettenregel aus Satz 5).

$$\begin{aligned} P'(L,S) &= P'(L) \bullet P(S) \\ &= \begin{array}{cc} l_1 & l_2 \\ \hline 0.4905 & 0.5095 \end{array} \bullet \begin{array}{c|c} s_1 & 0.5 \\ s_2 & 0.5 \end{array} \\ &= \begin{array}{c|cc} & l_1 & l_2 \\ \hline s_1 & 0.24525 & 0.25475 \\ s_2 & 0.24525 & 0.25475 \end{array} \end{aligned} \tag{6}$$

Ganz offensichtlich ist $P(L, S) \neq P'(L, S)$ d.h. dieses vereinfachte Modell ist nicht mehr in der Lage die Daten abzubilden. Das liegt natürlich daran, dass in diesem Modell L und S *unabhängig* voneinander sind. Diese Unabhängigkeit ist dadurch gekennzeichnet, dass sich die Verbundwahrscheinlichkeit $P'(L, S)$ durch die Multiplikation der Einzelwahrscheinlichkeiten $P'(L)$ und $P(S)$ ergibt, während

im abhängigen Fall $P(L,S) \neq P(L) \bullet P(S)$ gilt.

Definition 16 *Zwei Merkmale A und B werden genau dann* **unabhängig** *genannt, wenn gilt:*

$$P(A,B) \quad = \quad P(A) \bullet P(B)$$

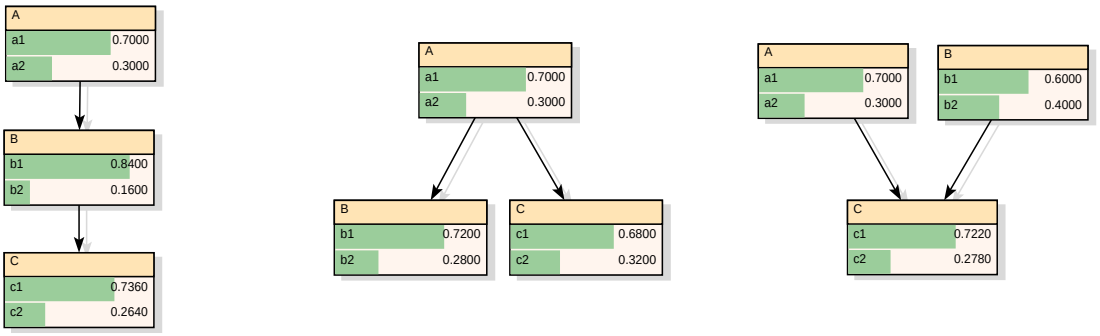


Abbildung 7: *Verbindungstypen: links: seriell, mitte: divergierend, rechts: konvergierend*

In einem komplexeren PN gibt es *bedingte Unabhängigkeiten* und *bedingte Abhängigkeiten*. Um dies zu verdeutlichen sollen nun die verschiedenen basalen Verbindungstypen untersucht werden.

4.2 Serielle Verbindung

In einer **seriellen Verbindung** von drei Knoten Ⓐ → Ⓑ → Ⓒ (siehe Abbildung 4.1 links) sind A und C abhängig, wenn der Zwischenknoten B unbeobachtet ist, werden jedoch unabhängig bei gegebenem B.

$$\begin{aligned} P(A,C) \quad &\neq \quad P(A) \bullet P(C) \\ P(A,C|B) \quad &= \quad P(A|B) \bullet P(C|B) \end{aligned}$$

Alle Tafeln entstehen durch Marginalisierung aus der mittels Kettenregel (5) gewonnenen Verbundwahrscheinlichkeit $P(A,B,C)$, sowie durch Division durch $P(B)$ (Satz von der bedingten Wahrscheinlichkeit 1) im Falle der bedingten Tafeln.

$$P(A,C|B) = P(A|B) \bullet P(C|B) \tag{7}$$

Der *Informationsfluss* von A nach C wird **blockiert**, wenn B gegeben ist. Dieser Sachverhalt stimmt auch mit der Interpretation kausaler Abhängigkeiten überein, denn wenn die Ursache B eines Merkmals C direkt Bekannt ist, so bringt es keinen weiteren Aufschluss für C auch noch die Ursache der Ursache - also A - zu kennen. Ohne Kenntnis über B lässt sich aber von A auf B und dadurch wiederum auf C schließen.

4.3 Divergierende Verbindung

Auch in einer **divergierenden Verbindung** Ⓑ ← Ⓐ → Ⓒ (siehe Abbildung 4.1 mitte) sind die äusseren Knoten B und C abhängig wenn keine Information über A vorliegt und werden unabhängig wenn A gegeben ist. Auch dieser Zusammenhang entspricht der intuitiven Vorstellung, denn wenn die Ursache A einer Auswirkung B bekannt ist so bringt es keine zusätzliche Information bezüglich B auch noch weitere Auswirkungen (C) der Ursache A zu kennen. Wenn A aber unbekannt ist, dann kann aus einer Wirkung C die Ursache A diagnostiziert und aus dieser wiederum die andere Auswirkung B vorhergesagt werden.

4.4 Konvergierende Verbindung

Anders als in den beiden zuvor beschriebenen Verbindungstypen verhält sich der **konvergierende Typ** der Form: Ⓐ → Ⓒ ← Ⓑ (siehe Abbildung 4.1 rechts). Hier sind A und B zunächst unabhängig, werden aber abhängig wenn C gegeben ist. Es gilt daher:

$$P(A,B) = P(A) \bullet P(B) \tag{8}$$

$$P(A,B|C) \neq P(A|C) \bullet P(B|C) \tag{9}$$

Eine Ursache (A) erlaubt zunächst keinen Rückschluss auf die jeweils andere mögliche Ursache (B) einer unbekannten Wirkung C. Wenn aber die Auswirkung C bekannt ist und eine Ursache (z.B. A) ausgeschlossen werden kann, so steigt die Wahrscheinlichkeit anderer Ursachen (B). Umgekehrt fällt diese wenn bereits eine mögliche Ursache gefunden wurde. Dieser Effekt nennt sich "**Explaining Away**". Im Gegensatz zu anderen Verfahren ist diese Eigenschaft in PN's besonders gut abgebildet.

Während bei den anderen beiden Typen der Zustand des Zwischenknotens sicher bekannt sein muss um die Abhängigkeit über diesen Pfad zu blockieren, reicht es im Falle der konvergierenden Verbindung aus unsichere Information über den Zwischenknoten zu haben, um eine Abhängigkeit zu etablieren (**freizuschalten**). Solche unsichere Information bezüglich des Zwischenknotens kann z.B. aus der Beobachtung der direkten und indirekten Auswirkungen - also von Nachfolgern des Zwischenknotens - resultieren.

4.5 Bedingte Unabhängigkeiten

Im allgemeinen können Abhängigkeiten zwischen Knoten über Pfade bestehen, die mehr als drei Knoten umfassen. Damit eine Abhängigkeit zwischen zwei Knoten über einen bestimmten Pfad besteht müssen alle Knoten auf diesem Pfad paarweise abhängig sein.

Satz 6 *Wenn alle Elternknoten* $\mathbf{Pa}(\mathbf{A})$ *gegeben sind ist ein Knoten A unabhängig von allen sonstigen Nicht-Nachfolgern* $\overline{\mathbf{De}(\mathbf{A})} = \mathbf{X} \setminus (\mathbf{De}(\mathbf{A}) \cup \mathbf{Pa}(\mathbf{A}))$.

$$P(A|\mathbf{An}(\mathbf{A})) = P(A|\mathbf{Pa}(\mathbf{A})) \tag{10}$$

Dieses ist leicht zu erkennen, denn jeder Pfad zwischen einem Knoten aus $\overline{\mathbf{De}(\mathbf{A})}$ führt notwendigerweise über eine serielle oder divergierende Verbindung mit A als einem Endknoten und einem

Elternknoten **Pa**(**A**) als Zwischenknoten - welcher gegeben ist. Diese Erkenntnis hilft aber zur Veranschaulichung der Abhängigkeitsstrukturen in nicht trivialen Netzen nur bedingt weiter. Hilfreicher ist das Konzept der *d-separation* bzw. *d-connection*.

Definition 17 *Zwei Knoten A und B sind* **d-connected**, *wenn es (mindestens) einen Pfad zwischen A und B gibt, auf dem keine serielle oder divergierende Verbindung geblockt ist und alle konvergierenden Verbindungen freigeschaltet sind. Zwei Knoten sind* **d-separiert**, *wenn sie nicht d-connected sind.*

Satz 7 *Zwei d-separierte Knoten A und B sind unabhängig gegeben die aktuelle Evidenz.*

Anmerkung: Zwischen zwei Knoten A und B die *d-connected* sind, besteht im allgemeinen eine Abhängigkeit gegeben die aktuelle Evidenz, allerdings können die bedingten Wahrscheinlichkeitstafeln im PN so gewählt werden dass A und B dennoch unabhängig sind. In diesem Fall könnte allerdings dieselbe Gesamtverteilung $P(\mathbf{X})$ auch mit einem um eine oder mehrere Kanten reduzierten Netz repräsentiert werden.

5 Junction Tree

5.1 Finding-Vektor

Im allgemeinen besteht die Aufgabe eines PN's darin aus bekannten Merkmalen die *A-Posteriori-Wahrscheinlichkeitsverteilungen* der übrigen unbekannten Merkmale zu folgern. Im Schalter-Lampe-Modell könnte z.B. aus der Beobachtung, dass die Lampe nicht leuchtet auf die Schalterstellung - oder genauer die A-Posteriori-Wahrscheinlichkeiten der möglichen Schalterstellungen "an" und "aus" - geschlossen werden. In diesem sehr einfachen Fall reicht dafür die Anwendung der Bayes'schen Formel (Gleichung 3), mit deren Hilfe aus den gegebenen CPT's für $P(L|S)$ und $P(L)$ die Wahrscheinlichkeitstafel $P(S|L)$ (Tabelle 2) berechnet werden kann. Da nun beobachtet wurde, dass die Lampe nicht leuchtet ist nur $P(S|l_2)$ gesucht, was der rechten Spalte der Tafel 2 entspricht. Diese Einschränkung von $P(S|L)$ auf $P(S|l_2)$ kann vorgenommen werden indem man zunächst alle ungültigen Einträge in $P(S|L)$ auf "0" setzt und anschließend über L marginalisiert. Das Null-Setzen kann durch die Multiplikation mit einem Vektor erreicht werden, der für den beobachteten Zustand eine eine 1 und ansonsten Nullen enthält. Ein solcher die Evidenz repräsentierender Vektor heisst im weiteren *Finding* (F).

$$\sum_L (P(S|L) \bullet F(L)) \quad = \quad P(S|L = l_2)$$

$$\iff \quad \sum_L \left(\begin{array}{c|cc} & l_1 & l_2 \\ \hline s_1 & 0.999 & 0.0196 \\ s_2 & 0.001 & 0.9804 \end{array} \bullet \begin{array}{|cc} l_1 & l_2 \\ \hline 0 & 1 \end{array} \right) \quad = \quad \begin{array}{c|c} & \\ \hline s_1 & 0.0196 \\ s_2 & 0.9804 \end{array} \qquad (11)$$

Im Sinne der weiteren Verallgemeinerung ist es wünschenswert die Findings in der Formel für die

Kettenregel zu berücksichtigen.

Satz 8 *Die Verbundwahrscheinlichkeit aller Knoten* $\mathbf{X}$ *bei gegebener Zustandskonfiguration* $\mathbf{b}$ *der beobachteten Knoten* $\mathbf{B} \subseteq \mathbf{X}$ *ergibt sich durch die* **modifizierte Kettenregel***:*

$$P(\mathbf{X}|\mathbf{B}=\mathbf{b}) = \frac{\prod_{A\in\mathbf{X}} P(A|\mathbf{Pa}(\mathbf{X}))}{P(\mathbf{B})} \bullet \prod_{B\in\mathbf{B}} F(B) \tag{12}$$

Anmerkung: $\mathbf{B}=\mathbf{b}$ ist eine Abkürzung für $B^1 = b_i^1, B^2 = b_j^2, \ldots$.

