



Intelligenza Artificiale

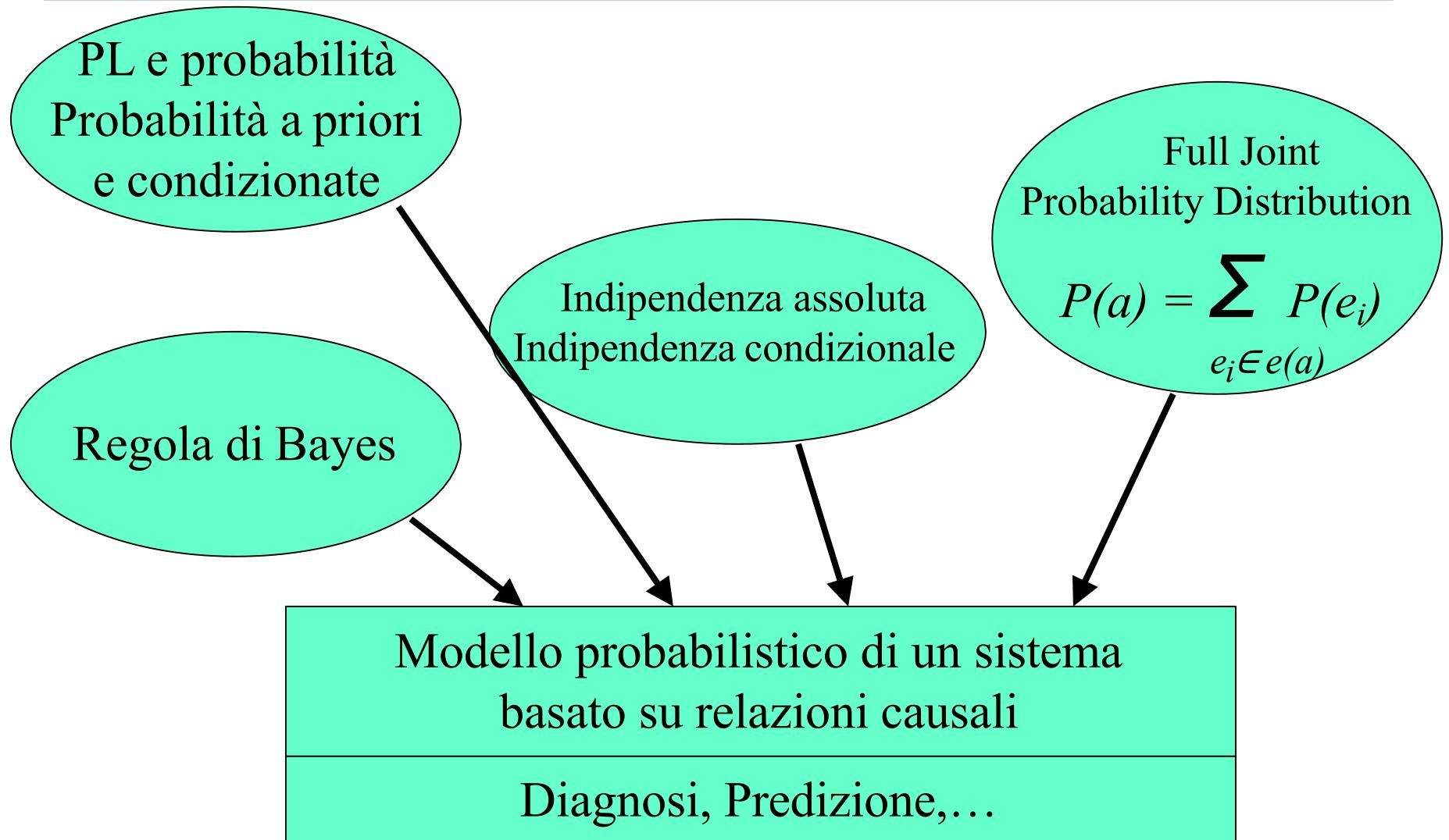
Ragionamento Probabilistico



Outline

- ▶ Reti bayesiane - sintassi
- ▶ Semantica delle reti bayesiane
- ▶ Costruzione di una rete
- ▶ Inferenza esatta su reti bayesiane

Reti Bayesiane



Reti Bayesiane

- ▶ Basic laws: $0 \leq P(\omega) \leq 1$, $\sum_{\omega \in \Omega} P(\omega) = 1$, $P(A) = \sum_{\omega \in A} P(\omega)$
- ▶ Random variable $X(\omega)$ has a value in each ω
 - ▶ Distribution $P(X)$ gives probability for each possible value x
 - ▶ Joint distribution $P(X,Y)$ gives total probability for each combination x,y
- ▶ Summing out/marginalization: $P(X=x) = \sum_y P(X=x, Y=y)$
- ▶ Conditional probability: $P(X|Y) = P(X,Y)/P(Y)$
- ▶ Chain rule: $P(X_1, \dots, X_n) = \prod_i P(X_i | X_1, \dots, X_{i-1})$
- ▶ Bayes Rule: $P(X|Y) = P(Y|X)P(X)/P(Y)$
- ▶ Independence: $P(X,Y) = P(X) P(Y)$ or $P(X|Y) = P(X)$ or $P(Y|X) = P(Y)$
- ▶ Conditional Independence: $P(X|Y,Z) = P(X|Z)$ or $P(X,Y|Z) = P(X|Z) P(Y|Z)$

Reti Bayesiane

$P(MalDiDenti, Prende, Carie) =$

$P(MalDiDenti | Carie) P(Prende | Carie) P(Carie)$

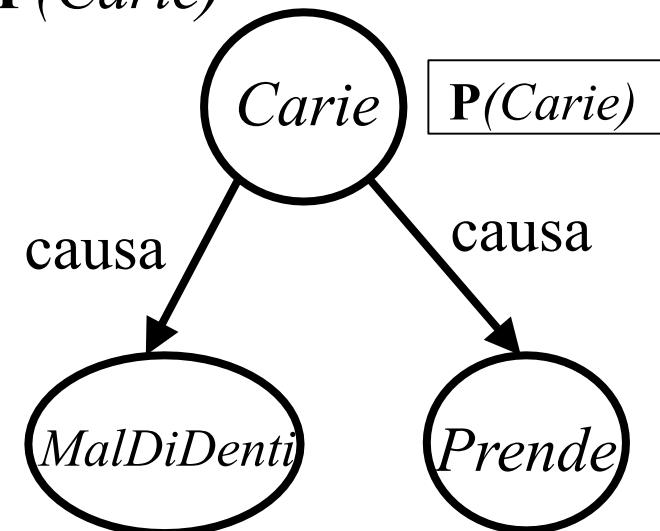
$P(MalDiDenti, Prende, Carie)$

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	.108	.012	.072	.008
\neg <i>cavity</i>	.016	.064	.144	.576

Bayes

$P(Cause|Effect) =$

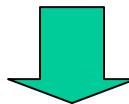
$$\frac{P(Effect|Cause) P(Cause)}{P(Effect)}$$



$P(MalDiDenti | Prende, Carie) = P(MalDiDenti | Carie)$
relazione di indipendenza condizionata

Reti Bayesiane

- ▶ Distribuzione Congiunta di Probabilità (Full Joint Distribution)
 - ▶ Risponde (con crescita esponenziale) a qualunque query probabilistica
 - ▶ Grande quantità di dati per ricavare le stime di ogni evento atomico
 - ▶ Modo innaturale di acquisire la conoscenza
 - ▶ L'indipendenza assoluta e condizionale riduce il numero di stime di probabilità richieste



- ▶ Reti bayesiane
(belief networks, probabilistic networks, causal networks)
 - ▶ Una notazione grafica semplice per le asserzioni di indipendenza condizionale e quindi per la specificazione compatta di una full joint distribution

Reti Bayesiane: Sintassi

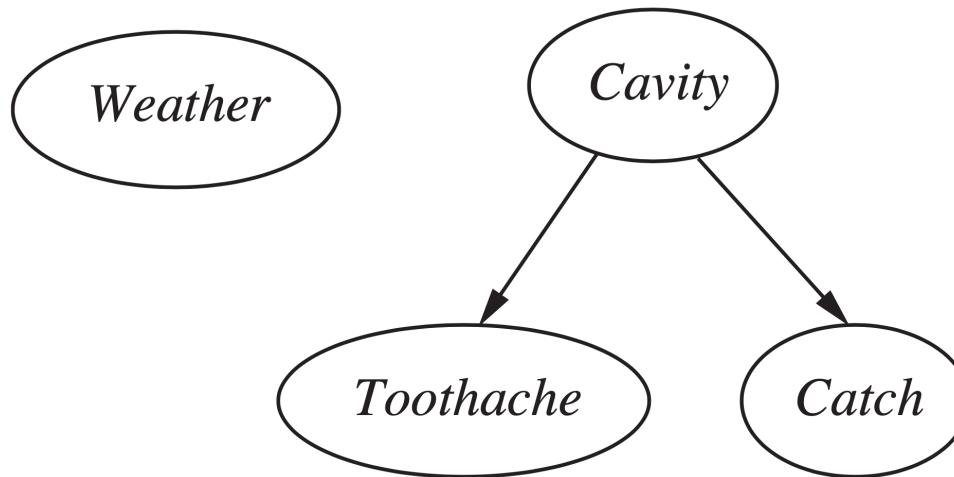
- ▶ Grafo orientato aciclico annotato con distribuzioni di probabilità condizionate
 - ▶ Un insieme di nodi, uno per variabile causale
 - ▶ Un insieme di archi orientati che connettono coppie di nodi. Se c'è un arco dal nodo X al nodo Y si dice che X è *parent* di Y (X ha un'influenza diretta su Y)
 - ▶ Il grafo non ha cicli
 - ▶ Ad ogni nodo X_i è associata una distribuzione di probabilità condizionale che quantifica l'effetto dei *parents* sul nodo

$$P(X_i \mid \text{Parents}(X_i))$$

- ▶ Per variabili discrete, la distribuzione condizionale è rappresentata come una tabella, (CPT Conditional Probability Table) che fornisce la distribuzione su X_i per ogni combinazione dei valori dei nodi *parents* (in direzione causale)