Da nur genau eine Zustandskombination der beobachteten Merkmale $\mathbf{B}$ übrig bleibt, reduziert sich $1/P(\mathbf{B})$ auf einen skalaren Normierungsfaktor. Da $\sum_{\mathbf{X}} P(\mathbf{X}|\mathbf{B}=\mathbf{b}) = 1$ sein muss ergibt sich dieser Faktor auch auf anderem Wege:

$$P(\mathbf{X}|\mathbf{B}=\mathbf{b}) = \frac{Q(\mathbf{U}|\mathbf{B}=\mathbf{b})}{\sum Q(\mathbf{X}|\mathbf{B}=\mathbf{b})} \tag{13}$$

wobei sich die unnormierte Tafel Q berechnet durch:

$$Q(\mathbf{U}|\mathbf{B}=\mathbf{b}) = \prod P(X|\mathbf{Pa}(\mathbf{X})) \bullet \prod_{Y\in\mathbf{B}} F(Y) \tag{14}$$

Die A-Posteriori-Tafel eines Knotens A (A kann beobachtet oder unbeobachtet sein) ergibt sich durch:

$$P(A|\mathbf{B}=\mathbf{b}) = \sum_{\mathbf{X}\setminus A} \alpha Q(\mathbf{U}|\mathbf{B}=\mathbf{b}) \tag{15}$$

Dabei ist α der Normierunsgfaktor, der nach dem *Distributivgesetz* auch aus Marginalisierungen ausgeklammert werden kann:

$$P(A|\mathbf{B}=\mathbf{b}) \quad = \quad \frac{\sum_{\mathbf{X}\setminus A} Q(\mathbf{U}|\mathbf{B}=\mathbf{b})}{\alpha}$$

Dieses Herangehensweise ist insbesondere auch unabhängig von der Frage wie die Finding-Vektoren aufgebaut sind, was im Folgenden noch wichtig wird. Mit Hinblick darauf soll die Schreibweise modifiziert werden: Das bislang mit $\mathbf{B}=\mathbf{b}$ bezeichnete und durch Finding-Vektoren beschriebene Wissen soll nun mit β abgekürzt werden (z.B. $P(\mathbf{X}|\beta)$ statt $P(\mathbf{X}|\mathbf{B}=\mathbf{b})$). Mit β wird nicht mehr nur exaktes Wissen ("A ist in Zustand a_1") repräsentiert, sondern auch Wissen wie z.B.: "A ist nicht in Zustand a_1, kann aber noch in Zustand a_2 oder a_3 sein". Wie solche Information in einen Finding-Vektor zu übersetzen ist wird noch behandelt, es gilt aber:

$$P(\mathbf{Z}|\beta) = P(\mathbf{Z}) \bullet \prod_{A\in\mathbf{Z}} F(A), \quad \mathbf{Z} \subseteq \mathbf{X} \tag{16}$$

5.2 Komplexitätsreduktion durch Zwischenmarginalisierungen

Dieses Verfahren über die (modifizierte) Kettenregel scheitert aber in nicht trivialen PN's an der Berechnung der Verbundwahrscheinlichkeit $P(\mathbf{X}|\beta)$ (bzw. $Q(\mathbf{X}|\beta)$), da diese Wahrscheinlichkeitstafel mit der Anzahl der Knoten im PN exponentiell wächst. Bei N Knoten, die alle (nur) zwei Zustände haben umfasst diese Tafel 2^N Einträge. Das sind z.B. für $N = 20$ Knoten schon $2^{20} = 1048576$ Werte. Allerdings kann die maximal auftretende Tafelgröße, wie auch die Anzahl der Rechenoperationen oft erheblich reduziert werden, wenn die bereits "zwischendurch" marginalisiert wird. Grundlage dafür ist einfach das *Distributivgesetz* von Addition und Multiplikation. Für die hier verwendeten Wahrscheinlichkeitstafeln gilt z.B.:

$$\sum_D (P(A,B,C) \bullet P(B,C,D,E)) = P(A,B,C) \sum_D P(B,C,D,E) \tag{17}$$

Da D ausschliesslich in $P(B,C,D,E)$ auftritt kann der Rest ($P(A,B,C)$) aus der Marginalisierung (Summierung) über D ausgeklammert werden. Dies ist unabhängig davon, ob wie hier angegeben mit Verbundwahrscheinlichkeiten oder aber (auch) mit bedingten Wahrscheinlichkeiten gerechnet wird[4]. Allerdings ist offensichtlich, dass niemals über eine Variable A die (noch) im Gegeben-Teil steht marginalisiert werden kann, da dazu mindestens der Term $P(A|\mathbf{Pa}(\mathbf{A}))$ ausgeklammert werden müsste.

In einem Netz mit drei Knoten in einer seriellen Verbindung der Form Ⓧ → Ⓩ → Ⓨ gilt z.B.:

$$\begin{aligned} P(Y|\beta) &= \sum_{X,Z} (P(X|\beta) \bullet P(Z|X,\beta) \bullet P(Y|Z,\beta)) \\ &= \sum_Z \left[\left(\sum_X P(X|\beta) \bullet P(Z|X,\beta) \right) \bullet P(Y|Z,\beta) \right] \end{aligned} \tag{18}$$

Analog zur seriellen Verbindung ist das Beispiel für die divergierende Verbindung der Form Ⓧ ← Ⓩ → Ⓨ:

$$\begin{aligned} P(Y|\beta) &= \sum_{X,Z} (P(X|Z,\beta) \bullet P(Z,\beta) \bullet P(Y|Z,\beta)) \\ &= \sum_Z \left[\left(\sum_X P(X|Z,\beta) \bullet P(Z|\beta) \right) \bullet P(Y|Z,\beta) \right] \end{aligned} \tag{19}$$

In beiden Fällen bleiben, obwohl die Beispiele noch trivial sind, die Tafeln in der Form mit der Zwischenmarginalisierung über X bereits kleiner (maximal 2×2) als ohne (maximal $2 \times 2 \times 2$). Entsprechend werden auch weniger Rechenoperationen benötigt. Sowohl in der Gleichung 18 als auch in der Gleichung 19 ist ein Berechnungsschema wie in Abbildung 5.2 zu erkennen.

Hierbei erfolgt eine *Informationsweiterleitung* zwischen zwei Knotenmengen $\mathbf{C_1} = \{X, Z\}$ und $\mathbf{C_2} = \{Y, Z\}$ - den sogenannten *Cliquen* - über die Schnittmenge beider Cliquen - den sogennanten *Separator* $\mathbf{S_2^1} = \{Z\} = \mathbf{C_1} \cap \mathbf{C_2}$.

Prinzipiell kann auch in einer konvergierenden Verbindung der Form Ⓧ → Ⓩ ← Ⓨ über eine Zwischenmarginalisierung gearbeitet werden, um die Tafel $P(Y|\beta)$ zu berechnen.

$$P(Y|\beta) = \sum_{X,Y} (P(Z|X,Y,\beta) \bullet P(X|\beta) \bullet P(Y|\beta))$$

[4]Ganz allgemein können $P(A,B,C)$ und $P(B,C,D,E)$ auch durch zwei beliebige Funktionen $theta_1(A,B,C)$ und $theta_2(B,C,D,E)$ ersetzt werden

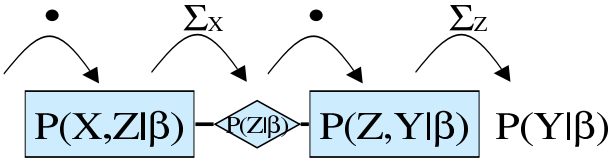


Abbildung 8:

$$= \sum_Z \left[\left(\sum_X P(Z|X,Y,\beta) \bullet P(X|\beta) \right) \bullet P(Y|\beta) \right] \tag{20}$$

Hierbei wird aber weder die maximale Tefelgrösse reduziert, da $P(Z|X,Y,\beta)$ bereits alle drei Knoten umfasst, noch gibt es einen Separator im obigen Sinne. Daher wird ein Knoten mit (mindestens) allen seinen Eltern immer in einer Clique zusammengefasst.

Es ergibt sich ein Verfahren das Schrittweise Knoten durch Marginalisierungen *eliminiert*. Um z.B. $P(F|\beta)$ zu berechnen muss die in Abbildung 5.2 dargestellte Reihenfolge von Marginalisierungen eingehalten werden (die beiden mit 1 nummerierten Schritte können gleichzeitig ausgeführt werden). Um z.B. Knoten $P(A|\beta)$ zu berechnen ergäbe sich eine andere Reihenfolge, jedoch bei gleicher Cliquen- Separatoren-Struktur. Diese Abhängigkeit der Berechnung von der zu Berechnenden A-Posteriori-Verteilung ist ein Problem, ein weiteres ist die Frage, wie die Cliquen bestimmt werden.

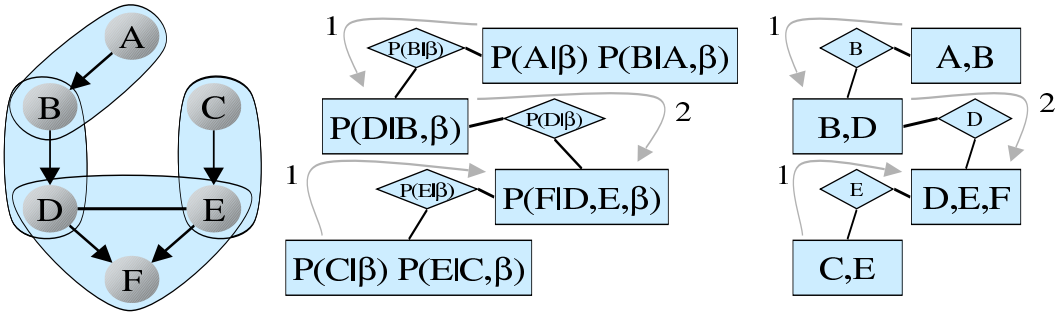


Abbildung 9: *links: Einfaches Netz mit Cliquen, mitte: Cliquen und Separatoren, rechts: vereinfachte Darstellung*

In Abbildung 5.2 sind die Elternknoten D, E des gemeinsamen Kindknotens F durch eine ungerichtete Kante verbunden worden. Diesen Vorgang alle Elternknoten gemeinsamer Kindknoten paarweise durch ungerichtete Kanten zu verbinden nennt man *Moralisierung*, den entstehenden Graph *moralischen Graph*[5]. Die gerichteten Kanten des DAG's werden im moralischen Graph normalerweise durch ungerichtete Kanten ersetzt. Maximal vollständig verbundene Teilgraphen des moralischen Graphen bilden die Cliquen. Zusammen mit einer geeigneten Wahl der Separatoren zur Verbindung

[5] über die Namensgebung kann man streiten, sie hat sich aber so durchgesetzt

der Cliquen entsteht ein sogenannter *Junction Tree*.

Definition 18 *Ein Teilgraph G ist* **vollständig verbunden**, *wenn für je zwei Knoten A und B aus G gilt, dass sie durch eine Kante verbunden sind.*

Definition 19 *Ein vollständig verbundener Teilgraph G heisst* **maximal vollständig verbunden** *wenn jeder weitere Knoten den man in G aufnähme, zu einer Verletzung der Bedingung vollständig verbundener Graphen führen würde.*

Satz 9 *In jedem vollständig verbundenen Teilgraphen G des moralischen Graphen existiert ein Knoten, der alle anderen Knoten in G als Elternknoten hat.*

Leider lassen sich die Cliquen so nur bestimmen, wenn der moralische Graph keine nicht abkürzbaren Zyklen der Länge > 3 enthält.

Definition 20 *Ein Zyklus mit den Knoten A und B ist* **abkürzbar**, *wenn es im Graph eine Kante $A - B$ gibt, die nicht Teil des Zyklus' ist.*

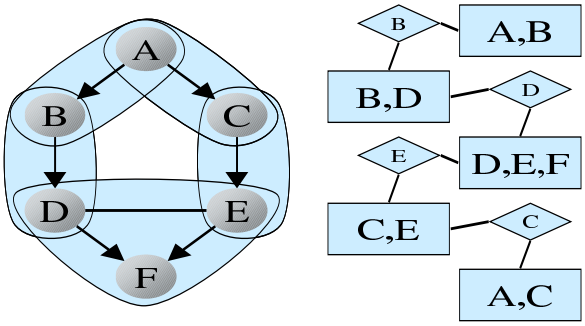


Abbildung 10: *DAG mit Zyklus, moralischer Kante $D - E$ und Cliquen* - **nicht** *korrekter Junction-Tree*

Es gibt keine Zwischenmarginalisierungsreihenfolge, die mit den in Abbildung 5.2 dargestellten Cliquen korrespondiert. Der dargestellte **nicht** korrekte Junction Tree ist nur eine von mehreren Möglichkeiten die Cliquen zu verbinden, das Problem bleibt jedoch immer das gleiche. Jeder dieser

Junction Trees verletzt die sogenannte *Running-Intersection-Eigenschaft.*

Definition 21 *Ein Junction-Tree erfüllt die* **Running-Intersection-Eigenschaft**, *genau dann wenn ein Knoten der in zwei Cliquen vorkommt auch in jeder Clique - und damit auch jedem Separator - auf dem Pfad zwischen diesen beiden Cliquen enthalten ist.*

Im hier gezeigten Junction-Tree erfüllt z.B. der Knoten A diese Eigenschaft nicht. Um diesem Problem zu begegnen wird der Graph vor der Cliquenbestimmung durch Einfügen weiterer Kanten - den sogenannten *Fill-Ins* - in einen *triangulierten Graphen* ohne entsprechende Zyklen überführt. Es gibt ein einfaches Verfahren um einen Graphen zu triangulieren und dabei gleichzeitig die Cliquen zu bestimmen. Dieses Verfahren *eliminiert* Schrittweise Knoten aus dem Graphen und funktioniert wie folgt[6]:

- Wähle einen zu eliminierenden Knoten X^e
- Verbinde alle Nachbarn von X^e paarweise durch eine Fill-In-Kante, sofern noch keine verbindende Kante existiert
- X^e und alle Nachbarn bilden eine neue Clique sofern nicht schon alle in einer vorher gefundenen Clique enthalten sind
- Entferne (eliminiere) X^e aus dem Graph
- Wiederhole die Schritte 1-4 bis keine Knoten mehr vorhanden sind.

Entscheidend für die Grösse der Cliquen und damit für die Qualität des späteren Junction Trees ist die Reihenfolge in der die Knoten eliminiert werden. Leider wäre die Suche nach der optimalen Lösung *NP-aufwändig* und muss im allgemeinen durch eine heuristische Suche ersetzt werden. Die einfachste Methode besteht darin immer einen Schritt voraus zu planen und in jedem Schritt den Knoten zur Eliminierung so auszuwählen, dass daraus eine Clique $\mathbf{C_i}$ mit der kleinsten Cliquentafel $P(\mathbf{C_i})$ resultiert - sofern überhaupt eine neue Clique entsteht. Als Unterkriterium bietet sich an eine Eliminierung so zu wählen, dass möglichst wenige Fill-Ins entstehen[7]. Die Verwendung dieses Unterkriteriums führt im Beispiel aus Abbildung 5.2 dazu dass F zuerst eliminiert wird. Andernfalls wäre auch jeder andere Knoten in Frage gekommen. Dieses Verfahren liefert in aller Regel gute Ergebnisse und man muss schon recht komplizierte Strukturen entwerfen um schwerwiegende Einbußen gegenüber einer optimalen Lösung zu erleiden.

Um dann aus allen Cliquen einen *Junction Tree* zu konstruieren bietet sich folgendes Verfahren

[6]Die Graphen-Triangulierung ist dabei eigentlich nur eine anschauliche Übersetzung der Schrittweisen Knoten-Eliminierung durch Zwischenmarginalisierungen

[7]teilweise wird dieses auch als Hauptkriterium vorgeschlagen

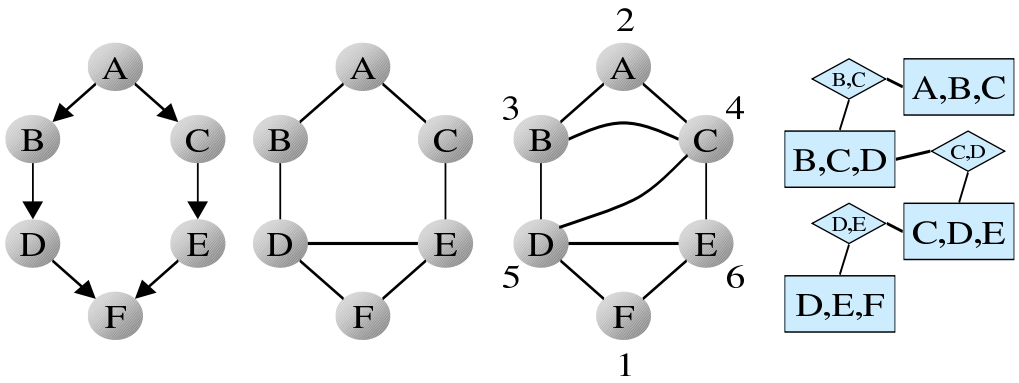


Abbildung 11: *DAG (mit Zyklus) - moralischer Graph - triangulierter Graph (mit Eliminierungs-Reihenfolge) - Junction Tree*

an:

- Bilde alle Separatoren für alle Cliquenpaare
- Markiere eine beliebige Clique als dem Junction Tree zugehörig und alle anderen als frei
- Wähle den mächtigsten[8] Separator, der eine Clique, die bereits dem Junction Tree angeschlossen ist mit einer noch freien Clique verbindet und markiere die noch freie Clique als ebenfalls zum Junction-Tree gehörig
- Wiederhole den vorigen Schritt bis alle Cliquen dem Junction-Tree angehören

Dieses Verfahren stellt insbesondere auch sicher, dass die *Running-Intersection-Eigenschaft* gewährleistet ist.

5.3 Inferenz im Junction-Tree

Es bleibt das zweite Problem, dass die gesamte Berechnung für jeden Knoten dessen A-Posteriori-Verteilung zu berechnen ist wiederholt werden muss. Angenommen man hätte für jede Clique $\mathbf{C^i}$ die Verbundwahrscheinlichkeitstafel $P(\mathbf{C^1})$. Information bezüglich der Knoten $A, B \in \mathbf{C^i}$ wird mittels der Multiplaktion mit den entsprechenden Findingvektoren eingebracht ($P(\mathbf{C^i}|\beta) = P(\mathbf{C^i}) \bullet F(A) \bullet F(B)$). Nun muss die Information an die benachbarten Cliquen z.B. $\mathbf{C^j}$ über den Separator $\mathbf{S^i_j}$ weitergegeben werden. Danach muss dann die empfangende Clique ihrerseits die Information an ihre übrigen Nachbarcliquen weiterleiten usw. Umgekehrt kann auch erst jede Information in den Junction-Tree eingebracht werden, indem jeder Finding-Vektor in eine das betreffende Merkmal enthaltende Clique einmultipliziert wird. Um danach die A-Posteriori-Verteilung eines Merkmals A zu erhalten muss die das Merkmal A enthaltende Clique $\mathbf{C^i}$ Information von ihren Nachbarcliquen empfangen, welche zuvor ihrerseits Information aller ihrer (übrigen) Nachbarcliquen empfangen haben

usw.

Definition 22 *Informationsweiterleitung von Clique* $\mathbf{C^j}$ *zur Clique* $\mathbf{C^i}$*:*

$$\begin{aligned} P(\mathbf{C^i}|\beta_t) &= \frac{P(\mathbf{C^i}|\beta_{t-1})}{\sum_{\mathbf{C^i}\setminus\mathbf{S^i_j}} P(\mathbf{C^i}|\beta_{t-1})} \bullet \sum_{\mathbf{C^j}\setminus\mathbf{S^i_j}} P(\mathbf{C^j}|\beta_t) \\ &= \frac{P(\mathbf{C^i}|\beta_{t-1})}{P(\mathbf{S^i_j}|\beta_{t-1})} \bullet P(\mathbf{S^i_j}|\beta_t) \end{aligned}$$

Dabei ist β_t *das Wissen zu einem Zeitpunkt* t*. Vor dem ersten Schritt* ($t = 0$) *liegt noch kein Wissen vor, diese "Null-Evidenz" sei mit* β_0 *bezeichnet* ($P(A|\beta_0) = P(A)$)*.*

Anmerkung: Die etwas komplizierte Schreibweise soll verdeutlichen, dass diese *Informationsweiterleitung* mehrfach ausgeführt werden kann z.B. wenn für weitere zuvor noch unbeobachtete Knoten Evidenz vorliegt. Dabei ist das neue $P(\mathbf{S^i_j}|\beta_{t-1})$ gleich dem alten $P(\mathbf{S^i_j}|\beta_t)$ aus der letzten Informationsweiterleitung - daher kann hier Aufwand vermieden werden wenn diese Tafel für weitere Informationsweiterleitungen aufgehoben wird. Ohne neue Evidenz verändert eine erneute Anwendung der Gleichung 21 dagegen offensichtlich nichts, da dann $P(\mathbf{S^i_j}|\beta_t) = P(\mathbf{S^i_j}|\beta_{t-1})$ gilt.

Auch dieses Verfahren hat noch einen Nachteil: Entweder muss jede Clique für die Information vorliegt diese einzeln über den gesamten Junction-Tree *propagieren* oder die gesamte Information muss **zu** jeder Clique aus der man die A-Posteriori-Verteilung eines Knotens ableiten will propagiert werden.

Es ist jedoch möglich beide Ansätze zu kombinieren, so dass nach zweimaliger Informationsweiterleitung über alle Separatoren alle Evidenz über den gesamten Junction-Tree verteilt ist. Dazu wird eine beliebige Clique zur *Root-Clique* des Junction-Trees ernannt. Nachdem alle Information in den Junction-Tree eingebracht ist wird eine Sequenz von Informationsweiterleitungen in der Form einer umgekehrten Tiefensuche zu dieser Root-Clique hin durchgeführt (*Collect-Phase*). Danach wird in umgekehrter Richtung von der Root-Clique zurück zu den entferntesten Knoten (*Leaves*) propagiert (*Distribute-Phase*).

Diese Form der Informationsweiterleitung über den gesamten Junction-Tree heisst *HUGIN-Propagation*. Die fällige Renormierung kann dadurch erfolgen, dass zwischen Collect- und Distribute-Phase die Tafel der Wurzel-Clique $\mathbf{C^w}$ renormiert wird ($P(\mathbf{C^w}|\beta) = Q(\mathbf{C^w}|\beta)/\sum Q(\mathbf{C^w}|\beta)$).

Nach einer solchen Propagierung besitzen zwei Cliquen $\mathbf{C^1}$ und $\mathbf{C^2}$ die das selbe Merkmal A

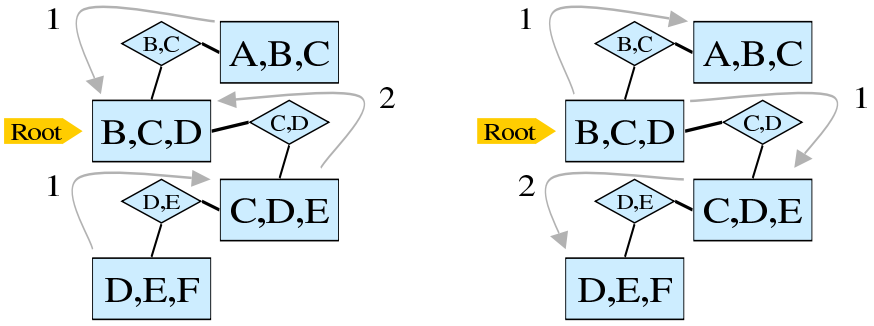


Abbildung 12: *links: Collect-Phase; rechts: Distribute-Phase*

enthalten auch dieselbe Information über A, d.h. der Junction-Tree befindet sich im *Equilibrium*.

Definition 23 *Der Junction-Tree befindet sich im* **Equilibrium**, *wenn für je zwei Cliquen* $\mathbf{C^1}$ *und* $\mathbf{C^2}$ *und jeden Knoten* $A \in \mathbf{C^1} \cap \mathbf{C^2}$ *gilt:*

$$\sum_{\mathbf{C^1} \setminus A} P(\mathbf{C^1}|\beta) = \sum_{\mathbf{C^2} \setminus A} P(\mathbf{C^2}|\beta) \tag{21}$$

5.4 Junction-Tree Initialisierung

Offen ist noch die Frage wie die Verbundwahrscheinlichkeitstafeln der Cliquen und Separatoren zu erhalten sind. Dazu werden alle Cliquen- und Separatorentafeln mit "1" initialisiert. Danach werden alle bedingten Wahrscheinlichkeitstafeln der einzelnen Knoten zu der Wahrscheinlichkeitstafel **einer** Clique multipliziert, die sowohl den Knoten selbst als auch alle seine Eltern enthält. Dass es mindestens eine solche Clique gibt ist dadurch gewährleistet, dass durch die Moralisierung ein Knoten und seine Elternknoten zu einem vollständig verbundenen Teilgraphen werden, dieser muss wiederum Teilmenge eines maximalen vollständig verbundenen Teilgraphen - also einer Clique - sein. Eine solche Clique wird auch als *Home-Clique* bezeichnet.

Wie auch anhand des Beispiels in Abbildung 5.4 zu erkennen ist muss nach dieser Initialisierung nur einmal Propagiert werden, um alle Tafeln in die gesuchten Verbundwahrscheinlichkeiten umzuwandeln und das *A-Priori-Equilibrium* herzustellen[9].

5.5 Unverbundene Teilgraphen

Unverbundene Teilgraphen ergeben ebenso unverbundene *moralische* und *triangulierte Graphen*. Damit gibt es bei der Baumgenerierung irgendwann keinen Separator $\mathbf{S_j^i} \neq \emptyset$ mehr, obwohl noch nicht

[9]Der entscheidende "Trick" besteht in der Initialisierung der Separatorentafeln mit dem Wert "1". Dadurch bleiben die bedingten Wahrscheinlichkeits-Tafeln zuächst von der Divison durch die Separatoren-Tafeln unverändert.

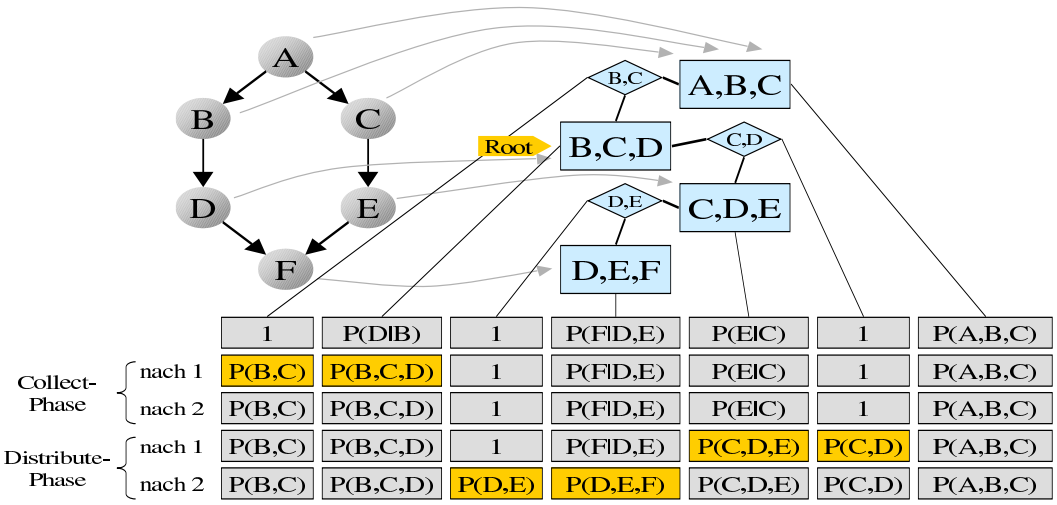


Abbildung 13: *Tafeln des Junction-Trees nach der Initialisierung und den einzelnen Propagations-Phasen*

alle Cliquen dem Baum angeschlossen sind. Das Problem kann gelöst werden indem man $\mathbf{S_j^i} = \emptyset$ als gültigen Separator zulässt und die dazugehörige Tafel $P(\mathbf{S_j^i}) = P(\emptyset) := 1$ setzt[10]. Da ein solcher Separator ohnehin keine Information weiterleiten würde (vgl. Gleichung 21) kann man unverbundene Teilgraphen auch von vorn herein getrennt betrachten und einen *Wald* von einzelnen *Junction Trees* pro Teilgraph aufstellen.