Reti Bayesiane

- ▶ La topologia della rete codifica le asserzioni di indipendenza condizionale



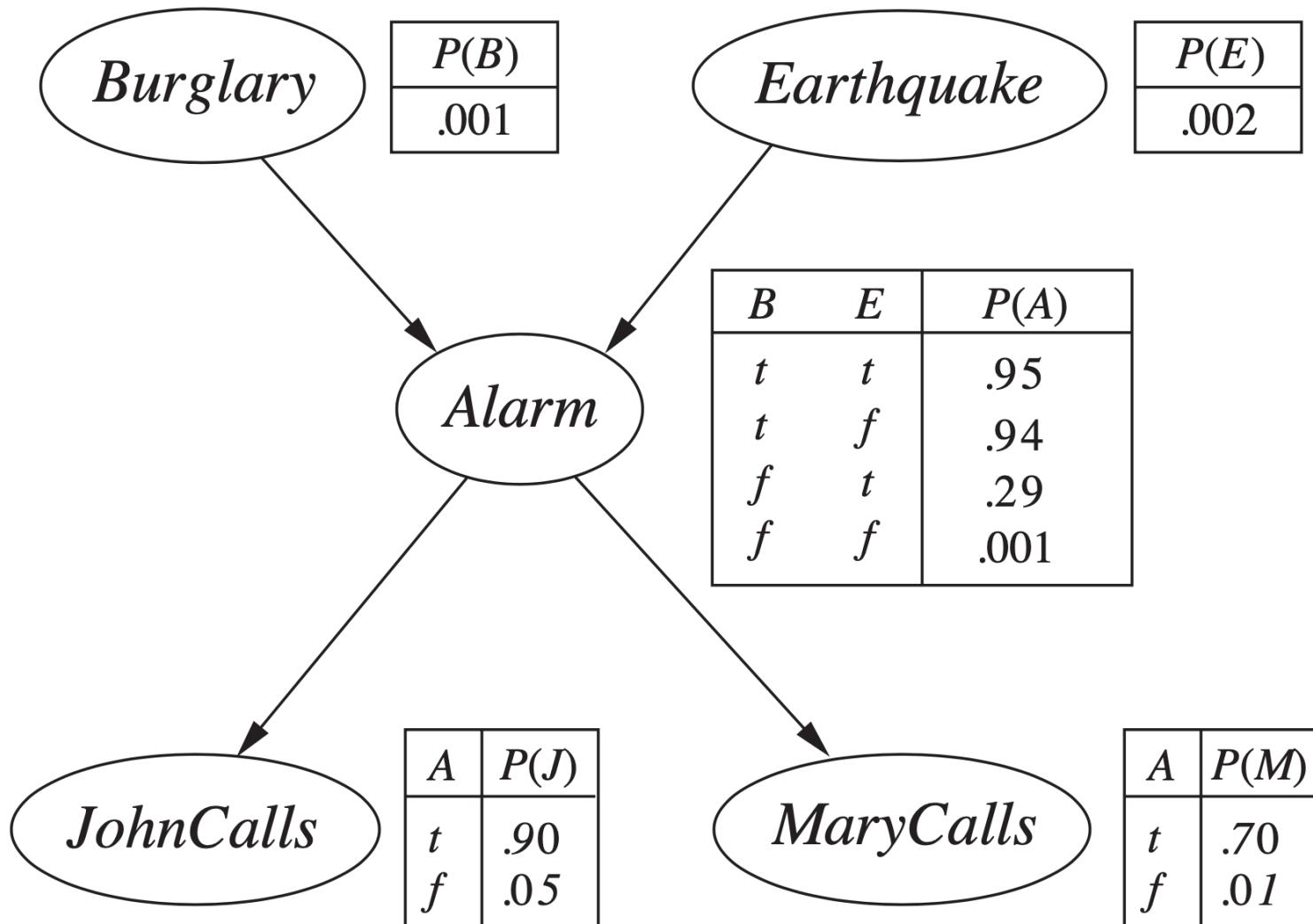
- ▶ *Meteo* è indipendente dalle altre variabili
- ▶ *MalDiDenti* e *Prende* sono condizionalmente indipendenti data *Carie* (ognuna dipende da *Carie* ma non c'è relazione causale tra le due)

Reti Bayesiane

Esempio

Now consider the following example, which is just a little more complex. You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes. (This example is due to Judea Pearl, a resident of Los Angeles—hence the acute interest in earthquakes.) You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm. John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too. Mary, on the other hand, likes rather loud music and sometimes misses the alarm altogether. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Reti Bayesiane



Reti Bayesiane

- ▶ Variabili:

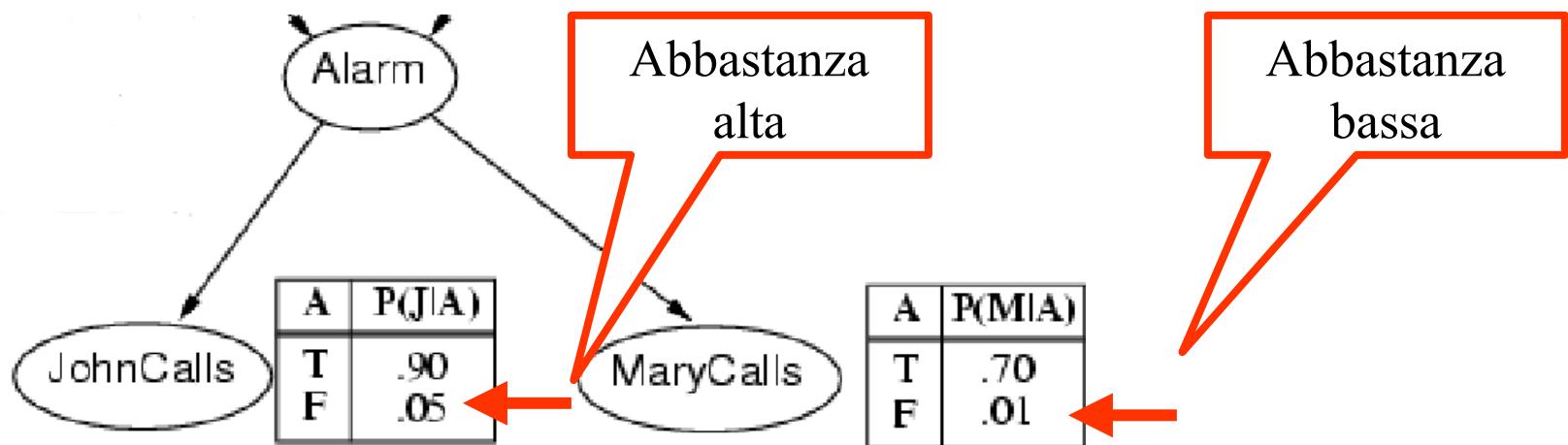
Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

Non è un fatto di sintassi, ma definisce quanto il modello è un buon modello della realtà (le relazioni di indipendenza condizionale codificate nel modello sono una sufficiente, per gli scopi dati, approssimazione della realtà)

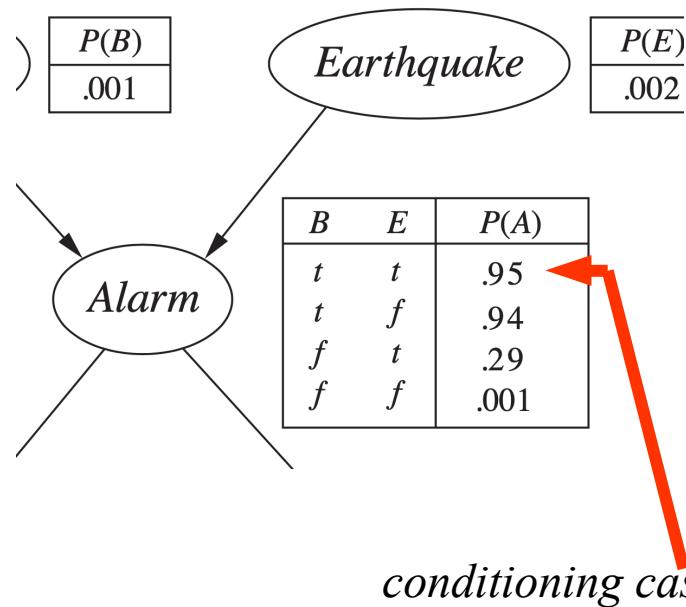
- ▶ La topologia della rete riflette la conoscenza causale
 - ▶ Il furto con scasso ed il terremoto possono attivare l'allarme
 - ▶ Il fatto che Mary o John possano chiamare dipende solo dall'allarme
 - ▶ Mary e John non sono attivati dal furto o dal terremoto e non interagiscono tra di loro

Reti Bayesiane

- ▶ Rappresentazione della conoscenza incerta
- ▶ La rete non modella il fatto che Mary ascolta musica ad alto volume e che John confonde il suono del telefono con quello dell'allarme
- ▶ Tutti questi fattori (e altri potenzialmente infiniti) sono riassunti dall'incertezza associata ai link tra *alarm*, *Mary* e *John*



Reti Bayesiane



CPT Conditional Probability Table (per variabili discrete)

Ogni riga contiene, per ogni valore del nodo, la probabilità condizionale per un *conditioning case*
(una possibile combinazione di valori dei nodi *parents*)

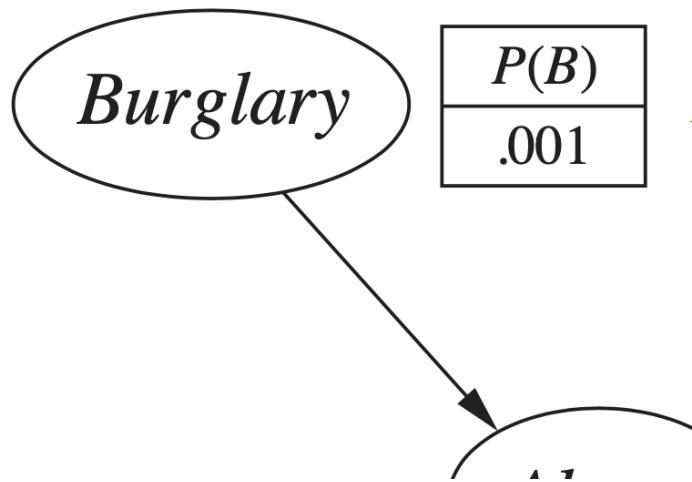
B	E	$P(A B,E)$	$P(\neg A B,E)$
T	T	.95	.05

Ogni riga ha somma 1
delle probabilità condizionate

↑
p

↑
Omesso. Per variabile booleana è 1-p

Reti Bayesiane



Un nodo senza *parents*
ha una sola riga che rappresenta la
probabilità a priori di ogni possibile
valore della variabile