5.6 Evidenz-Arten

Bisher wurden Merkmale immer als beobachtet d.h. der Zustand ist mit Sicherheit bekannt oder als unbeobachtet angenommen. Wissen über ein beobachtetes Merkmal wurde durch ein Finding-Vektor ausgedrückt der genau eine "1" und ansonsten Nullen enthielt. Mit mehr als einer "1" kann ausgedrückt werden dass mehrere Zustände möglich sind und nur diejenigen ausgeschlossen werden können, für die der Finding-Vektor eine "0" enthält.[11].

Ein solcher Finding-Vektor differenziert nicht zwischen den nicht ausgeschlossenen Zuständen. Durch unterschiedliche positive und von Null verschiedene Werte im Finding-Vektor kann der Benutzer seinen Glauben in die einzelnen Möglichkeiten (möglichen Zustände) spezifizieren. Ein in dieser Form gegebenes Wissen soll als *Soft-Evidence* bezeichnet werden. Für die bedingten (Un-)Abhängigkeiten in einem PN bedeutet Soft-Eviden dass:

- in einer seriellen oder divergierenden Verbindung der Informationsfluss nicht geblockt (sondern nur "abgeschwächt") wird.
- eine Abhängigkeit über eine konvergierende Verbindung hergestellt wird.

Die Angabe einer Soft-Evidenz ist dabei identisch zu einer "harten" Evidenz an einem zusätzlichen (gedachten) Kindknoten. Das wird in Abbildung 5.6 verdeutlicht. Das Verhältnis 0.8 : 0.2 im Finding

[10] Dieser Ansatz ist allerdings unschön, insofern dass normalerweise $P(\emptyset) = 0$ gilt. Im Gegensatz zu dem *unmöglichen Ereignis* ist "$\emptyset$" hier aber als "über alle Knoten marginalisiert" zu verstehen, denn es gilt: $\sum_{\mathbf{V}} P(\mathbf{V}) = 1$.

[11] Damit kann ein Finding auch totales Unwissen über eine (unbeobachtete) Variable wiederspiegeln, indem es für alle Zustände eine "1" enthält.

für Knoten A links entspricht dem Verhältnis 0.4 : 0.1 der linken Spalte der CPT des Knotens $SoftEvid_A$ rechts. Diese Spalte wurde durch die harte Evidenz für diesen Kind-Knoten ausgewählt.

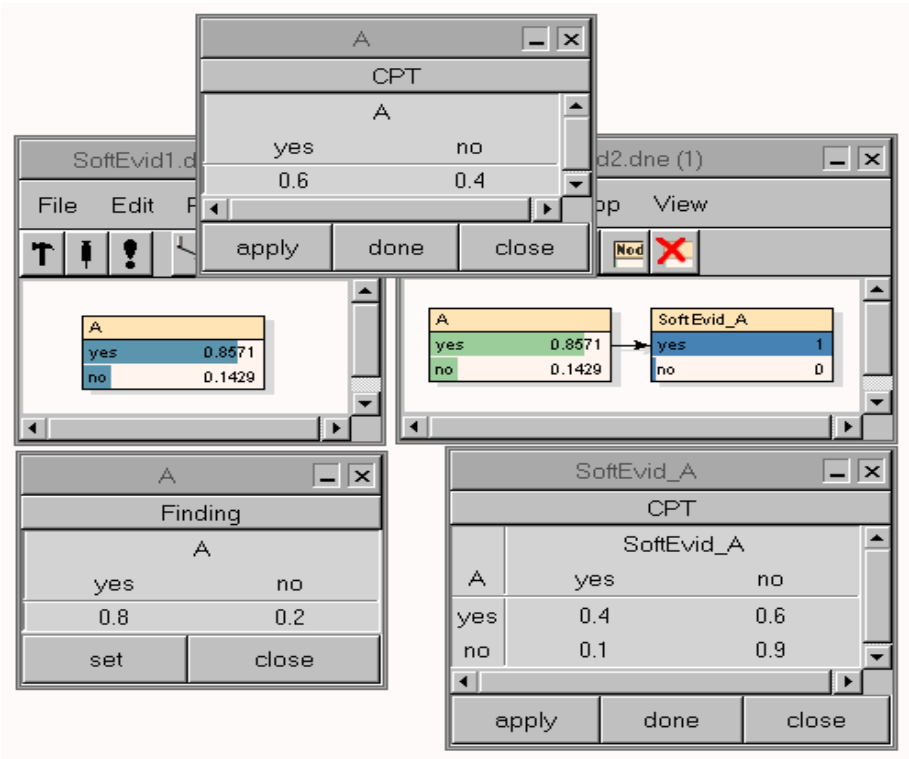


Abbildung 14: *Soft Evidenz links, Soft Evidenz als zusätzlicher Kindknoten rechts*

6 Dynamische Probabilistische Netze

In den vergangenen Abschnitten wurden nur statische Probleme behandelt, d.h. zu einem Zeitpunkt liegen Beobachtungen vor und nur aus diesen sollen Rückschlüsse gezogen werden. Viele Probleme sind aber dynamischer Natur, d.h. der momentane Zustand hängt neben den aktuellen Einflüssen auch von einer Historie vergangener Einflüsse ab.

6.1 Markov-Bedingung

Dabei können oft bestimmte einschränkende Annahmen über die zeitlichen Abhängigkeiten gemacht werden. Ist in einem Modell die Zukunft unabhängig von der Vergangenheit gegeben den gegenwärtigen (Gesamt-)Zustand so ist die *Markov Bedingung erster Ordnung*, oder kurz die *Markov Bedingung* erfüllt. Gilt die Unabhängigkeit zwischen Zukunft und Vergangenheit nur bei vollständiger Kenntnis des Gesamtzustands von $i > 1$ Zeitschritten, so spricht man auch von der *Markov Bedingung i-ter Ordnung*.

Dynamische Probabilistische Netze (DPN's) wie sie hier verwendet werden erfüllen immer die *Markov Bedingung (erster Ordnung)*. Sie unterscheiden sich von statischen PN's dadurch, dass die Knoten bestimmten Zeitschritten (*Zeitscheiben*) zugeordnet sind und es kausale Verbindungen von Knoten eines Zeitschritts zu Knoten des nächsten Zeitschritts gibt. In einem (dynamichen) PN sind

normalerweise nicht - wie in der Markov Bedingung formuliert - alle Merkmale eines Zeitschritts beobachtet, denn es ist ja gerade die Aufgabe bestimmte unbeobachtete Merkmale zu schätzen. Daher kann eine Abhängigkeit von vergangenen Zeitpunkten über eine quasi unendliche Anzahl von Zeitscheiben bestehen[12].

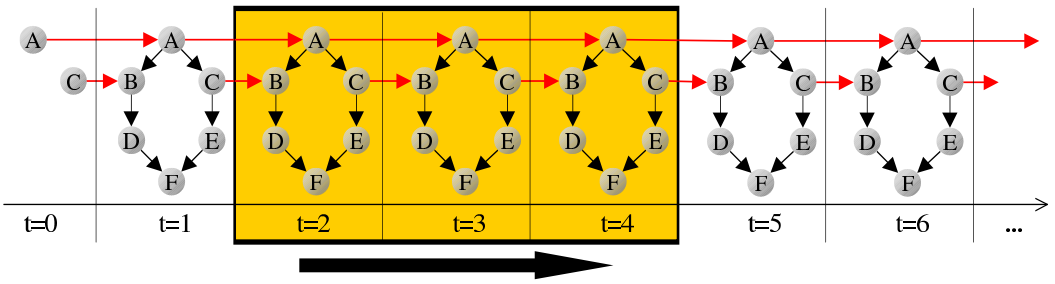


Abbildung 15: *DPN mit Zeitfenster, links Prior-Slice ($t = 0$), weitere Zeitschritte ($t > 0$) sind stationär*

6.2 Stationäre Systeme

Die hier behandelten Systeme sind *stationär* - d.h. dass sich in allen *Zeitscheiben* dieselben Knoten und dieselbe Struktur mit denselben bedingten Wahrscheinlichkeiten wiederfinden (siehe Abbildung 6.1). Ausgenommen davon ist nur die erste Zeitscheibe ($t = 0$) die sogenannte *Prior-Zeitscheibe*, die sich schon alleine deshalb unterscheiden muss weil die Knoten hier nicht von einer früheren Zeitscheibe abhängen können. Es soll aber auch gestattet sein für die Knoten innerhalb dieser Zeitscheibe eine komplett andere Struktur vorzugeben. Minimal erforderlich sind nur diejenigen Knoten die Kindknoten in der zweiten Zeitscheibe ($t = 1$) haben, wie in Abbildung 6.1 dargestellt.

6.3 Aufrollen eines Netzes

Zur Definition eines dynamischen PN's werden demnach nur *Prior-* und zweite Zeitscheibe sowie die Anzahl der Zeitscheiben T benötigt. Danach könnte das Netz für diese Anzahl von Zeitscheiben expandiert werden, indem die Zeitscheibe für $t = 1$ entsprechend oft kopiert und dem Netz hinzugefügt wird. Der Juction-Tree für dieses *aufgerollte* Netz wird dann wie bei einem statischen Netz aufgestellt. Problematisch wird dieses Verfahren aber dann, wenn über eine lange oder sogar quasi-unendliche Zeitspanne gerechnet werden soll. Dazu müsste ein gigantisches Netz aufgebaut und in jedem Zeitschritt durchpropagiert werden. D.h. es würde bereits aus den Informationen zum Zeitpunkt $t = 0$ eine Prognose der Knoten der letzten Zeitscheibe berechnet und umgekehrt würde auch zum letzten Zeitpunkt der Einfluss bis in die Prior-Zeitscheibe zurückverfolgt.

6.4 Zeitfenster

Bei langen Zeitreihen ist es normalerweise ausreichend lediglich die Knoten eines bestimmten *Zeitfensters* zu bestimmen (Abbildung 6.1). Das bedeutet jedoch nicht, dass der Einfluss von Zeitscheiben, die aus dem Zeitfenster rutschen verloren ginge, es wird nur nicht mehr der Informationsgehalt

[12] i.d.R. werden diese aber nach wenigen Zeitschritten extrem klein

späterer Beobachtungen für solche Zeitscheiben zurückverfolgt. Das bedeutet auch das Evidenz immer nur für das aktuelle Zeitfenster eingebracht werden kann. Die Ergebnisse zu einem Zeitpunkt t im aktuellen Zeitfenster sind jeweils identisch mit denen eines entsprechend groß aufgerollten Netzes.

In Abbildung 6.1 ist ein drei Zeitscheiben umfassendes Zeitfenster dargestellt. Zu einem Zeitpunkt t besteht immer nur das Teilnetz für das aktuelle Zeitfenster. Rückt dieses Zeitfenster um einen Zeitschritt weiter, so fällt "hinten" eine Zeitscheibe heraus und "vorne" muss eine hinzugefügt werden. Das macht aber nur Sinn, wenn es möglich ist den dazugehörigen Junction Tree ebenfalls dynamisch hinten um einzelne Zeitscheiben kürzen und vorne entsprechend erweitern zu können.

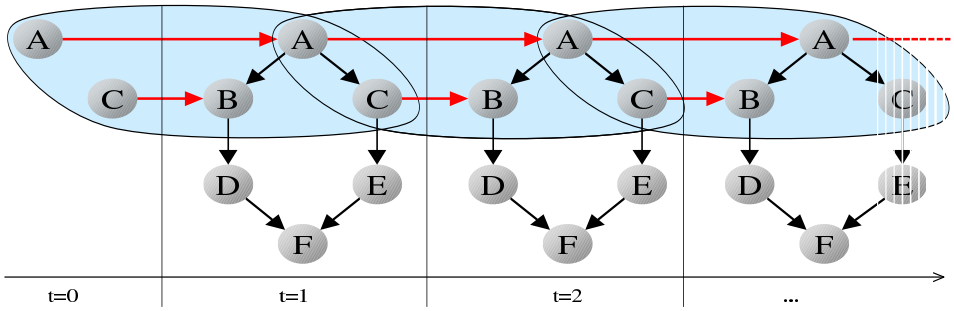


Abbildung 16: *DPN mit "dynamischer Clique"*

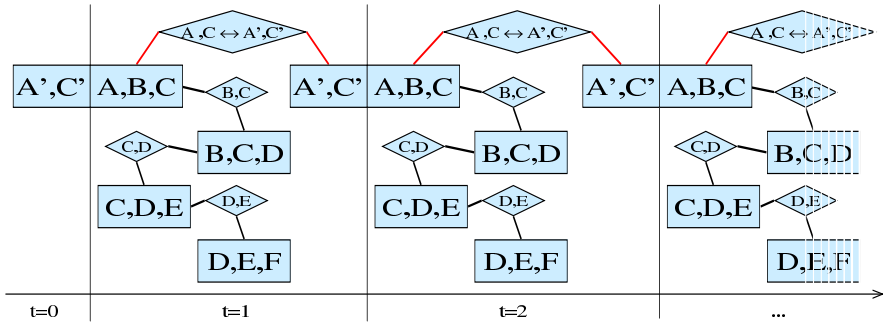


Abbildung 17: *Dynamischer Junction Tree mit "dynamischer Clique" und "dynamischem Separator"*

6.5 Replizierbarer Junction-Tree

Der Junction Tree muss - wie das DPN auch - eine für jede Zeitscheibe ($t > 0$) identisch replizierbare Form haben. Obwohl immer nur der Junction-Tree für das aktuelle Zeitfenster tatsächlich gebildet wird muss der (virtuelle) Gesamt-Junction-Tree (für alle Zeitscheiben) natürlich trotzdem die Bedingungen an eine Baumstruktur erfüllen. Dieses kann nur gewährleistet werden, wenn die Teil-Junction-Trees einzelner Zeitscheiben über einen einzigen Separator verbunden werden. Das aber erfordert eine Clique, die alle relevanten Knoten zweier benachbarter Zeitscheiben in sich vereinigt. Das sind alle Knoten die im moralischen Graph Verbindungen zur nächsten/vorigen Zeitscheiben haben. Werden diese Knoten paarweise verbunden ist das Vorhandensein der benötigten Clique nach Anwendung des Eliminierungs-Algorithmus' gewährleistet.

In Abbildung 6.5 ist für die Zeitscheibe $t = 2$ eine zusätzliche Clique[13] mit den Knoten $A(t = 2)$ und $B(t = 2)$ dargestellt. Ohne diese Clique bräuchte man aber zwei Teil-Junction-Trees um alle Knoten der Zeitscheibe $t = 2$ gleichzeitig zur Verfügung zu haben. Um dieses Problem zu vermeiden wird die Zusatz-Clique gebildet, die alle Knoten der aktuellen Zeitscheibe, die auch in der dynamischen Clique der nächsten Zeitscheibe auftreten beinhaltet. Ein solche Clique $\mathbf{C^1} = \{A, B\}$ im Junction-Tree, die eigentlich nur Teilmenge einer anderen Clique $\mathbf{C^2} = \{A', B', A, B\}$ ist kann problemlos verwendet werden, solange die Knoten in $\mathbf{C^1}$ keine Elternknoten haben, die nur in $\mathbf{C^2}$ und nicht in der $\mathbf{C^1}$ vorkommen. Das ist durch die Einschränkung der Richtung dynamischer Kanten offensichtlich der Fall.

Dieses Verfahren hat den Nachteil, dass eine zusätzliche Clique ($\{A, B\}$) und damit ein zusätzlicher Separator ($\{B\}$) entsteht, so dass zwei Informationsweiterleitungen (für Collect- und Distribute-Phase) mehr pro Zeitschritt erforderlich werden. Auf der anderen Seite müsste aber ein Junction-Tree benutzt werden der um eine Zeischeibe länger ist als das eigentlich zu betrachtende Zeitfenster um sicher alle Knoten dieses Zeitfensters zur Verfügung zu haben, was insbesondere bei einem Zweitfenster welches nur eine Zeitscheibe umfasst der weitaus größere Nachteil wäre.

Der replizierbare Junction-Tree wird mit folgendem Verfahren aus einem PN der ersten beiden Zeitscheiben $t = 0$ und $t = 1$ gebildet:

- Moralisieren des PN's
- paarweise Verbindung aller Knoten in $t = 1$ ($t = 0$) die Verbindungen zu $t = 0$ ($t = 1$) haben.
- Verbindung der Knoten in $t = 1$ die auch in $t = 0$ verbunden wurden (Sicherstellen einer Anknüpfungs-Clique)
- Triangulation und Cliquenbildung mit dem Eliminierungs-Algorithmus'.
- Abstpalten der Cliquen, die nur Knoten aus $t = 0$ enthalten, um den replizierbaren Junction-Tree für alle Zeitscheiben $t > 0$ zu erhalten.

7 Ungerichtete Kanten

Die Verwendung dynamischer PN's löst wenigstend zum Teil die Problematik von gegenseitigen kausalen Abhängigkeiten, die wegen des Verbots gerichteter Zyklen nicht direkt zu realisieren sind. In einen DPN können die Zustände zweier Variablen X und Y zu einem Zeitpunkt t aber durchaus beide von beiden Vorgängerzuständen des Zeitschritts $t-1$ abhängen - ohne dass hier ein gerichteter Zyklus entstünde.

Eine andere - auch für statische PN's geeignete - Methode gegenseitige (nicht kausale) Abhängigkeiten abzubilden besteht in det Verwendung eines Zusaztknotens. Dieser ist gemeinsamer Kindknoten der in Relation zu setzenden Knoten. Er wird immer mit derselben Evidenz instanziiert, so dass die einzelnen Werte für diesen vorgegebenen Zustand über alle Elternzustände die Relation der Elternknoten abbilden. So zu sehen in der umrandeten Spalte der CPT des Verbindungsknotens AB in Abbildung 7. Der Knoten AB wird immer mit dem Zustand *yes* instanziiert, um die Abhängigkeit zwischen A und B in dieser konvergierenden Verbindung herzustellen.

[13] Dies ist keine "echte" Clique, da sie keinen **maximal** vollständigen Subgraph bildet.

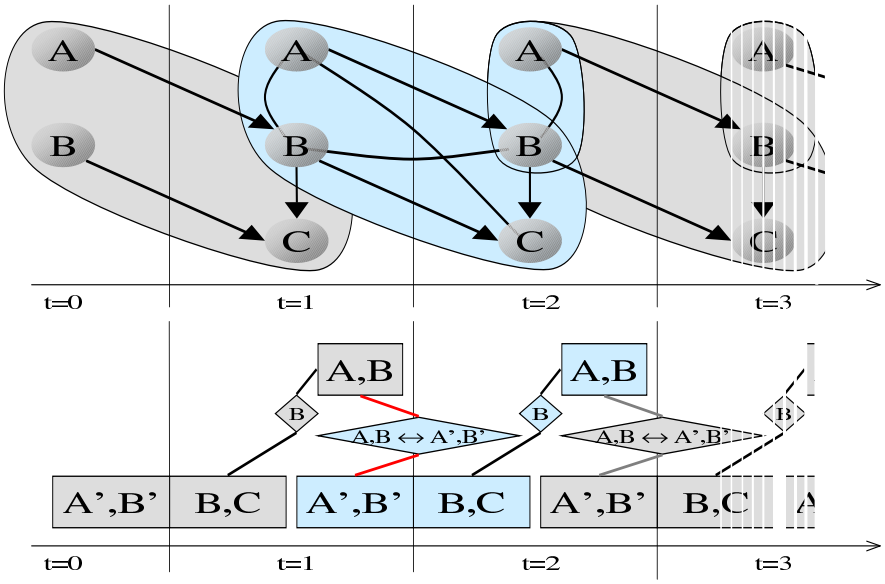


Abbildung 18: *Temporärer Knoten T zur Erzeugung der "Dynamische Clique"*

8 Modularisierte Netze

Mit einem ähnlichen Verfahren wie für dynamische PN's können auch *modularisierte* Netze entstehen. *Modularisiert* bedeutet dass Teilnetze getrennt aufgestellt werden und auch die entsprechenden Teil-Junction-Trees für jedes Teilnetz getrennt gebildet werden können, bevor diese Module zu einem Gesamtmodell zusammengefügt werden. Dadurch können Teile von Probabilistischen Netzen leichter wiederverwendet werden. Es wird dadurch auch möglich dynamisch verschiedene Teilnetze zu koppeln, d.h. zu einem Zeitpunkt z.B. das (Haupt-)Modul M^1 mit einem Modul M^2 und zu einem anderen Zeitpunkt mit einem Modul M^3 zu koppeln. Um zwei Teilnetze (Module) mit den Graphen G^1 und G^2 koppeln zu können müssen beide Netze eine Menge von gemeinsamen Knoten enthalten. Diese Menge muss alle Knoten umfassen, die im moralischen Graphen des Gesamtgraphs $G = G^1 \oplus G^2$ verbunden sind. Die Menge dieser Verbindungsknoten sei mit $\mathbf{V}$ bezeichnet. Bei der separaten Erzeugung der Junction-Trees JT^1 und JT^2 zu den beiden Teilgraphen muss gewährleistet sein, dass die Knoten in $\mathbf{V}$ in mindestens einer Clique $\mathbf{C^1}$ bzw. $\mathbf{C^2}$ beider Junction-Trees JT^1 bzw. JT^2 vereint sind. Damit können dann die Cliquen $\mathbf{C^1}$ und $\mathbf{C^2}$ über einen Separator verbunden werden. Um beide Junction-Trees JT^1 und JT^2 einzeln erzeugen zu können dürfen die Verbindungsknoten $\mathbf{V}$ jeweils nur in (maximal) einem Teilgraph (z.B. G^1) Elternknoten haben. Die Clique C^1 des Junction-Trees JT^1 des Teilgraphs G^1 enthält dann die korrekte A-Priori-Verteilung der Verbindungskonten $\mathbf{V}$ ($P(\mathbf{V}) = \sum_{\mathbf{C^1}\setminus\mathbf{V}} P(\mathbf{C^1})$. Es ist daher nach der Verbindung der Junction-Trees JT^1 und JT^2 eine initiale Informationsweiterleitung von C^1 nach C^2 (nicht umgekehrt) über den Separator $\mathbf{S_2^1} = \mathbf{V}$ vorzunehmen. In einem DPN kann nur ein Teilnetz (Modul) zeitscheibenübergreifende Kanten haben, da sonst der dynamische Separator nicht korrekt gebildet werden kann und der "Gesamt-Junction-Tree" keine Baumstruktur mehr hätte.

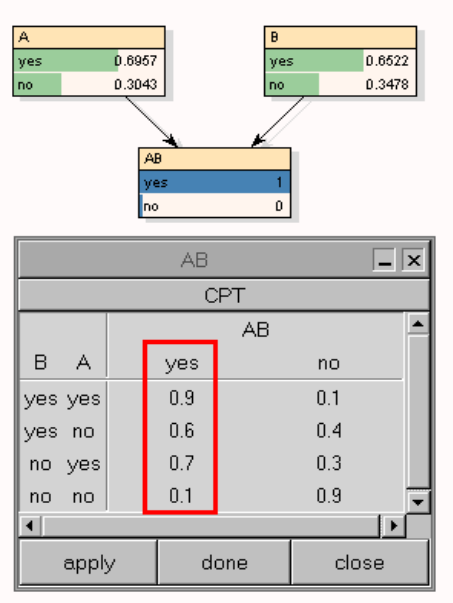


Abbildung 19: *"ungerichtete" Relation zwischen A und B über gemeinsamen Kindknoten AB*

9 Gewinnung eines konkreten Modells

Um ein PN für eine konkrete Problemstellung zu definieren muss:

- eine Auswahl der Merkmale getroffen werden
- eine kausale Struktur der Merkmale untereinander festgelegt werden
- ein Wahrscheinlichkeitsmodell quantifiziert werden, indem die bedingten Wahrscheinlichkeitstafeln entsprechend der kausalen Struktur aufgestellt und mit Werten gefüllt werden

Alle Schritte können von einem oder mehreren Experten vorgenommen werden oder unter Zuhilfename von Lernverfahren aus Daten bestimmt werden. Die Möglichkeit das Modell komplett durch menschliche Experten bestimmen zu können stellt einen großen Vorteil gegenüber anderen Verfahren, wie z.B. *Neuronalen Netzen* dar die als *Black-Box-Verfahren* keinen intuitiv verständliche Repräsentation ihres Wissens zur Verfügung stellen. Alle Bestandteile eines PN sind jedoch intuitiv erfassbar. Sowohl die *globale* Information gegeben durch die Netzstruktur, als auch die *lokale* Information bezüglich der bedingten Wahrscheinlichkeiten[14] kann aus Expertenwissen bestimmt oder durch Experten überprüft werden. Hauptschwierigkeiten dabei bereiten oft die elternlosen Knoten, da die erforderlichen A-Priori-Wahrscheinlichkeiten oft nicht genau angegeben werden können (das betrifft auch die Informationsgewinnung aus Daten). Der intuitiven Bewertung entziehen sich ausserdem noch die CPT's der Knoten die Elternknoten in der vorigen Zeitscheibe eines DPN's haben. Diese müssen oft in einem iterativen Prozess den Erwartungen angenähert werden. Auf die Problematik dynamischer PN's wird im Abschnitt **??** noch näher eingegangen.

Bei der Verwendung von Datenmaterial ist es am einfachsten die bedingten Wahrscheinlichkeitstafeln zu gewinnen, insbesondere wenn die Daten *vollständig* sind, d.h. wenn alle Merkmale des PN's zu jedem Zeitpunkt beobachtet sind. Dann entspricht das "Lernen" der Tafeln einem Auszählen der Fälle

[14] lokal, weil jeweils nur ein Knoten und seine Elternknoten betrachet werden

und normieren der so gewonnenen Tafeln. Nicht mehr ganz so trivial wird es wenn Merkmale nie oder zeitweise nicht beobachtet wurden. Hierfür existieren Verfahren wie der *Expectation-Maximization-Algorithmus (EM-Algorithmus)* oder auch *gradienten-basierte* Verfahren.