$P(B)$	$P(\neg B)$
.001	.999

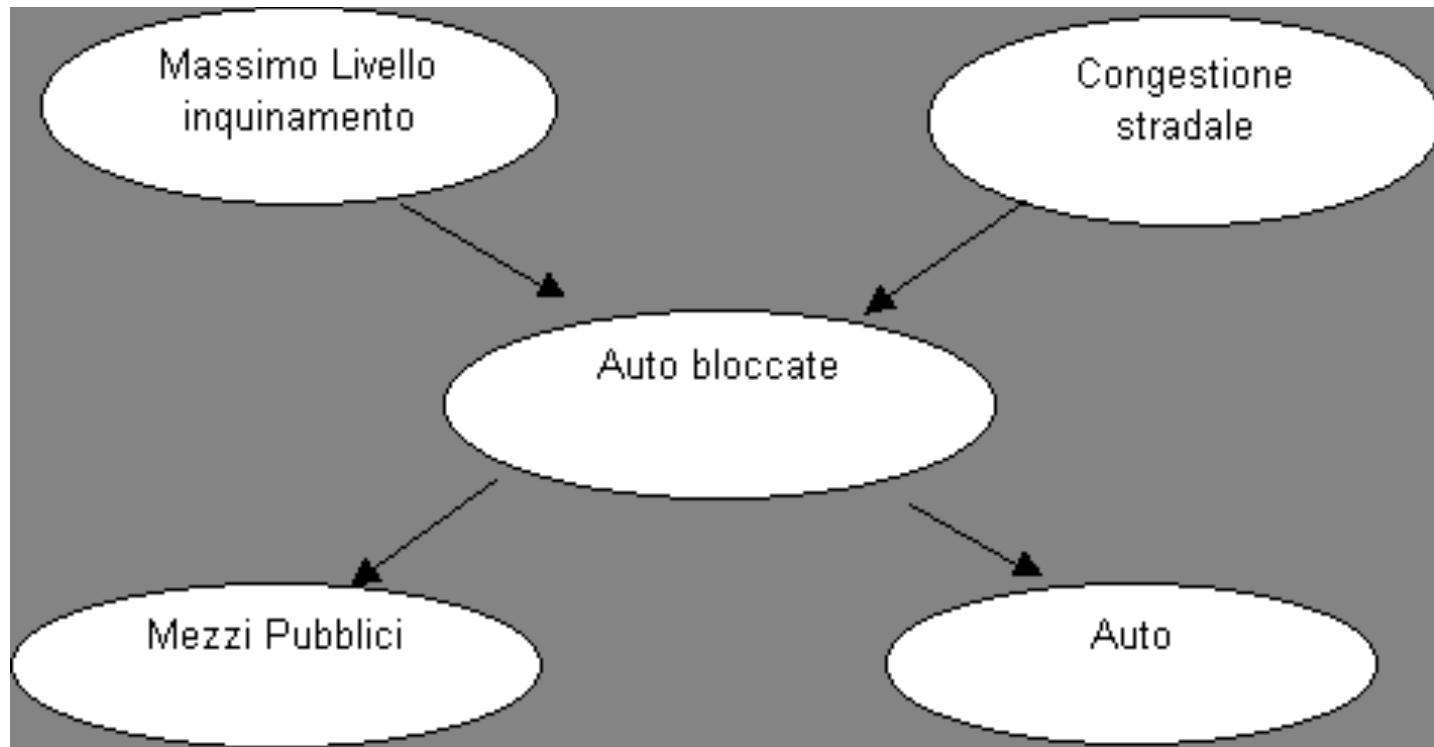
Esempio 1: Gestione del traffico

Un Comune può decidere se bloccare o no le auto per una giornata nel caso in cui si verifichi uno dei seguenti casi:

- ▶ viene raggiunto il livello massimo di inquinamento,
- ▶ si verifica una congestione delle strade.

A seconda della situazione i cittadini dovranno decidere se spostarsi con i mezzi pubblici o prendere la macchina.

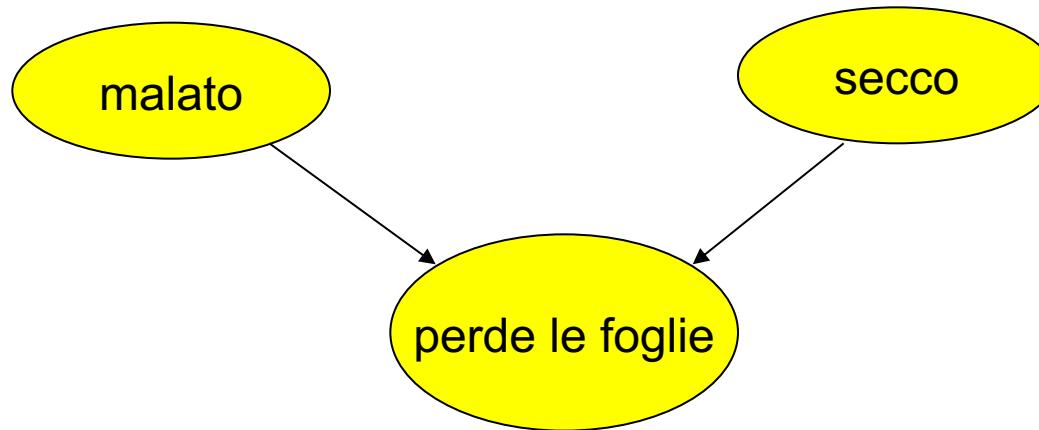
Esempio 1: Gestione del traffico



Esempio 2: L'albero di Jack

- ▶ Un giorno Jack si accorge che il suo albero di mele perde le foglie.
- ▶ Jack sa che se l'albero è secco allora è normale che perda le foglie.
- ▶ Ma Jack sa anche che la perdita delle foglie può essere sintomo di malattia per il suo albero.

La rete per l'albero di Jack



La rete consiste di 3 nodi:

- ▶ Malato, Secco, e Perde le foglie
- ▶ Malato può essere "malato" o "no"
- ▶ Secco può essere "secco" o "no"
- ▶ Perde le foglie può essere "si" o "no".

La rete per l'albero di Jack

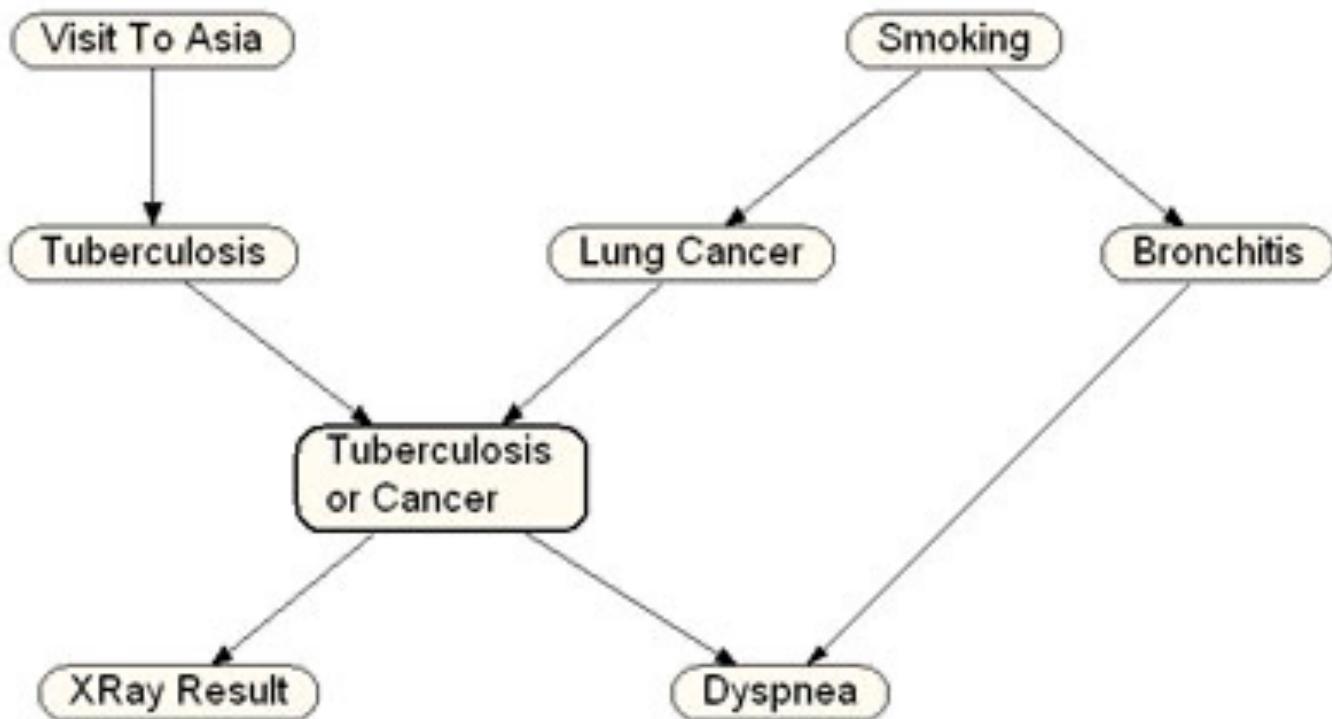
- ▶ La dipendenza casuale è tra *Malato* e *Perde le foglie* e *Secco* e *Perde le foglie*.
 - ▶ Ad ogni nodo è associata una tabella di probabilità, che possono essere a priori o condizionate. Ad esempio:
-
- ▶ $P(\text{Malato}=\text{"malato"})=0.1$
 - ▶ $P(\text{Malato}=\text{"no"})=0.9$
 - ▶ $P(\text{Secco}=\text{"secco"})=0.1$
 - ▶ $P(\text{Secco}=\text{"no"})=0.9$

La rete per l'albero di Jack

	Secco=“secco”		Secco=“No”	
	Malato= “Malato”	Malato= “No”	Malato= “Malato”	Malato= “No”
Perde=“si”	0.95	0.85	0.90	0.02
Perde=“no”	0.05	0.15	0.10	0.98

$P(\text{Perde le foglie} \mid \text{Malato, Secco})$

Diagnosi di una malattia polmonare



I dati

- ▶ Statisticamente, in un campione rappresentativo di popolazione si conoscono i seguenti dati.

- ▶ Il 50% dei pazienti fuma.
- ▶ Il 1% ha la tubercolosi.
- ▶ Il 5.5% ha un cancro al polmone (⌚).
- ▶ Il 45% ha una qualche forma di bronchite.

Come si procede

Si costruisce la rete bayesiana e si analizzano i sintomi mostrati dal paziente. Ad esempio, supponiamo che:

- ▶ Il paziente lamenta dispnea.
- ▶ Il paziente è stato recentemente in Asia.
- ▶ Il paziente è un fumatore.

Inoltre

- ▶ Si effettua una radiografia da cui si vede che:
- ▶ 1. Risultato negativo

Sembra che la diagnosi migliore sia che il paziente ha una semplice bronchite.

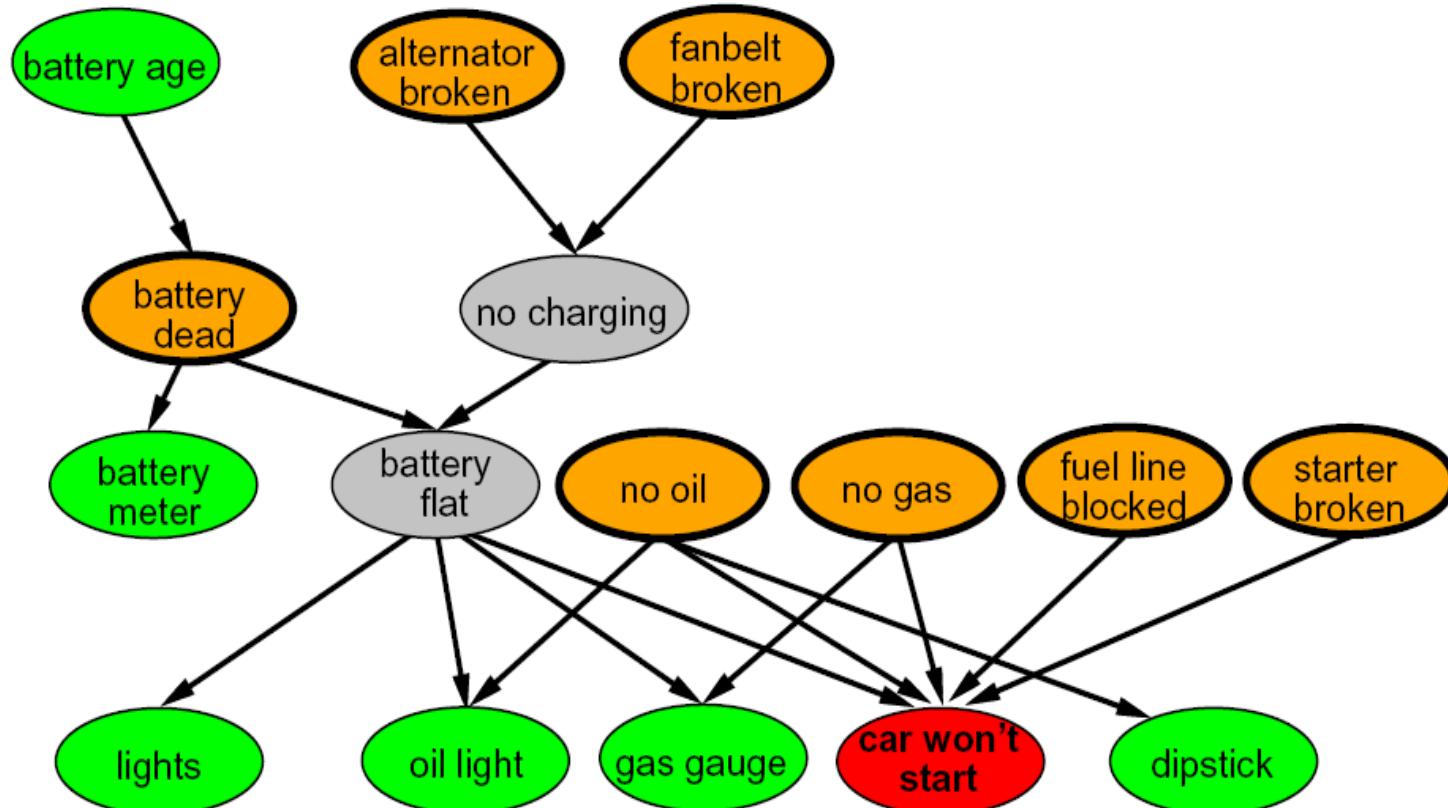
- ▶ 2. Risultato positivo
- La probabilità di un cancro o della tubercolosi sono aumentate enormemente. E' comunque necessario effettuare nuovi esami.

Reti Bayesiane

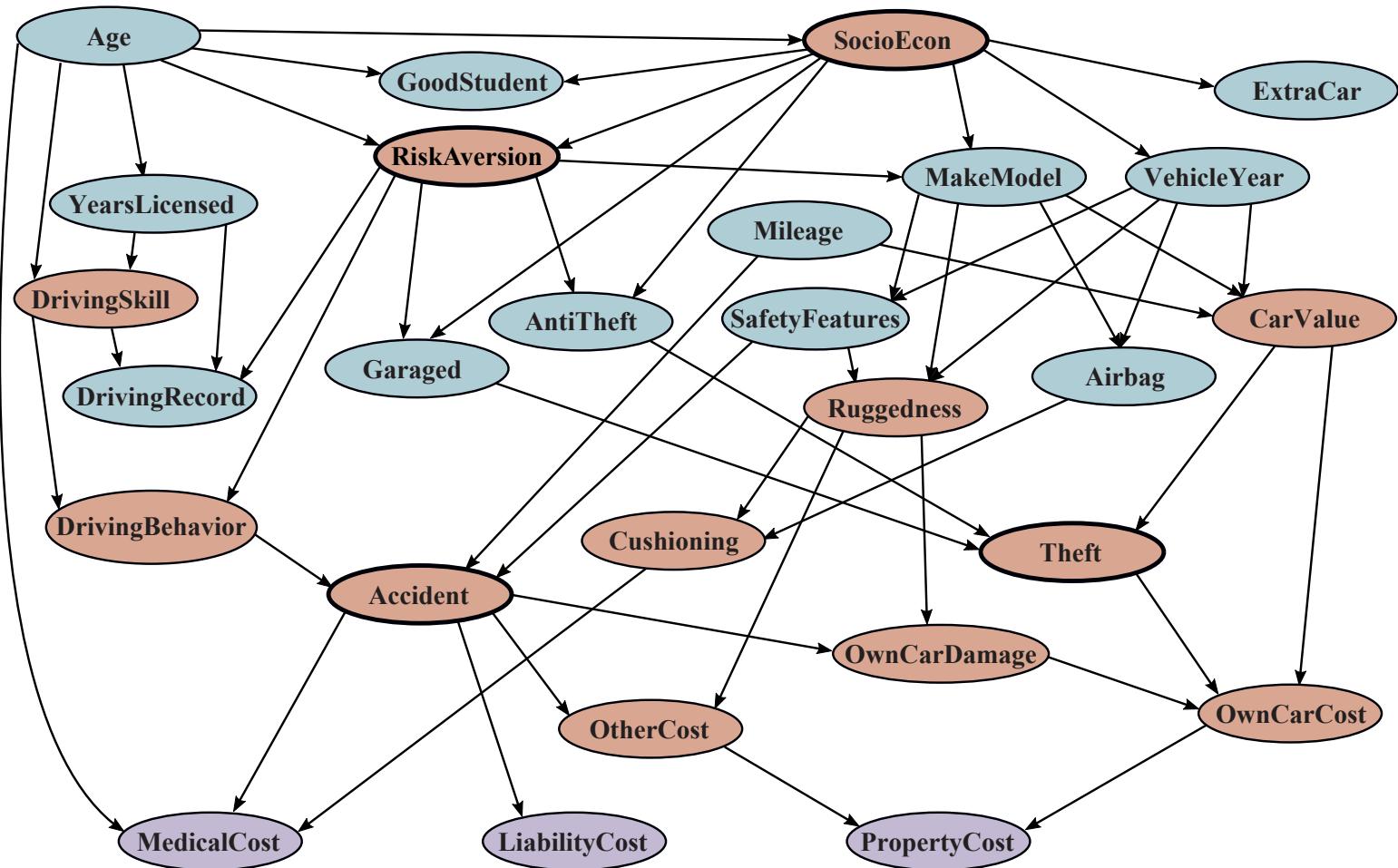
Evidenza iniziale: l'auto non parte

Variabili verificabili (verde), variabili "rotte, quindi aggiustabili" (arancione)

Le variabili nascoste (in grigio) assicurano una struttura sparsa, riducono i parametri



Reti Bayesiane



Assicurazione di automobile

Semantica Reti Bayesiane

- ▶ **Semantica:** una rete è una rappresentazione di una distribuzione di probabilità congiunta
- ▶ Uno specifico elemento di una distribuzione di probabilità congiunta (evento atomico) è definito come:

$$P(X_1=x_1, \wedge \dots \wedge X_n=x_n) \text{ abbreviato in } P(x_1, \dots, x_n)$$

- ▶ Il valore dell'elemento è :

$$P(x_1, \dots, x_n) = \prod_{i=1,n} P(x_i | Parents(X_i))$$

- ▶ Una distribuzione di probabilità congiunta è definita come il **prodotto** delle distribuzioni condizionali locali (CPT), date le asserzioni di indipendenza condizionale codificate dalla topologia della rete
- ▶ Le CPT sono la **decomposizione** di una distribuzione di probabilità congiunta

Probabilità dell'EVENTO ATOMICO

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(x_i | Parents(X_i))$$


La probabilità di x_i condizionata dai nodi genitori di x_i

- ▶ $Parents(X_i)$ denota i valori specifici delle variabili in X_i definiti in x_1, \dots, x_n (l'elemento corrispondente della CPT di x_i)

Esempio

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$

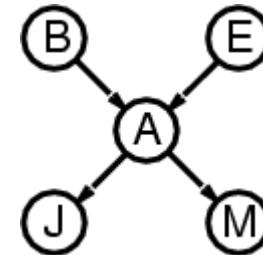
$$\begin{aligned} &P(x_i | Parents(X_i)) \\ &P(a | \neg b, \neg e) \end{aligned}$$

- Suonato allarme
- Non c'è stato furto e nemmeno terremoto
- Sia Mary che John hanno chiamato

Semantica Reti Bayesiane

► Esempio di evento atomico:

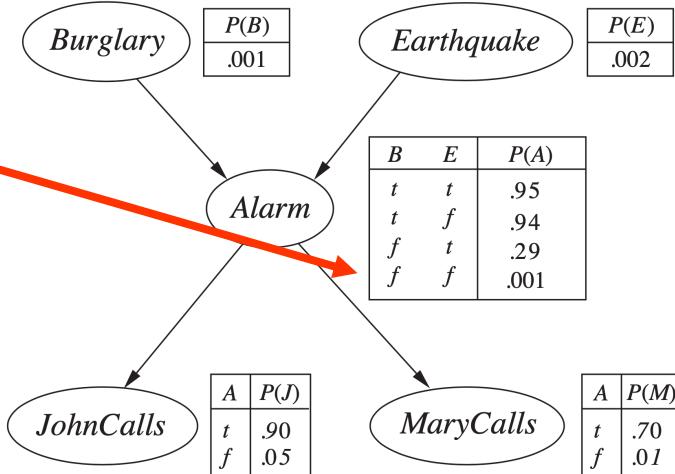
$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$



$$P(x_i | Parents(X_i))$$

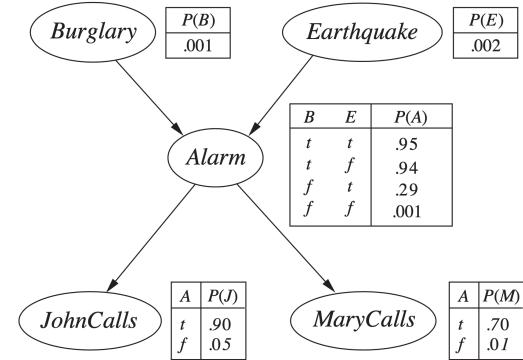
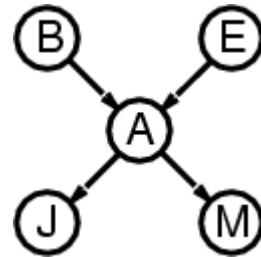
La probabilità di a condizionata dai nodi genitori di a è:

$$P(a | \neg b, \neg e)$$



Semantica Reti Bayesiane

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$



Applico iterativamente la product rule $P(a \wedge b) = P(a | b) P(b)$

.....

$$= P(j | m \wedge a \wedge \neg b \wedge \neg e) P(m | a \wedge \neg b \wedge \neg e) P(a | \neg b \wedge \neg e) P(\neg b | \neg e) P(\neg e)$$

Utilizzo le relazioni di indipendenza condizionale codificate nel grafo:

$$P(j | m \wedge a \wedge \neg b \wedge \neg e) = P(j | a) \quad j \text{ dipende solo da } a$$

.....