Problematischer ist das Finden der kausalen Struktur aus Daten. Auch hierfür exitstieren Lernverfahren (???), die Versuchen eine optimale Struktur aus Daten abzuleiten. Schwierigkeiten bereitet dabei die Interpretation der Kanten als kausale Abhängigkeit, da z.B. die beiden Netze Ⓐ → Ⓑ und Ⓐ ← Ⓑ genau dieselbe zugrundeliegende Verteilung $P(A, B)$ repräsentieren können, d.h. sie sind *equivalent*. Welches der beiden Netze (wenn überhaupt) der wahren kausalen Richtung entspricht kann aufgrund von Beobachtungen bezüglich A und B nicht unterschieden werden. In komplexeren Modellen sind viele Strukturen trotz sich nur in einzelnen Kantenrichtungen unterscheidenden Verbindungen nicht *equivalent* und daher prinzipiell auch unterscheidbar. So ist ein Drei-Knoten-Netz mit einer seriellen/divergierenden Verbindung nicht *equivalent* zu einem Netz mit einer konvergierenden Verbindung. Jedoch sind z.B. die zwei möglichen seriellen Verbindungen Ⓐ → Ⓑ → Ⓒ und Ⓐ ← Ⓑ ← Ⓒ und die divergierende Verbindung Ⓐ ← Ⓑ → Ⓒ wieder *equivalent* zueinander. Insbesondere sind auch immer alle vollständig verbundenen Strukturen *equivalent* zueinander. Das Problem der wahren kausalen Struktur ist daher mit Lernverfahren nicht vollständig lösbar. Ein weiteres Problem ist nicht auf PN's begrenzt, sondern gilt praktisch für alle Lernverfahren, die die Modellkomplexität (hier gegeben durch die Netzstruktur) variieren. Das Problem besteht darin, dass Daten praktisch immer durch zufällige Einflüsse verrauscht sind. Dadurch wird z.B. die Unabhängigkeitsbedingung für zwei Variablen A und B nie genau erfüllt sein. Im allgemeinen gilt auch für zwei unabhängige Merkmale:

$$\hat{P}(A) \bullet \hat{P}(B) \neq \hat{P}(A, B) \tag{22}$$

Wobei $\hat{P}(.)$ die aus den Daten geschätzten Verteilungen sind. Dies ist z.T. auch schon durch numerische Probleme begründet[15]. Würde man aber alle *Scheinabhängigkeiten* in das Modell aufnehmen so würde man im Extremfall einen vollständig verbundenen Graphen - mindestens jedoch einen mit deutlich zu vielen Kanten erhalten. Ein solches Modell wäre nicht nur zu komplex um damit effizient rechnen zu können, sondern es würde auch schlechtere Prognosen liefern als ein Modell, in dem nur die tatsächlichen Abhängigkeiten repräsentiert sind. Eine solche Überinterpretation von nur scheinbaren Abhängigkeiten durch Lernverfahren wird oft als *overfitting* bezeichnet. Das Problem ist also echte von scheinbaren Abhängigkeiten zu unterscheiden. Scheinabhängigkeiten können dabei als schwach angenommen werden. Was "schwach" ist hängt aber entscheidend von der Anzahl der Beobachtungen ab, da der Einfluss zufälligen Rauschens wie auch die rein numerischen Probleme mit zunehmendem Beobachtungsumfang abnehmen (*Gesetz der großen Zahlen*)

$$\lim_{N \to \infty} \hat{P}(\mathbf{X}) \to P(\mathbf{X}) \tag{23}$$

Dabei ist $\hat{P}(\mathbf{X})$ die aus den Daten abgeleitete Verteilung und $P(\mathbf{X})$ die zugrundeliegende wahre Verteilung.

Beispiel: Mit einem Würfel bei sechs Würfen zweimal die eins jedoch keine sechs zu würfeln sagt wenig darüber aus, ob der Würfel in Ordnung ist (jede Seite kann trotzdem mit der Wahrscheinlichkeit 1/6 geworfen werden), denn die Wahrscheinlichkeit bei sechs unabhängigen Würfen mit einem regulären Würfel keine sechs zu Würfeln beträgt: $(5/6)^6 \approx 1/3$ Wenn man jedoch nach 60 Würfen immer noch keine einzige sechs geworfen hat, so kann der Würfel mit großer Sicherheit als nicht in Ordnung angenommen werden. Die Wahrscheinlichkeit für eine solche Reihe beträgt bei einem

[15] Man stelle sich einen Datensatz vor, für den die Unabhängikeitsbedingung der Variablen A und B erfüllt ist - kommt nun noch eine Beobachtung hinzu wird die Unabhängikeitsbedingung i.d.R. nicht mehr erfüllt sein.

regulären Würfel nur noch $(5/6)^6\,00 \approx 1.8 \cdot 10^{-5}$
Auch für das Problem des Strukturlernens existieren verschiedene Algorithmen. Wird dabei ausserdem mit unvollständigen Daten gearbeitet müssen diese Verfahren mit den zuvor beschriebenen Verfahren kombiniert werden.

Für das letzte Problem der Merkmalsauswahl muss mindestens eine Grundauswahl vom Benutzer getroffen werden - schon allein durch die Auswahl der Merkmale die beobachtet werden. Je nach dem Beitrag für die Prognose der letztlich interessierenden Merkmale können dann u.U. auch automatisch wieder einzelne Variablen entfernt werden. Auch für den umgekehrten Weg existieren Verfahren. Hierbei wird versucht durch Hinzunahme neuer (natürlich unbeobachteter) Merkmale eine Komplexitätsreduktion zu erreichen. Wenn z.B. eine Menge von Merkmalen **W** von einer gemeinsamen Ursache U abhängen, die jedoch selbst im Netz nicht repräsentiert ist, so liefert ein Struktur-Lernverfahren oft einen vollständig verbundenen Teilgraphen für die Knotenmenge **W**. Wird ein gemeinsamer Elternknoten U eingeführt und die Verbindungen untereinander entfernt ergibt sich in diesen Fällen ein weniger komplexes die wahren Abhängikeiten jedoch besser widerspiegelndes Modell. Durch die Hinzunahme unbeobachteter Knoten müssen wiederum die erwähnten Verfahren zur Schätzung solcher Merkmale zum Einsatz kommen.

10 Wahrscheinlichkeitsrechnung

Axiom 4 *Die Wahrscheinlichkeitsfunktion $P(.)$ ordnet jedem Ereignis a des Ereignisraums Ω eine nicht negative reelle Zahl zu*

$$0 \leq P(a) \leq 1$$

Axiom 5 *:*

$$P(\Omega) \quad = \quad 1$$

Ω *ist das* **sichere Ereignis**

Axiom 6 *Für disjunkte Ereignisse a und b gilt:*

$$P(a \vee b) \quad = \quad P(a) + P(b)$$

Wobei die Ereignisse a und b Teilmengen des Ereignisraums darstellen ($a, b \subseteq \Omega$).

Satz 10 *Im allgemeinen gilt für zwei Ereignisse a und b (***Additionssatz***):*

$$P(a \vee b) \quad = \quad P(a) + P(b) - P(a, b)$$

Wobei $P(a, b)$ eine vereinfachte Schreibweise von $P(a \cap b)$ (Mengenschreibweise) bzw. $P(a \wedge b)$ (logisch) ist.

Definition 24 *Ein* **Merkmal** *umfasst eine Menge disjunkter Ereignisse*

Definition 25 *Zwei Ereignisse heissen* **(stochastisch) unabhängig**, *genau dann wenn gilt:*

$$P(a \wedge b) = P(a, b) \quad = \quad P(a)P(b)$$

Zwei Merkmale heissen **(stochastisch) unabhängig**, *genau dann wenn gilt:*

$$P(A \wedge B) = P(A, B) \quad = \quad P(A) \bullet P(B)$$

Satz 11 *Im allgemeinen gilt für zwei Ereignisse a und b bzw. zwei Merkmale A und B (***Multiplikationssatz***):*

$$\begin{aligned} P(a \wedge b) = P(a,b) &= P(a) \cdot P(b|a) = P(b) \cdot P(a|b) \\ P(A \wedge B) = P(A,B) &= P(A) \bullet P(B|A) = P(B) \bullet P(A|B) \end{aligned}$$

Satz 12 *Die bedingte Wahrscheinlichkeit für das Ereignis a, unter der Vorraussetzung dass b eingetreten (gegeben) ist, bzw. die bedingte Wahrscheinlichkeitsverteilung für das Merkmal A gegeben B ist:*

$$\begin{aligned} P(a|b) &= \frac{P(a,b)}{P(b)} \\ P(A|B) &= \frac{P(A,B)}{P(B)} \end{aligned}$$

Satz 13 Bayes'sche Formel:

$$P(A|B) = \frac{P(B|A) \bullet P(A)}{P(B)} = \frac{P(B|A) \bullet P(A)}{\sum_A P(B|A) \bullet P(A)}$$

Satz 14 *Die Verbundwahrscheinlichkeitsverteilung aller Merkmale einer Merkmalsmenge* $\mathbf{X}$ *eines PN's kann über die* **Kettenregel** *berechnet werden:*

$$P(\mathbf{X}) = \prod_i P(X^i|\mathbf{Pa}(\mathbf{X^i}))$$

Wenn (wie in PN's) gilt: $P(A|\mathbf{An}(\mathbf{A})) = P(A|\mathbf{Pa}(\mathbf{A}))$

11 Rechenregeln

Definition 26 Multiplikation von Wahrscheinlichkeits-Tafeln:

$$P(A,B) = P(A|B) \bullet P(B) \equiv P(a_i,b_j) = P(a_i|b_j)P(b_j), \quad \forall_{i,j}$$

Definition 27 Division von Wahrscheinlichkeits-Tafeln:

$$P(A|B) = \frac{P(A,B)}{P(B)} \equiv P(a_i|b_j) = \frac{P(a_i,b_j)}{P(b_j)}, \quad \forall_{i,j}, \quad \frac{0}{0} := 0$$

Definition 28 Marginalisierung:

$$P(B|C) = \sum_A P(A,B|C) \equiv P(b_j|c_k) = \sum_i P(a_i,b_j|c_k)$$

Satz 15 Normierung:

$$\begin{aligned} P(\mathbf{U}) &= \frac{Q(\mathbf{U})}{\sum_{\mathbf{U}} Q(\mathbf{U})} \\ P(\mathbf{U}|\mathbf{V}) &= \frac{P(\mathbf{U},\mathbf{V})}{P(\mathbf{V})} = \frac{P(\mathbf{U},\mathbf{V})}{\sum_{\mathbf{U}} P(\mathbf{U},\mathbf{V})} \\ P(\mathbf{U}|\mathbf{V}) &= \frac{Q(\mathbf{U},\mathbf{V})}{Q(\mathbf{V})} = \frac{Q(\mathbf{U},\mathbf{V})}{\sum_{\mathbf{U}} Q(\mathbf{U},\mathbf{V})} \end{aligned}$$

Wobei $Q(.)$ *eine unnormierte Tafel darstellt.*

12 Beweise und Definitionen

Satz 16 *In einem DAG existiert mindestens ein elternloser und ein kinderloser Knoten.*

Definition 29 *Zwei Knoten heissen* **benachbart** *oder* **Nachbarn**, *wenn sie durch eine Kante direkt verbunden sind.*

Definition 30 *Der zugehörige* **moralische Graph** *entsteht aus einem gerichteten Graph, indem je zwei Elternknoten eines gemeinsamen Kindknotens durch eine ungerichtete Kante verbunden werden und die gerichteten Kanten in ungerichtete Kanten überführt werden (durch Vernachlässigung der Richtungsinformation).*

Definition 31 *Ein Zyklus mit den Knoten A und B ist* **abkürzbar**, *wenn es im Graph eine Kante $A - B$ gibt, die nicht Teil des Zyklus' ist.*

Definition 32 *Ein Graph ist* **trianguliert**, *wenn jeder Zyklus von mehr als drei Knoten Länge abkürzbar ist.*

Lemma 2 *Ein Zyklus dem der Knoten A angehört gehören auch genau zwei Nachbarn von A an.*

Definition 33 *Ein zyklenfreier Graph heisst* **Baum** *(oder* **Tree***).*

Definition 34 *Ein Graph in dem alle Knotenpaare über (mindestens) einen Pfad verbunden sind heisst* **verbunden**.

Definition 35 *Ein* **Subgraph** *G' ist ein Graph, der aus einer Teilmenge der Knoten $\mathbf{X}' \subset \mathbf{X}$ eines anderen Graphen G sowie allen Kanten aus G besteht, die nur Knoten in $\mathbf{X}'$ verbinden.*

Definition 36 *Ein (Sub-)Graph heisst* **vollständig verbunden**, *wenn jedes Knotenpaar dieses (Sub-)Graphen durch eine Kante verbunden ist.*

Definition 37 *Ein Knoten heisst* **eliminierbar**, *wenn die Menge der Nachbarknoten des Knotens einen vollständig verbundenen Subgraphen bildet.*

Definition 38 *Ein Graph heisst* **eliminierbar**, *wenn es eine Eliminierungsreihenfolge der Knoten gibt, so dass alle Knoten eliminierbar sind. Dabei wird jeweils nur der Subgraph der noch nicht eliminierten Knoten betrachtet.*

Satz 17 *Ein Subgraph eines triangulierten Graphen ist ebenfalls trianguliert*[16].

Beweis 1 *Jeder Zyklus eines Subgraphen ist auch im Graphen selbst enthalten. Da ein solcher Zyklus im Graphen abkürzbar ist, ist er dies auch im Subgraphen.*

Satz 18 *Jeder eliminierbare Graph ist trianguliert.*

Beweis 2 *Wird ein Knoten A (entsprechend der Eliminierungsreihenfolge) eliminiert, so ist jeder Zyklus von mehr als drei Knoten Länge dem der Knoten A angehört abkürzbar, da alle Nachbarn von A paarweise verbunden sein müssen. Weil der Knoten A eliminiert wird und daher fortan in keinem Zyklus mehr auftritt, ist die Abkürzbarkeit für alle Zyklen dieses Knotens gezeigt. Entsprechendes gilt für alle Knoten in der Eliminierungsreihenfolge.*

[16] Ein Subgraph ist z.B. auch der Restgraph nach der Eliminierung eines Knotens

Definition 39 *Eine Knotenmenge* **S separiert** *zwei Knoten A und B wenn jeder Pfad von A nach B über mindestens einen Knoten aus* **S** *läuft.*

Definition 40 *Eine zwei Knoten A und B separierende Knotenmenge* **S** *heisst* **minimal**, *wenn aus* **S** *kein Knoten entfernt werden kann ohne dass ein Pfad von A nach B ohne Zwischenknoten aus* **S** *möglich wird. Das bedeuted auch, dass jeder Pfad von A nach B nur über genau einen Knoten aus* **S** *verläuft.*

Lemma 3 *Zwei nicht benachbarte Knoten sind in einem triangulierten Graph durch mindestens eine minimale, vollständig verbundene Knotenmenge separiert.*

Beweis 3 *Offensichtlich werden zwei nicht benachbarte Knoten durch die Menge aller anderen Knoten separiert. Diese Knotenmenge kann minimalisiert werden, um der Definition des minimalen Separators zu genügen*[17]*. Zu zeigen ist, dass eine solche minimale separierende Knotenmenge vollständig verbunden ist. Seien die Knoten A und B nicht benachbart und* **S** *eine sie separierende minimale Knotenmenge. Sind C und D zwei Knoten aus* **S** *so ist zu zeigen dass C und D benachbart sind. Da es mindestens einen Pfad von A nach B über C (aber nicht D) sowie mindestens einen über D (aber nicht C) gibt, existiert ein Zyklus mit den Knoten C und D und mindestens jeweils einem Knoten E bzw. F der von A bzw. B ausgehend vor erreichen des Separators auf beiden Pfaden besucht wird (diese Knoten können auch A und B selbst sein). Da der Graph trianguliert ist umfasst der kleinste solche Zyklus genau die vier Knoten C, D, E und F. Dieser muss aber auch noch abkürzbar sein - und die einzige Möglichkeit dafür ist eine Kante $C - D$. Eine Kante $E - F$ würde* **S** *umgehen und damit der Definition des Separators widersprechen.*

Satz 19 *Ist in einem durch Separierung entstandenen Subgraphen bestehend aus dem Separator* **S** *und der Menge der vom Rest separierten Knoten* **A** *ein Knoten* $A_i \in \mathbf{A}$ *eliminierbar, dann ist er auch im Gesamtgraphen eliminierbar.*

Beweis 4 *Der Subgraph* G^{AS} *mit den Knoten* $\mathbf{A} \cup \mathbf{S} \subseteq \mathbf{X}$ *enthält für jeden Knoten* $A_i \in \mathbf{A}$ *auch alle seine Nachbarknoten. Damit sind die Nachbarn im Subgraphen genau dann vollständig verbunden, wenn sie es auch im Graphen G mit allen Knoten* **X** *sind.*

Satz 20 *Jeder vollständig verbundene Graph ist sowohl trianguliert als auch eliminierbar.*

Beweis 5 *Alle Knoten sind verbunden, daher sind auch alle Nachbarknoten eines Knoten verbunden, und es sind auch alle Zyklen abkürzbar.*

Satz 21 *Jeder triangulierte Graph ist eliminierbar*

Vorüberlegung In einem nicht vollständig verbundenen Graphen existieren mindestens zwei nicht benachbarte Knoten A_i und B_j. Diese sind nach Satz ??? in einem triangulierten Graphen durch (mindestens) eine vollständig verbundene Knotenmenge **S** separiert. Dabei sei **A** die Menge der Knoten die auf irendeinem Pfad von A_i nach B_j vor **S** liegen. Entsprechend sei **B** die Menge der Knoten die in umgekehrter Pfadrichtung vor **S** liegen. Die Knoten A_i und B_j sind dabei selbst in der jeweiligen Knoten-Menge enthalten ($A_i \in \mathbf{A}$ und $B_j \in \mathbf{B}$). Der Separator **S** separiert alle Knoten aus **A** von allen Knoten aus **B**. Es entstehen mindestens zwei Subgraphen G^{AS}, G^{BS} bestehend aus den Knoten $\mathbf{A} \cup \mathbf{S}$ bzw. $\mathbf{B} \cup \mathbf{S}$.

[17] Es sind in der Regel verschiedene minimale Separatoren denkbar

Beweis 6 *Für einen vollständig verbundenen Graphen gilt Satz ???.*
Induktions-Vorraussetzung: *Ein Graph mit nur zwei Knoten ist immer trianguliert und eliminierbar.*
Induktions-Annahme: *Jeder triangulierte nicht vollständig verbundene Graph mit mehr als zwei und weniger als N Knoten besitzt mindestens zwei eliminierbare und nicht benachbarte Knoten*[18]*.*
Induktions-Schritt: *Ein Graph mit N Knoten ist entweder vollständig verbunden und damit nach Satz ??? eliminierbar oder durch einen Separator* $\mathbf{S}$ *in midestens zwei Subgraphen mit* $< N$ *Knoten separierbar (*$\mathbf{A} \cup \mathbf{S}$ *bzw.* $\mathbf{B} \cup \mathbf{S}$*). Diese Subgraphen sind entweder vollständig verbunden, so dass jeder Knoten zuerst eliminierbar wäre. oder es existieren mindestens zwei nicht benachbarte eliminierbarere Knoten pro Subgraph (z.B.* A_1*,* A_2 *bzw.* B_1*,* B_2*). Da nicht beide in* $\mathbf{S}$ *liegen können (denn in* $\mathbf{S}$ *sind alle Knoten benachbart), liegt mindestens einer in* $\mathbf{A}$ *(bzw.* $\mathbf{B}$*), so dass dieser auch im Gesamtgraphen eliminierbar ist. Daher hat der Gesamtgraph mit N Knoten mindestens zwei (einer pro Subgraph) eliminierbare Knoten. Der Restgraph nach Eliminierung von einem der in Frage kommenden Knoten ist nach Satz ??? wiederum trianguliert und es greift die Induktionsannahme bzw. -Voraussetzung.*

Definition 41 *Ein vollständig verbundener Subgraph, der* **maximal** *ist in dem Sinne, dass ihm kein weiterer Knoten hinzugefügt werden kann ohne dass die Eigenschaft vollständig verbunden zu sein verloren ginge heisst* **Clique***.*

Definition 42 *Ein* **Junction-Graph** *ist ein Hyper-Graph aller Cliquen eines Graphen. Nicht leere Schnittmengen - die sogenannten* **Separatoren** *- zweier Cliquen bilden Kanten zwischen zwei Cliquen. D.h. die Kante repräsentiert die Schnittmenge.*

Definition 43 *Ein* **Junction Tree** *ist ein Junction-Graph mit zwei Eigenschaften.*
1. Es existiert genau ein Pfad zwischen zwei Cliquen, d.h. es existieren keine Zyklen.
2. Ein Knoten der in zwei Cliquen $\mathbf{C^1}$ *und* $\mathbf{C^2}$ *enthalten ist, ist auch in allen Cliquen auf dem Pfad von* $\mathbf{C^1}$ *nach* $\mathbf{C^2}$ *und damit auch in jedem Separtor auf diesem Pfad enthalten (***Running-Intersection-Eigenschaft***).*

Satz 22 *Jeder Triangulierte Graph hat einen Junction-Tree.*

Beweis 7 Induktions-Vorraussetzung: *Der Satz ist offensichtlich für jeden Graphen mit höchstens zwei Knoten wahr.* **Induktions-Annahme:** *Der Satz ist für alle Graphen mit weniger als N Knoten wahr.* **Induktions-Schritt:** *Ein triangulierter Graph G mit N Knoten hat nach Satz ??? mindestens einen eliminierbaren Knoten A. Dieser bildet zusammen mit seinen Nachbarknoten eine Clique* $\mathbf{C}$*, da die Nachbarn untereinander und mit A verbunden sind und alle weiteren Knoten zumindest nicht mit A verbunden sind. Derselbe Graph ohne A (*G^**) habe - laut Induktions-Annahme - einen Junction-Tree* JT^* *mit einer Clique* $\mathbf{C}^* \supseteq \mathbf{C} \setminus A$*. Da nämlich alle Nachbarn von A auch in* G^* *vollständig verbunden sind, müssen sie auch in einer gemeinsamen Clique* $\mathbf{C}^*$ *enthalten sein. Nun lässt sich der Junction-Tree JT für G aus* JT^* *konstruieren:*
Ist $\mathbf{C}^* = \mathbf{C} \setminus A$ *dann kann A einfach zur bestehenden Clique* C^* *hinzugefügt werden.*
Ist $\mathbf{C}^* \supset \mathbf{C} \setminus A$*, dann wird* $\mathbf{C}$ *als neue Clique eingefügt und mit* $\mathbf{C}^*$ *über den Separator* $\mathbf{S} = \mathbf{C} \cap \mathbf{C}^*$ *verbunden. Da A nur in* $\mathbf{C}$ *vorkommt und alle Knoten die sowohl in* $\mathbf{C}^*$*, als auch in* $\mathbf{C}$ *vorkommen auch in* $\mathbf{S}$ *enthalten sind ist der so gewonnene Junction-Tree JT korrekt.*

[18] Diese Annahme ist stärker als für den Beweis notwendig, denn auch ein eliminierbarer Knoten würde reichen

Satz 23 *Sind nicht alle Knoten mit dynamischen Kanten in einem Separator vereint, so entsteht kein für alle Zeitscheiben identischer Junction-Tree-Ausschnitt.*

Beweis 8 *Die Annahme es sei möglich einen in allen Zeitschritten identischen Junction-Tree mit mehr als einem dynamischen Separator zu erzeugen führt zu einem Widerspruch. Seien* $\mathbf{C_t^1}$ *und* $\mathbf{C_t^2}$ *zwei Cliquen, die beide Knoten aus den Zeitscheiben* $t-1$ *und* t *enthalten. Seien* $\mathbf{C_{t+1}^1}$ *und* $\mathbf{C_{t+1}^2}$ *die Entsprechungen dieser Cliquen mit denselben Knoten zum Zeitpunkt* t *und* $t+1$. $\mathbf{S^1}$ *und* $\mathbf{S^2}$ *seien zwei Separatoren zwischen* $\mathbf{C_t^1}$ *und* $\mathbf{C_{t+1}^1}$ *bzw. zwischen* $\mathbf{C_t^2}$ *und* $\mathbf{C_{t+1}^2}$. *Es muss ein Pfad* $\mathbf{W_t^{1,2}}$ *zwischen* $\mathbf{C_t^1}$ *und* $\mathbf{C_t^2}$ *existieren - evtl. auch über weitere Cliquen aus unterschiedlichen Zeitscheiben hinweg (Def. Junction-Tree). Dann müsste, um identische Junction-Tree-Ausschnitte zu erhalten auch ein Pfad* $\mathbf{W_{t+1}^{1,2}}$ *existieren.* $\mathbf{W_{t+1}^{1,2}}$ *geht dabei aus* $\mathbf{W_t^{1,2}}$ *durch Verschiebung um einen Zeitschrtitt hervor. Damit können* $\mathbf{W_t^{1,2}}$ *und* $\mathbf{W_{t+1}^{1,2}}$ *nicht in allen Kanten identisch sein. Das aber bedeutet, dass es zwei Pfade von* $\mathbf{C_t^1}$ *nach* $\mathbf{C_t^2}$ *geben müsste: 1. den über* $\mathbf{W_t^{1,2}}$ *und 2. den über* $\mathbf{S^1}$, $\mathbf{W_{t+1}^{1,2}}$ *und* $\mathbf{S^2}$. *Das aber ist ein Widerspruch zur Defition des Junction-Trees.*

Satz 24 *Ein Graph mit* N *Knoten und* $N-1$ *Kanten, in dem alle Knotenpaare durch einen Pfad verbunden sind ist zyklenfrei und heisst daher Baum.*

Beweis 9 Induktions-Vorraussetzung: *Der Satz gilt offensichtlich für Graphen mit einem Knoten* **Induktions-Annahme:** *Der Satz gilt für alle Graphen mit* $< N$ *Knoten* **Induktions-Schritt:** *Um einen weiteren Knoten* A *in einen Graphen mit* $N-1$ *Knoten und* $N-2$ *Kanten einzufügen wird eine weitere Kante benötigt, die* A *mit einem bereits im Graphen enthaltenen Knoten* B *verbindet. Damit hat der entstehende Graph* N *Knoten und* $N-1$ *Kanten. Dieser Graph ist offensichtlich zyklenfrei, da ein Zyklus nur über* A *und die Kante* $A-B$ *entstehen kann, es aber von* A *keine andere Kante gibt.*

Satz 25 *Aus Satz ??? folgt auch, dass zwei einzelne Bäume, die über eine Kante verbunden werden zusammen wieder einen Baum bilden.*

Beweis 10 *Baum* T^1 *habe* N^1 *Knoten und* N^1-1 *Kanten. Baum* T^2 *habe* N^2 *Knoten und* N^2-1 *Kanten. Seien die Knoten* A *aus* T^1 *und* B *aus* T^2 *durch eine zusätzliche Kante verbunden. Damit hat der entstehende Graph* T $N = N^1 + N^2$ *Knoten und* $N^1 - 1 + N^2 - 1 + 1 = N - 1$ *Kanten. Da es für alle Knoten in* T^1 *einen Pfad nach* A *und für alle Knoten in* T^2 *einen Pfad nach* B *gibt, ist der entstehende Graph offensichtlich verbunden.*

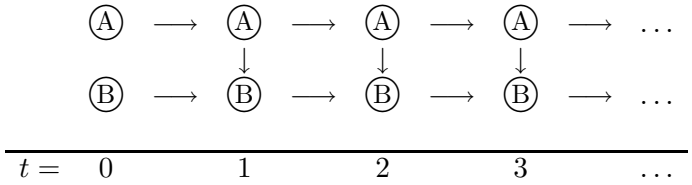
Satz 26 *Wenn die Subgraphen jeder Zeitscheibe trianguliert sind ist auch der Gesamtgraph trianguliert, wenn alle Knoten mit Verbindungen zwischen zwei Zeitscheiben* $t-1$ *und* t *vollständig verbunden sind.*