ed ottengo

$$= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e)$$

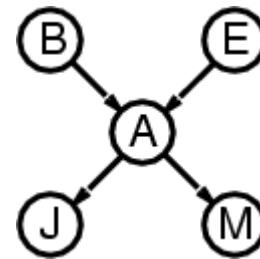
Ho applicato la formula:

$$P(x_1, \dots, x_n) = \prod_{i=1, n} P(x_i | \text{Parents}(X_i))$$

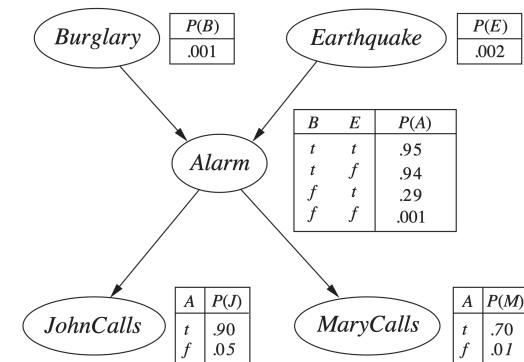
Semantica Reti Bayesiane

- Esempio-calcolo della probabilità dell'evento atomico:
 - Suonato allarme
 - Non c'è stato furto e nemmeno terremoto
 - Sia Mary che John hanno chiamato

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$



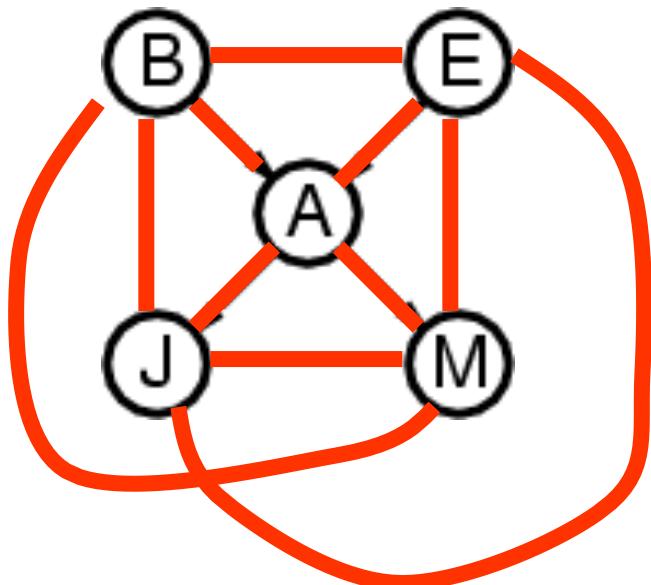
$$\begin{aligned} &= P(j | a) P(m | a) P(a | \neg b, \neg e) P(\neg b) P(\neg e) \\ &= 0,90 * 0,70 * 0,001 * 0,999 * 0,998 \\ &= 0,00062 \end{aligned}$$



Semantica Reti Bayesiane

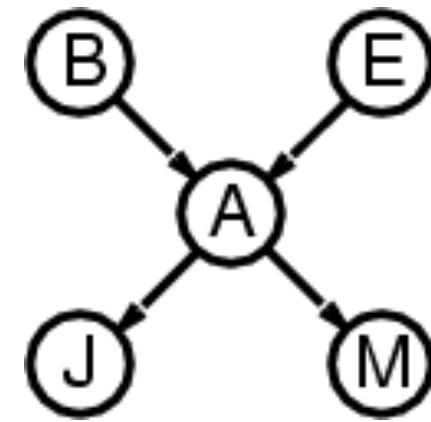
full joint distribution

Tavola di
32 (2^5) numeri



5 distribuzioni locali

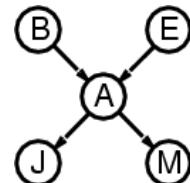
5 tavole
più piccole



Relazioni di indipendenza condizionale (archi causali)

Semantica Reti Bayesiane

- ▶ Compattezza
- ▶ La rete è una rappresentazione più compatta della distribuzione di probabilità condizionale
- ▶ Topologia + CPTs = rappresentazione compatta di una distribuzione di probabilità condizionale
- ▶ Se ogni variabile ha non più di k parents, la rete completa di n variabili (booleane) richiede $O(n \cdot 2^k)$ numeri
- ▶ A differenza della distribuzione di probabilità condizionale che cresce esponenzialmente: $O(2^n)$
- ▶ Per la rete di allarme precedente: $1 + 1 + 4 + 2 + 2 = 10$
- ▶ L'equivalente distribuzione di probabilità condizionale: $2^5 - 1 = 31$
- ▶ Se $n=30$ e $k=5$ abbiamo 960 numeri nella RB contro 1 miliardo



Costruzione di una rete

- ▶ Data la semantica della rete, come la si costruisce? Come si vede dall'esempio precedente:
- ▶ Applicando iterativamente la product rule si ottiene (*chain rule*):

$$P(X_1, \dots, X_n) = \prod_{i=1,n} P(X_i | X_{i-1}, \dots, X_1)$$

- ▶ Utilizzando l'equazione

$$P(X_1, \dots, X_n) = \prod_{i=1,n} P(X_i | \text{Parents}(X_i))$$

- ▶ Si ottiene

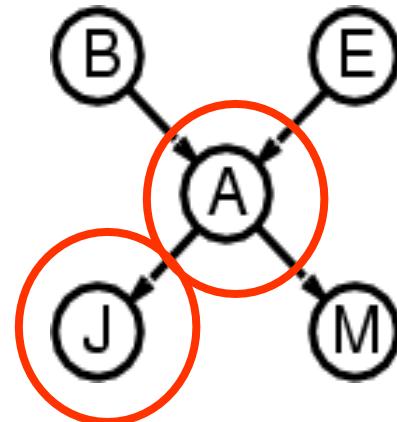
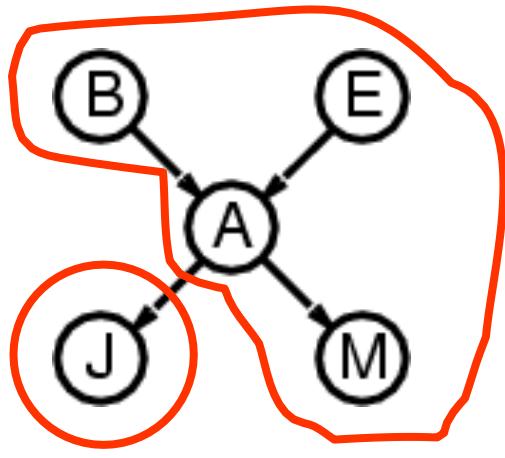
$$P(X_i | X_{i-1}, \dots, X_1) = P(X_i | \text{Parents}(X_i))$$

- ▶ La probabilità di X_i condizionata da tutti gli altri nodi è la probabilità di X_i condizionata dai nodi genitori

Costruzione di una rete

$$P(X_i | X_{i-1}, \dots, X_1) = P(X_i | \text{Parents}(X_i))$$

- ▶ La probabilità di X_i condizionata da tutti gli altri nodi è la probabilità di X_i condizionata dai nodi genitori



Esempio precedente:

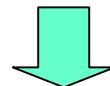
$$P(j | m \wedge a \wedge \neg b \wedge \neg e) = P(j | a)$$

j dipende solo da a

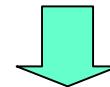
Costruzione di una rete

$$P(X_i | X_{i-1}, \dots, X_1) = P(X_i | \text{Parents}(X_i))$$

- ▶ Dato l'ordinamento dei nodi X_n, \dots, X_1 ,
- ▶ Per ogni nodo X_i , la probabilità condizionata del nodo rispetto a tutti gli altri nodi è la probabilità condizionata rispetto ai nodi genitori



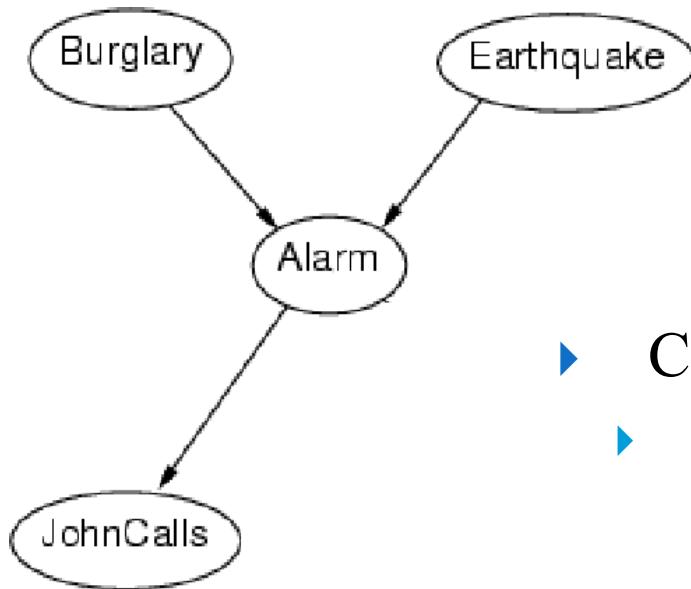
- ▶ Ogni nodo deve essere condizionalmente indipendente dai suoi predecessori nell'ordinamento dei nodi, dati i suoi nodi genitori



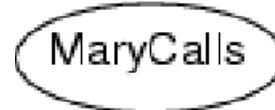
- ▶ I genitori di ogni nodo devono essere tutti e soli i nodi che influenzano direttamente il nodo

Costruzione di una rete

- I genitori di ogni nodo devono essere tutti e soli i nodi che influenzano direttamente il nodo



Devo aggiungere



- Conoscenza di dominio:
 - MaryCalls* può essere influenzata da *Burglary* o *Earthquake*, ma non direttamente
 - JohnCalls* non influenza *MaryCalls*
- Indipendenza condizionale:

$$\begin{aligned} P(\text{MaryCalls} \mid \text{JohnCalls}, \text{Alarm}, \text{Earthquake}, \text{Burglary}) &= \\ P(\text{MaryCalls} \mid \text{Alarm}) \end{aligned}$$

Costruzione di una rete

- ▶ Problema:
 - ▶ Rappresentare nella rete le relazioni di indipendenza condizionale in modo da avere una rappresentazione il più compatta possibile della distribuzione di probabilità congiunta
 - ▶ Rappresentare la conoscenza di dominio in modo il più possibile comprensibile

Costruzione di una rete

- ▶ L'ordine di scelta dei nodi nella costruzione porta a reti diverse
- ▶ L'ordine di scelta causale (dalle “cause prime” ai successivi effetti, causa di altri effetti,...) porta a costruire reti ottime

Costruzione di una rete

- ▶ Data la semantica della rete, come la si costruisce?
- ▶ Primo modo: Costruzione di una rete a partire dal significato “numerico”
 - ▶ Si sceglie un ordinamento dei nodi
 - ▶ Si scrive un nodo alla volta
 - ▶ Si verifica l’indipendenza condizionale del nuovo nodo dai precedenti
 - ▶ Conseguentemente si scrivono gli archi
$$P(X_i | X_{i-1}, \dots, X_1) = P(X_i | Parents(X_i))$$
 - ▶ La probabilità di X_i condizionata da tutti gli altri nodi è la probabilità di X_i condizionata dai nodi genitori

Costruzione di una rete

► Esempio

Ordine dei nodi: M, J, A, B, E

Inserisco il nodo M → non ha *parents*

Inserisco il nodo J

Ordine diagnostico

$$P(J | M) = P(J) ?$$

No, J non è condizionalmente indipendente da M

Se Mary chiama significa che probabilmente c'è stato un allarme e quindi è più probabile che anche John chiami

Sono portati a modellare relazioni tra effetti di eventuali cause comuni



Costruzione di una rete

▶ Esempio

Ordine dei nodi: M, J, A, B, E

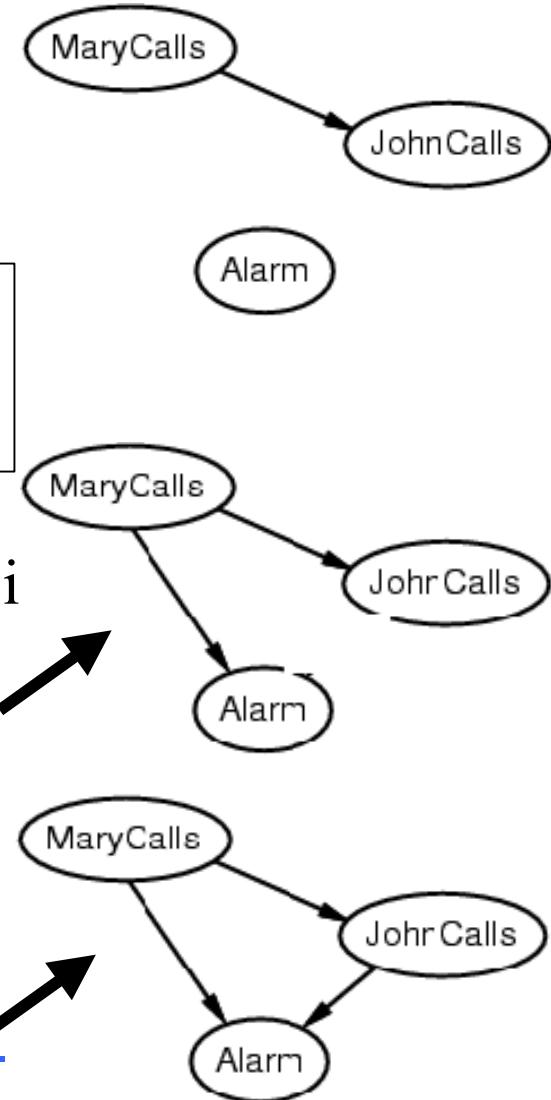
▶ $P(J | M) = P(J)$? **No**

A è condizionalmente
indipendente da M
dato J?

▶ Inserisco il nodo A

▶ $P(A | J, M) = P(A | J)$? **No** La probabilità che ci sia un allarme dipende dal fatto che Mary abbia chiamato

▶ $P(A | J, M) = P(A)$? **No** Se sia Mary che John chiamano, la probabilità che ci sia una allarme cambia



Costruzione di una rete

► Esempio

Ordine dei nodi: M, J, A, B, E

$P(J | M) = P(J)$? **No**

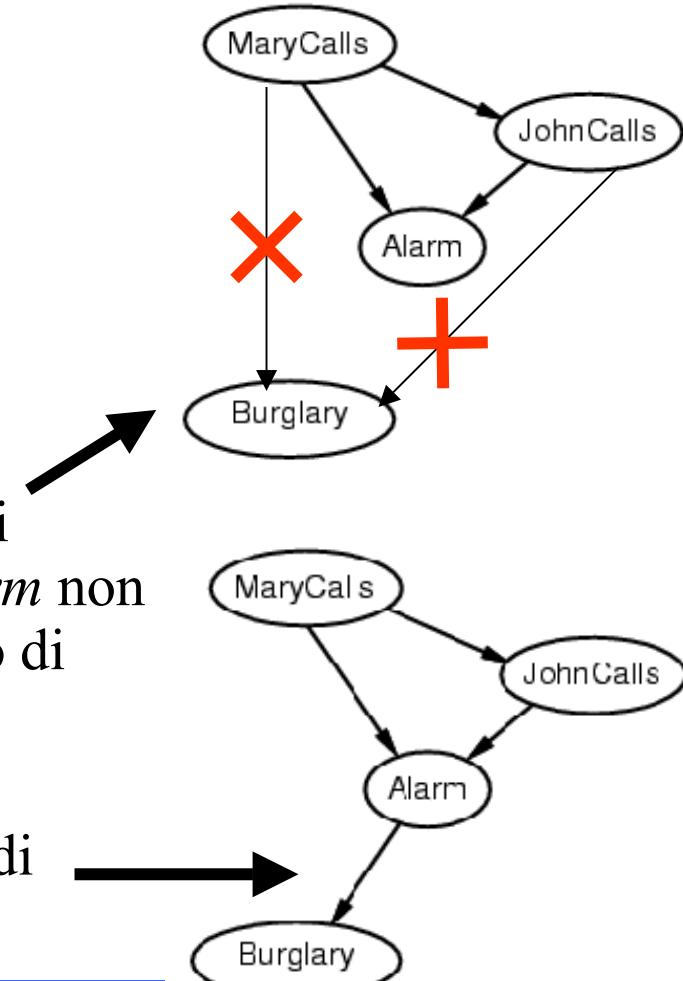
$P(A | J, M) = P(A | J)$? **No**

$P(A | J, M) = P(A)$? **No**

Inserisco il nodo B

$P(B | A, J, M) = P(B | A)$? **Yes** La probabilità di *Burglary* avendo come evidenza lo stato di *Alarm* non cambia se si aggiungono come evidenze lo stato di *MaryCalls* e *JohnCalls*

$P(B | A, J, M) = P(B)$? **No** La probabilità di *Burglary* cambia se si ha l'evidenza dello stato di *Alarm*



Costruzione di una rete

► Esempio

Ordine dei nodi: M, J, A, B, E

$$P(J | M) = P(J)? \text{No}$$

$$P(A | J, M) = P(A | J)? \text{No}$$

$$P(A | J, M) = P(A)? \text{No}$$

$$P(B | A, J, M) = P(B | A)? \text{Yes}$$

$$P(B | A, J, M) = P(B)? \text{No}$$

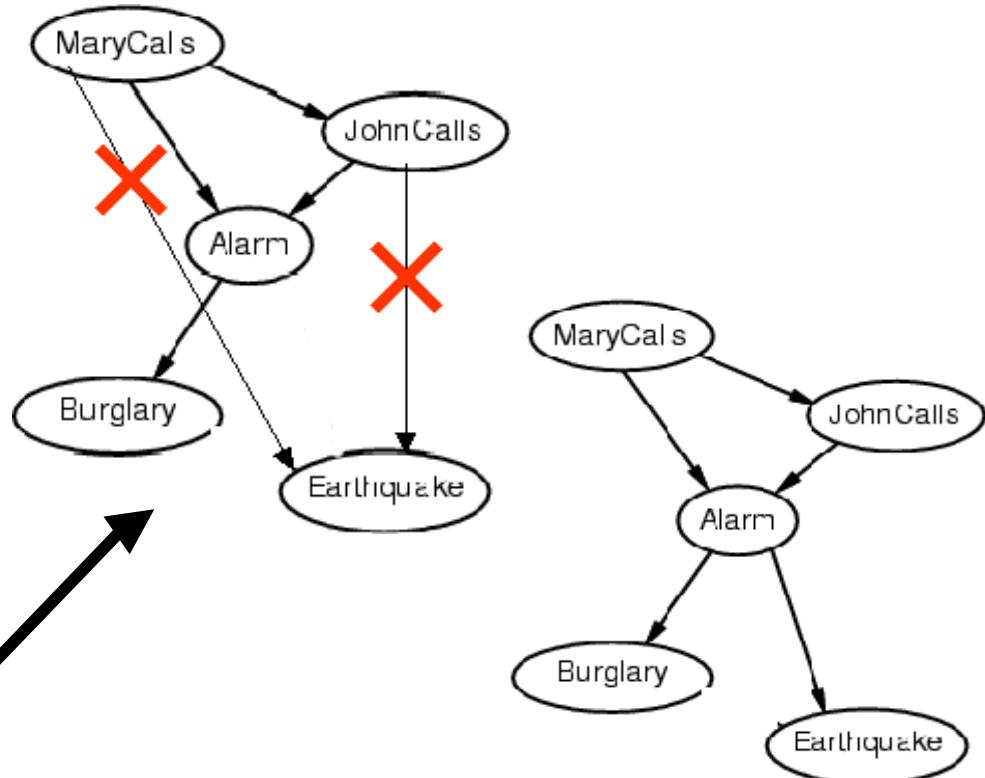
Inserisco il nodo E

$$P(E | B, A, J, M) = P(E | A, B)? \text{Yes}$$

Earthquake non cambia conoscendo lo stato delle chiamate di Mary e John

$$P(E | B, A, J, M) = P(E | B)? \text{No}$$

Se ho evidenza ci sia un *Alarm* la probabilità che ci sia un *Earthquake* cambia



Costruzione di una rete

► Esempio

Ordine dei nodi: M, J, A, B, E

$P(J | M) = P(J)$? **No**

$P(A | J, M) = P(A | J)$? **No**

$P(A | J, M) = P(A)$? **No**

$P(B | A, J, M) = P(B | A)$? **Yes**

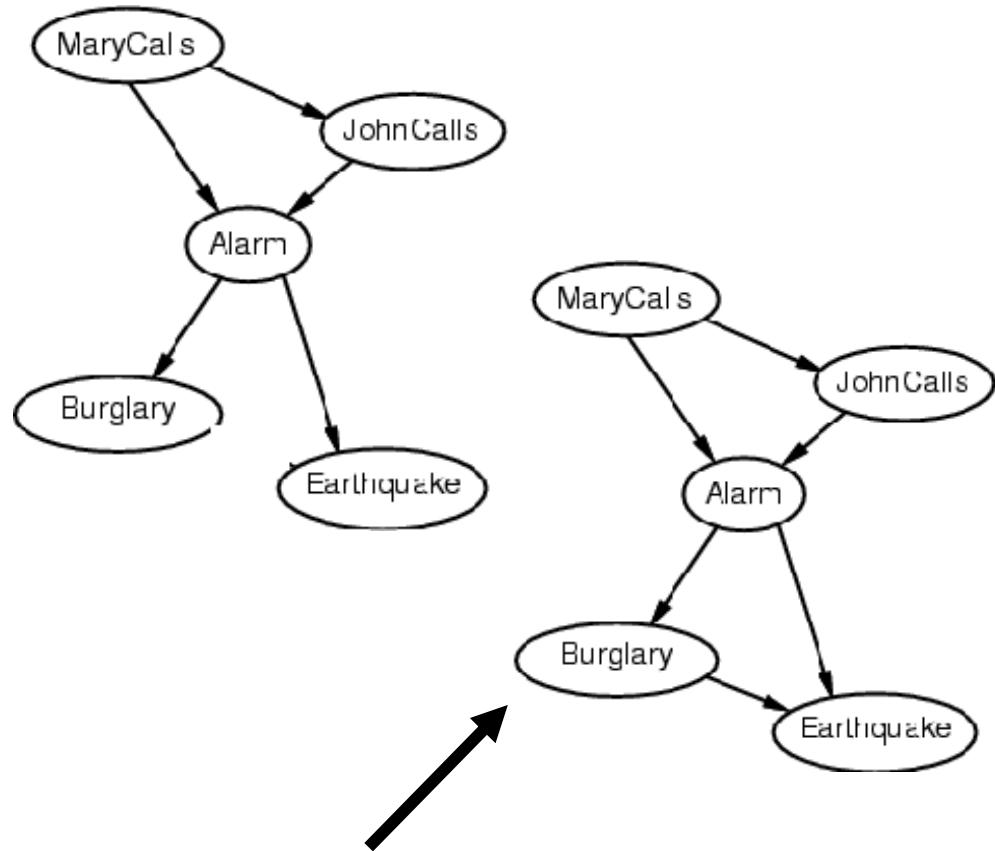
$P(B | A, J, M) = P(B)$? **No**

$P(E | B, A, J, M) = P(E | A)$? **Yes**

$P(E | B, A, J, M) = P(E | B)$? **No**

Inserisco il nodo E

► $P(E | B, A, J, M) = P(E | A)$? **No** Se so che c'è stato un *Burglary* in presenza di allarme, *Burglary* può spiegare l'allarme e quindi influenzare la probabilità di *Earthquake*