Beweis 11 *Da innerhalb einer Zeitscheibe* t *keine nichtabkürzbaren Zyklen der Länge* > 3 *auftreten müsste ein längerer nichtabkürzbarer Zyklus mit dem Knoten* A *über mindestens einen Knoten* D *einer anderen Zeitscheibe z.B.* $t-1$ *verlaufen. Um die Zeitscheibe* t *zu verlassen und wieder in* t *einzutreten müssen zwei verschiedene Knoten* B *und* C *mit* $t-1/t$*-Verbindung auf dem Zyklus-Pfad liegen. Dann ist der Zyklus aber abkürzbar, da laut Voraussetzung* B *und* C *verbunden sind.*

Satz 27 *Aus einem dynamischen Graphen in dem alle Knoten mit Zeitscheibenübergreifenden Kanten benachbart sind kann ein Junction-Tree konstruiert werden der für alle Zeitscheiben* $t > 0$ *identische Sub-Trees aufweist.*

Beweis 12 *Seien* $\mathbf{C_t^2} \ldots \mathbf{C_t^N}$ *Cliquen, die nur Knoten aus t enthalten, und* $\mathbf{C_t^1}$ *die einzige Clique mit Knoten aus t und* $t-1$*. Die Clique* $\mathbf{C_t^1}$ *existiert da alle Knoten mit Verbindungen von* $t-1$ *nach t benachbart sind. Dann ist diese Clique* $\mathbf{C_t^1}$ *auch die einzige mit einem nicht leeren Separator zu einer Clique* $\mathbf{C_{t-1}^i}$*. Alle anderen Cliquen* $\mathbf{C_t^2} \ldots \mathbf{C_t^N}$ *sind daher mit* $\mathbf{C_t^1}$ *über Pfade verbunden, die nur Cliquen aus* $\{\mathbf{C_t^2} \ldots \mathbf{C_t^N}\}$ *selbst sowie die Clique* $\mathbf{C_{t+1}^1}$ *enthalten. Da alle Zeitscheiben dieselbe Struktur aufweisen ist es offensichtlich, dass analoge Cliquen für alle Zeitscheiben definiert und in der selben Struktur untereinander verbunden werden können. Im Gesamt-Junction-Tree sind je zwei solcher Teil-Bäume durch den dynamischen Separator zwischen* $\mathbf{C_t^1}$ *und* $\mathbf{C_{t-1}^i}$ *verbunden.*

Anmerkung: Die Umkehrung gilt nicht. Es ist nicht in jedem Fall notwendig alle Knoten aus zwei benachbarten Zeitscheiben $t-1$ und t die Verbindungen zueinander haben in einer Clique zu vereinen, um einen replizierbaren Sub-Junction-Tree zu erhalten.



Wird aus diesem Graph (der moralische Graph ist trianguliert wegen der moralischen Kante $B(t-1) - A(t)$) normal (ohne zusätliche Fill-In-Kanten) ein Junction Tree erzeugt entsteht jeweils identische Sub-Junction-Trees mit den Cliquen $\mathbf{C^1(t)} = \{A(t-1), B(t-1), A(t)\}$ und $\mathbf{C^2(t)} = \{B(t-1), A(t), B(t)\}$, die über den Separator $\{B(t-1), A(t)\}$ verbunden sind. Zusätzlich ist $\mathbf{C^1(t)}$ mit $\mathbf{C^2(t-1)}$ über den Separator $\{A(t-1), B(t-1)\}$ verbunden. Weder $\mathbf{C^1(t)}$ noch $\mathbf{C^2(t)}$ enthält alle Knoten mit Verbindungen zwischen t und $t-1$ ($A(t-1)$, $B(t-1)$, $A(t)$, $B(t)$).

Satz 28 *Um eine Abhängigkeit gegeben die aktuelle Evidenz zwischen zwei Knoten X und Y zu erlauben, muss es einen Pfad von X nach Y geben auf dem kein serielles oder divergierendes Knoten-Triplet* $A - B - C$ *"geblockt" ist (geblockt = harte Evidenz bezüglich B) und auf dem alle konvergierenden Knoten-Triplets Information (harte oder weiche Evidenz) bezüglich B oder eines Nachfolgers von B enthalten.*

Definition 44 *Gibt es keinen solchen Pfad sind X und Y* **d-separiert**.

Beweis 13 (Beweisskizze) *Die d-separierten Fälle lassen sich jeweils relativ einfach zeigen.*
Serieller Verbindungstyp: Ⓐ → Ⓑ → Ⓒ *(d-separiert wenn B gegeben).*
Zu zeigen ist dass:

$$P(A, C|B) \quad = \quad P(A|B) \bullet P(C|B)$$

Über die Kettenregel und den Satz von der Bedingten Wahrscheinlichkeit erhalten wir:

$$\begin{aligned} P(A, C|B) &= \frac{P(A) \bullet P(B|A) \bullet P(C|B)}{\sum_{A,C} (P(A) \bullet P(B|A) \bullet P(C|B))} \quad \left(= \frac{P(A, B, C)}{\sum_{A,C} P(A, B, C)} \right) \\ &= \frac{P(A, B) \bullet P(C|B)}{P(B)} \\ &= \frac{P(A, B)}{P(B)} \bullet P(C|B) \\ &= P(A|B) \bullet P(C|B) \end{aligned}$$

Divergierender Verbindungstyp: Ⓐ ← Ⓑ → Ⓒ *(Unabhängigkeit wenn B gegeben).*
Zu zeigen ist wiederum dass:

$$P(A,C|B) \quad = \quad P(A|B) \bullet P(C|B)$$

Äquivalent zum vorigen Fall ergibt sich:

$$\begin{aligned} P(A,C|B) &= \frac{P(B) \bullet P(A|B) \bullet P(C|B)}{\sum_{A,C} (P(B) \bullet P(A|B) \bullet P(C|B))} \quad \left(= \frac{P(A,B,C)}{\sum_{A,C} P(A,B,C)} \right) \\ &= \frac{P(B)}{P(B)} \bullet P(A|B) \bullet P(C|B) \\ &= 1 \bullet P(A|B) \bullet P(C|B) \end{aligned}$$

Konvergierender Verbindungstyp: Ⓐ → Ⓑ ← Ⓒ *(Unabhängigkeit wenn keine Information über B vorliegt).*
Zu zeigen ist damit:

$$P(A,C) \quad = \quad P(A) \bullet P(C)$$

Laut Kettenregel gilt:

$$\begin{aligned} P(A,C) &= \sum_B (P(A) \bullet P(C) \bullet P(B|A,C)) \\ &= P(A) \bullet P(C) \bullet \sum_B P(B|A,C) \\ &= P(A) \bullet P(C) \bullet 1 \end{aligned}$$

Die Fällen in denen Knoten d-connected sind können durch Beispiele gezeigt werden.
Für den seriellen und divergierenden Verbindungstyp gilt:
Wenn B unbekannt / unsicher ist sind A und C im Allgemeinen nicht unabhängig (natürlich können die Tafeln im Einzelfall auch so gewählt sein dass doch eine Unabhängigkeit vorliegt).
Zu zeigen ist also dass bei entsprechender Wahl der Tafeln gilt:

$$\begin{aligned} P(A,C) &\neq P(A) \bullet P(C) \\ \sum_B P(A,B,C) &\neq \sum_{B,C} P(A,B,C) \bullet \sum_{A,B} P(A,B,C) \\ \sum_B P(A,B,C) \bullet F(B) &\not\propto \sum_{B,C} (P(A,B,C) \bullet F(B)) \bullet \sum_{A,B} (P(A,B,C) \bullet F(B)) \end{aligned}$$

Wobei $F(B)$ einen beliebigen Finding-Vektor mit mehr als einem Wert $\neq 0$ darstellt. Hierbei sind i.d.R. die Ergebnisse der linken und rechten Seite mit unterschiedlichen Faktoren zu renormieren. Das heisst die Resultierenden Tafeln wären gleich wenn linke und rechte Seite (vor Normierung) proportional sind. Annahme: A und B wären doch unabhängig. Gegenbeispiel: Gegeben sei folgende Tafel

$P(A,C) = \sum_B P(A,B,C)$ =

		b_1	+	b_2
c_1	a_1	s	+	t
c_1	a_2	u	+	v
c_2	a_1	w	+	x
c_2	a_2	y	+	z
	$\sum_{A,C}$		1	

Es müsste nun u.A. gelten: $P(a_1, b_1) = s + t = (s + t + w + x)(s + t + u + v)$ *Gegeben sei folgende Setzung für den seriellen Fall:*

$$\begin{array}{rcllllllllll}
s & = & P(a_1) & P(b_1|a_1) & P(c_1|b_1) & := & 0.3 & \bullet & 0.8 & \bullet & 0.1 & = & 0.024 \\
t & = & P(a_1) & P(b_2|a_1) & P(c_1|b_2) & := & 0.3 & \bullet & 0.2 & \bullet & 0.5 & = & 0.03 \\
u & = & P(a_2) & P(b_1|a_2) & P(c_1|b_1) & := & 0.7 & \bullet & 0.4 & \bullet & 0.1 & = & 0.028 \\
v & = & P(a_2) & P(b_2|a_2) & P(c_1|b_2) & := & 0.7 & \bullet & 0.6 & \bullet & 0.5 & = & 0.21 \\
w & = & P(a_1) & P(b_1|a_1) & P(c_2|b_1) & := & 0.3 & \bullet & 0.8 & \bullet & 0.9 & = & 0.216 \\
x & = & P(a_1) & P(b_2|a_1) & P(c_2|b_2) & := & 0.3 & \bullet & 0.2 & \bullet & 0.5 & = & 0.03
\end{array}$$

Damit ist aber $s + t = 0.054 \neq (s + t + w + x)(s + t + u + v) = 0.3 \bullet 0.292 = 0.0876$ *und damit die Annahme für die serielle Verbindung widerlegt.*

Für den divergierenden Fall gelte folgende Setzung:

$$\begin{array}{rcllllllllll}
s & = & P(a_1|b_1) & P(b_1) & P(c_1|b_1) & := & 0.8 & \bullet & 0.9 & \bullet & 0.3 & = & 0.216 \\
t & = & P(a_1|b_2) & P(b_2) & P(c_1|b_2) & := & 0.4 & \bullet & 0.1 & \bullet & 0.5 & = & 0.02 \\
u & = & P(a_2|b_1) & P(b_1) & P(c_1|b_1) & := & 0.2 & \bullet & 0.9 & \bullet & 0.3 & = & 0.054 \\
v & = & P(a_2|b_2) & P(b_2) & P(c_1|b_2) & := & 0.6 & \bullet & 0.1 & \bullet & 0.5 & = & 0.03 \\
w & = & P(a_1|b_1) & P(b_1) & P(c_2|b_1) & := & 0.8 & \bullet & 0.9 & \bullet & 0.7 & = & 0.504 \\
x & = & P(a_1|b_2) & P(b_2) & P(c_2|b_2) & := & 0.4 & \bullet & 0.1 & \bullet & 0.5 & = & 0.02
\end{array}$$

Damit ist aber $s + t = 0.236 \neq (s + t + w + x)(s + t + u + v) = 0.76 \bullet 0.32 = 0.2432$ *und damit die Annahme auch für die divergierende Verbindung widerlegt.*

Für den konvergierenden Verbindungstyp gilt, dass die Unabhängigkeitsbedingung nicht mehr erfüllt ist, wenn B *gegeben ist.*

$$\begin{array}{rcl}
\displaystyle\sum_B \frac{P(A,B,C)}{P(B)} = P(A,C|B) & \neq & \displaystyle\sum_C P(A,C|B) \bullet \sum_A P(A,C|B) \\
\displaystyle\sum_B (P(A,B,C) \bullet F(B)) & \not\propto & \displaystyle\sum_{BC} P(A,B,C) \bullet F(B) \bullet \\
 & & \displaystyle\sum_{BA} P(A,B,C) \bullet F(B)
\end{array}$$

$$P(A,C|B) \neq \left(\sum_C P(A,C|B)\right) \bullet \left(\sum_A P(A,C|B)\right)$$

		b_1	b_2
c_1	a_1	s	t
c_1	a_2	u	v
c_2	a_1	w	x
c_2	a_2	y	z
	$\sum_{A,C}$	1	1

$\neq$

		b_1	b_2
c_1	a_1	$(s+u)(s+w)$	$(t+v)(t+x)$
c_1	a_2	$(s+u)(u+v)$	$(t+v)(v+z)$
c_2	a_1	$(w+y)(s+w)$	$(x+z)(t+x)$
c_2	a_2	$(w+y)(u+v)$	$(x+z)(v+z)$
	$\sum_{A,C}$	1	1

Setzt man konkrete Werte für die Zelle $A = a_1$, $B = b_1$ *und* $C = c_1$ *ein, dann ergibt sich z.B.:*

$$\begin{array}{rcl}
s & = & k/c \\
u & = & l/c
\end{array}$$

$$
\begin{array}{rcl}
w & = & m/c \\
k & = & P(c_1)P(a_1)P(b_1|a_1,c_1) := 0.7 \cdot 0.1 \cdot 0.4 = 0.028 \\
l & = & P(c_1)P(a_2)P(b_1|a_2,c_1) := 0.7 \cdot 0.9 \cdot 0.8 = 0.504 \\
m & = & P(c_2)P(a_1)P(b_1|a_1,c_2) := 0.3 \cdot 0.1 \cdot 0.3 = 0.009 \\
n & = & P(c_2)P(a_2)P(b_1|a_2,c_2) := 0.3 \cdot 0.9 \cdot 0.5 = 0.135 \\
c & = & k + l + m + n = 0.676
\end{array}
$$

$$
\begin{array}{rcccl}
i) & & s & = & 0.028/0.676 \approx 0.041420 \\
ii) & & (s+u)(s+w) & = & 0.532/0.676 \cdot 0.037/0.676 \approx 0.0431 \\
i) \wedge ii) & \Rightarrow & (s+u)(s+w) & \neq & s
\end{array}
$$