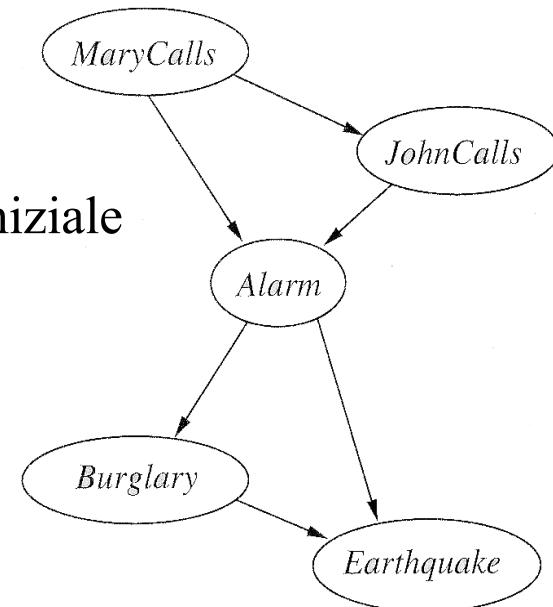
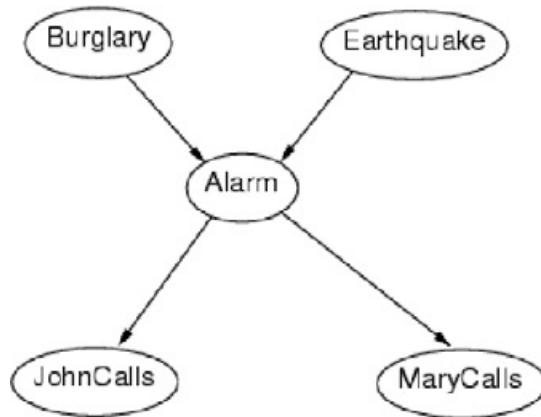
Costruzione di una rete

Risultato finale con l'ordinamento

M, J, A, B, E

sintomo, sintomo, causa intermedia, causa iniziale, causa iniziale

Un numero maggiore di link rispetto a:

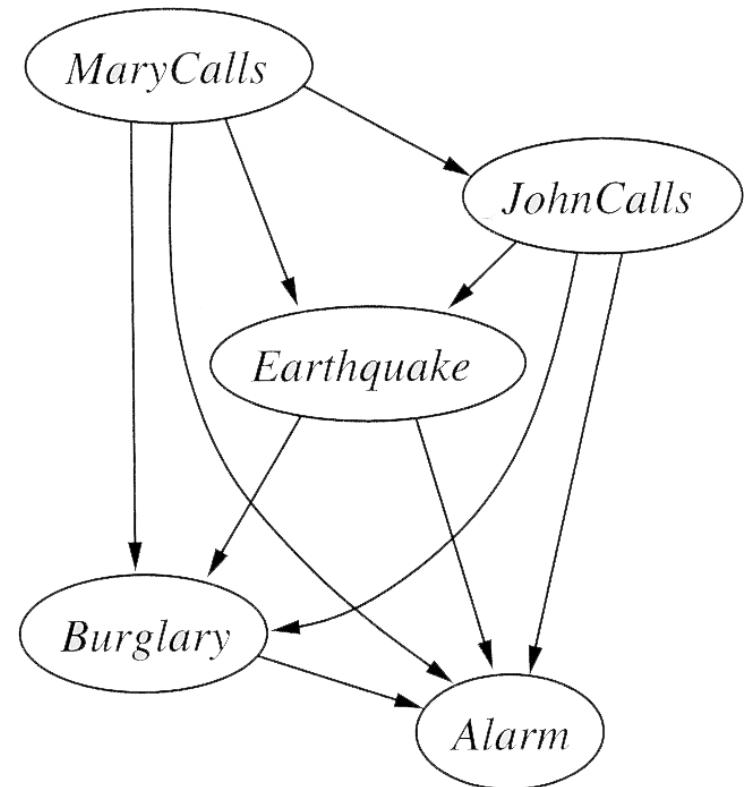
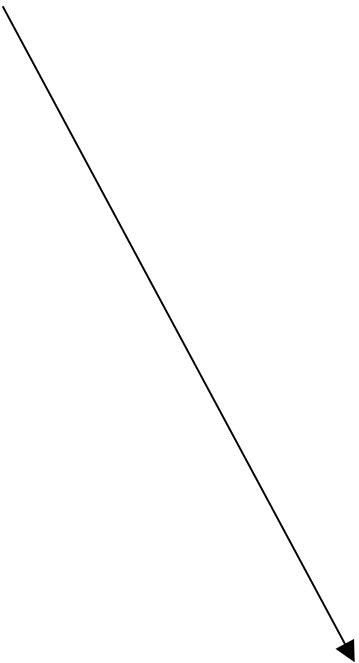


Alcuni link sono innaturali e di difficile stima
(Probabilità di *Earthquake*, dato *Alarm* e *Burglary*)

Costruzione di una rete

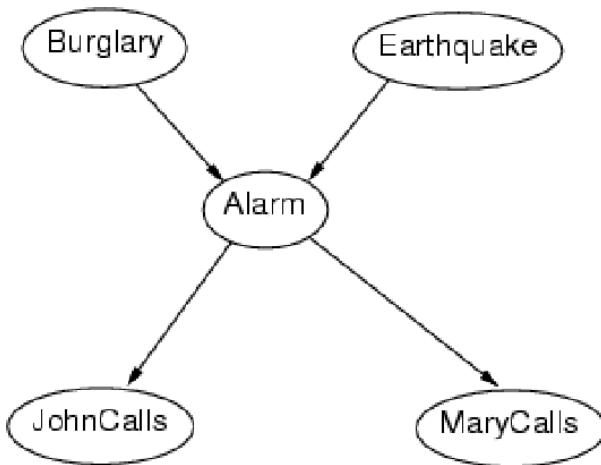
Risultato finale con l'ordinamento

M, J, E, B, A

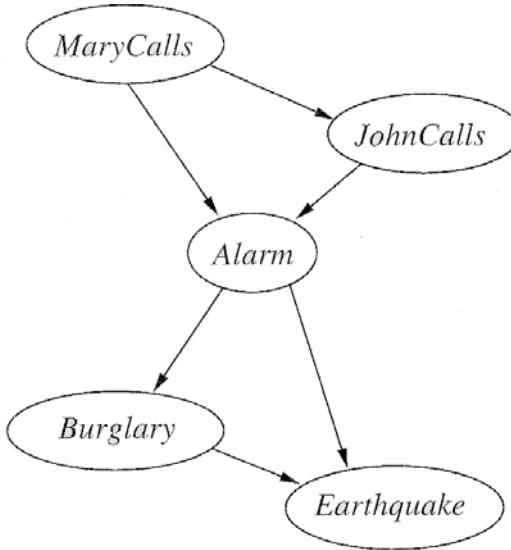


sintomo, sintomo, causa iniziale, causa iniziale, causa intermedia

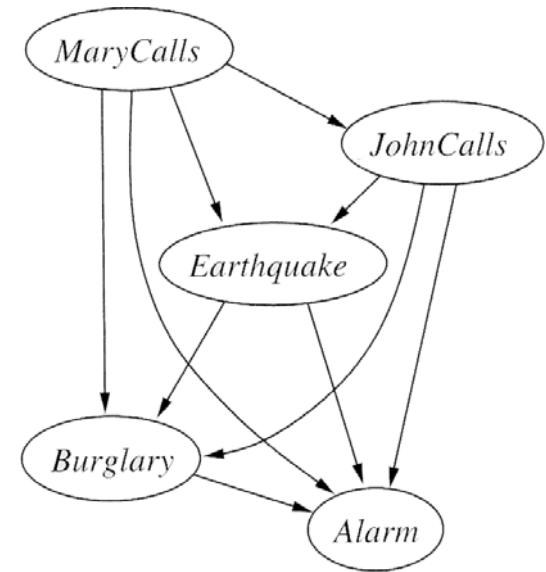
Costruzione di una rete



Ordine causale



Ordine diagnostico



Ordine diagnostico
“a salti”

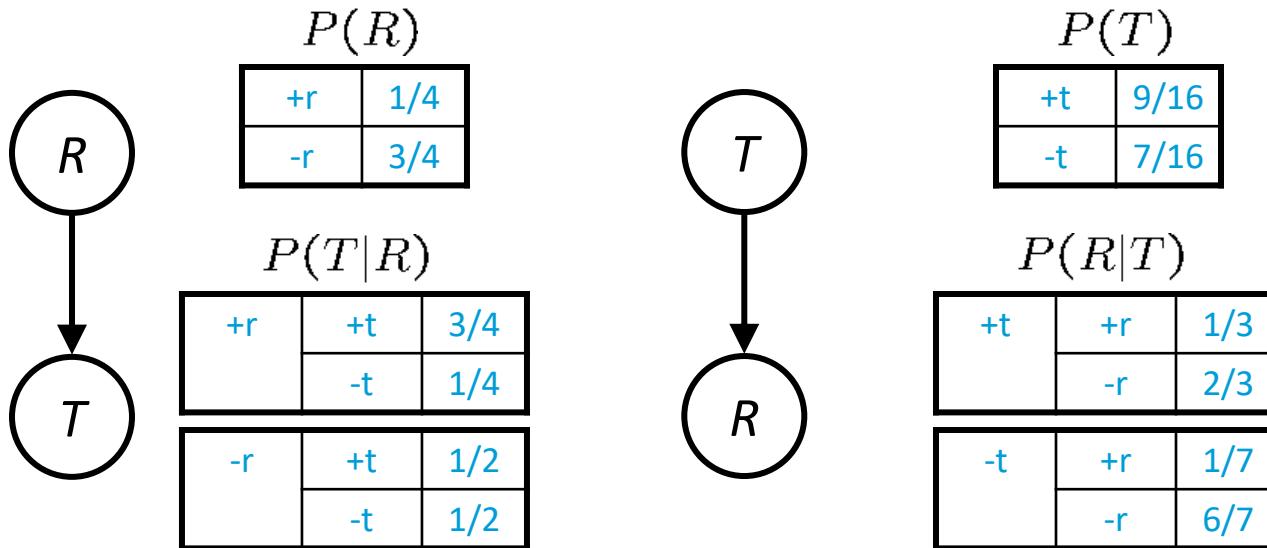
Costruzione di una rete

- ▶ Data la semantica della rete, come la si costruisce?
- ▶ Secondo modo: Costruzione di una rete causale
 - ▶ Si definiscono le cause prime
 - ▶ Si connettono le cause prime agli effetti diretti
 - ▶ Si procede allo stesso modo interpretando gli effetti come nuove cause
- ▶ L'ordine di scelta causale (dalle “cause prime” ai successivi effetti, causa di altri effetti,...) porta a costruire reti ottime

Causalità

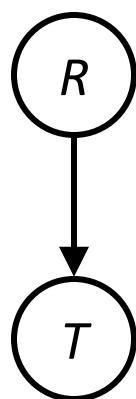
► BNs non devo per forza essere causali

- A volte non esiste una rete causale sul dominio (soprattutto se mancano variabili)
- E.g. consideriamo le variabili Traffic e Rain



Esempio: Traffico

► Direzione causale



$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

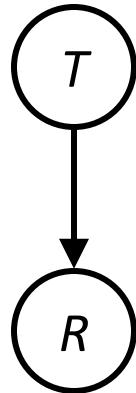
$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16



Esempio: Traffico Inverso

- ▶ Causalità inversa?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7



$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

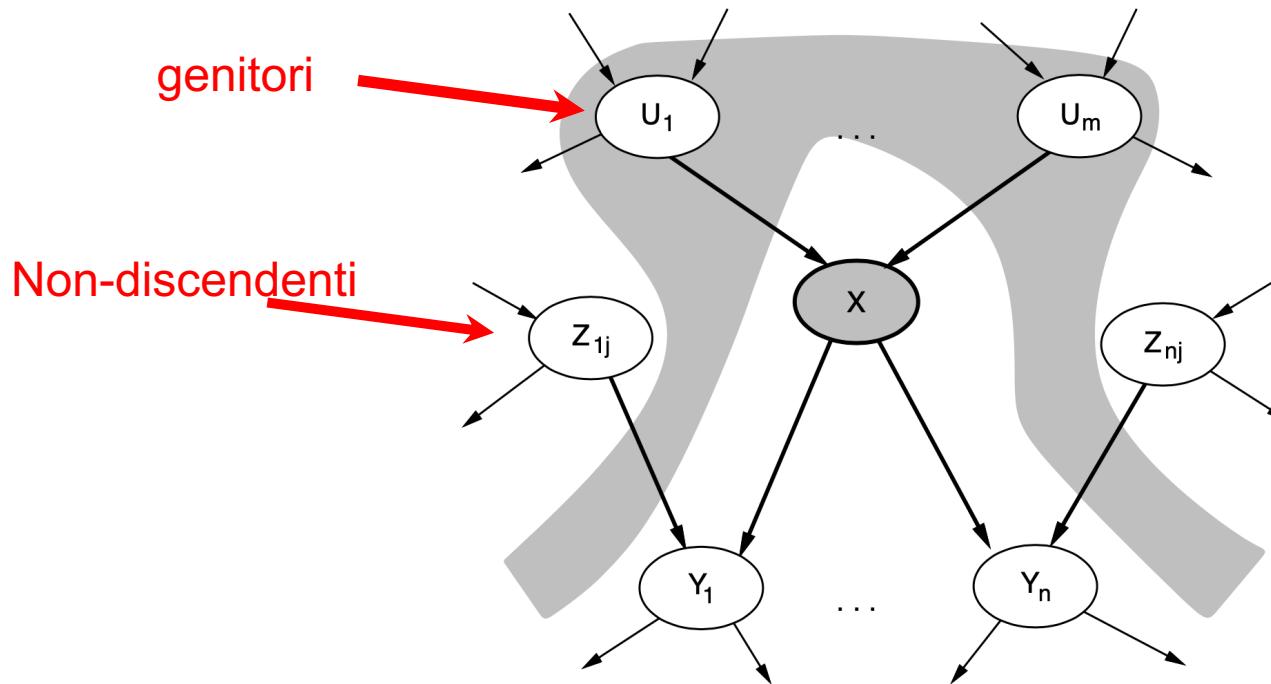
Causalità?

- ▶ Quando le reti di Bayes riflettono i veri modelli causali:
 - ▶ Spesso più semplici (i nodi hanno meno genitori)
 - ▶ Spesso più facili da pensare
 - ▶ Spesso è più facile stimare le probabilità dai dati
- ▶ I BN non devono essere effettivamente causali
 - ▶ A volte non esiste una rete causale sul dominio (soprattutto se mancano variabili)
 - ▶ Per esempio, consideriamo le variabili Traffico e Gocciolamento
 - ▶ Finisci con le frecce che riflettono la correlazione, non la causalità
- ▶ Cosa significano veramente le frecce?
 - ▶ Può capitare che la topologia codifichi la struttura causale
 - ▶ La topologia codifica davvero l'indipendenza condizionale

$$P(x_i|x_1, \dots x_{i-1}) = P(x_i|parents(X_i))$$

Semantica Locale

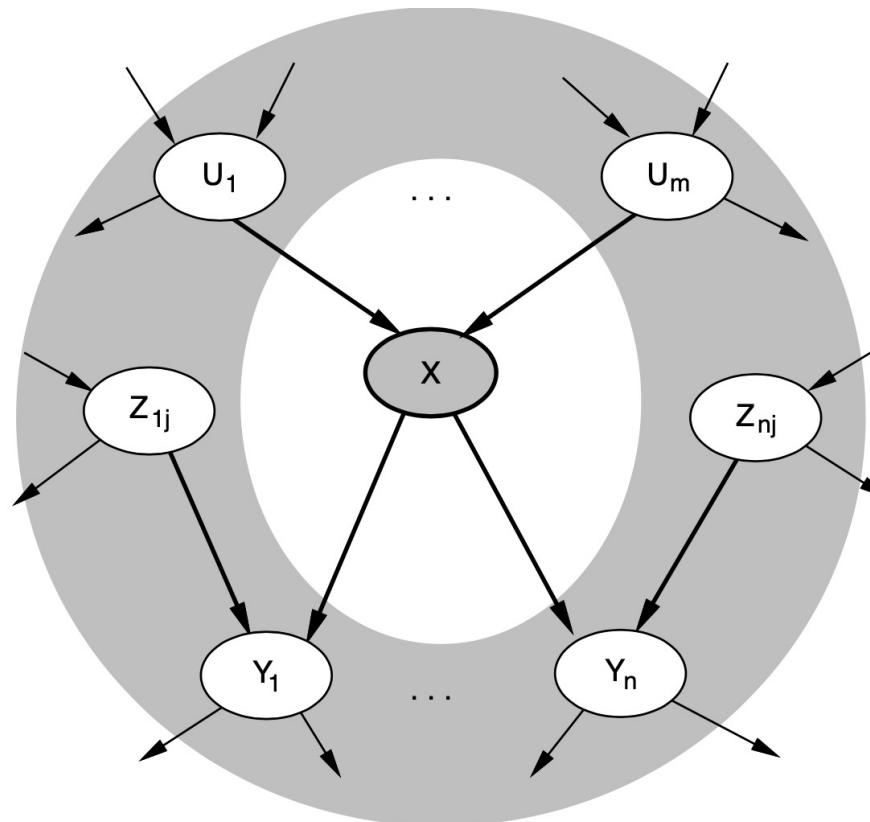
- ▶ Semantica locale (topologica): ogni nodo è condizionatamente indipendente dai suoi non-descendenti dati i suoi genitori



- ▶ Teorema: Local semantics \Leftrightarrow global semantics

Semantica Locale

- ▶ **Coperta di Markov:** ogni nodo è condizionatamente indipendente da tutti gli altri nodi data la sua Coperta di Markov: genitori + figli + genitori di figli



Inferenza esatta su reti bayesiane

- Inferenza esatta
- Calcolare la probabilità CONDIZIONATA (a posteriori) di un insieme di *query variables*, dato un evento (un assegnamento di valori ad un insieme di *evidence variables*)

Inferenza esatta su reti bayesiane

- Procedura di inferenza per rispondere a query probabilistiche per variabili discrete (query su singola variabile) (per enumerazione delle entries in una full joint distribution) – Utilizzando la product rule:
 \mathbf{X} Variabile Query, \mathbf{E} Variabili evidenza, \mathbf{e} valori osservati per \mathbf{E} , \mathbf{Y} Variabili Hidden i cui valori sono \mathbf{y}

$$\text{Query: } \mathbf{P}(\mathbf{X} | \mathbf{e})$$

Distribuzione di probabilità di \mathbf{X} condizionata da \mathbf{e}

$$\mathbf{P}(\mathbf{X} | \mathbf{e}) = \alpha \quad \mathbf{P}(\mathbf{X}, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(\mathbf{X}, \mathbf{e}, \mathbf{y})$$

Somma su tutti i possibili valori \mathbf{y} delle variabili non osservate

Inferenza esatta su reti bayesiane

$$P(X | e) = \alpha P(X, e) = \alpha \sum_y P(X, e, y)$$

In una rete bayesiana i termini $P(X, e, y)$ (probabilità dell'EVENTO ATOMICO) possono essere scritti come prodotti di probabilità condizionate prese dalla rete (CPT)



$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i | Parents(X_i))$$

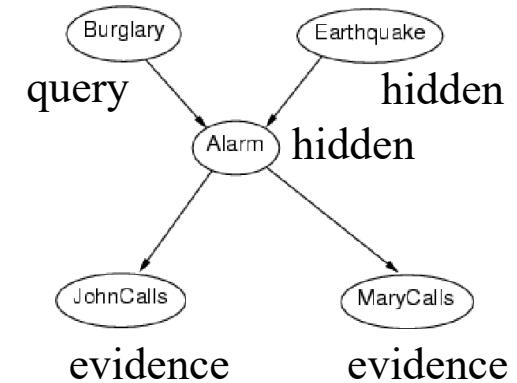
Inferenza esatta su reti bayesiane

Esempio Query:

$$P(\text{Burglary} \mid \text{JohnCalls}=\text{true}, \text{MaryCalls}=\text{true})$$

Inferenza in verso diagnostico utilizzando

CPT scritte in verso causale (Teorema di Bayes)



$$P(B \mid j, m) = \alpha P(B, j, m) = \alpha \sum_e \sum_a P(B, j, m, e, a)$$

Somma su tutti i possibili valori y
delle variabili hidden e, a

Calcolo per *Burglary* = true utilizzando: $P(x_i \mid \text{Parents}(X_i))$

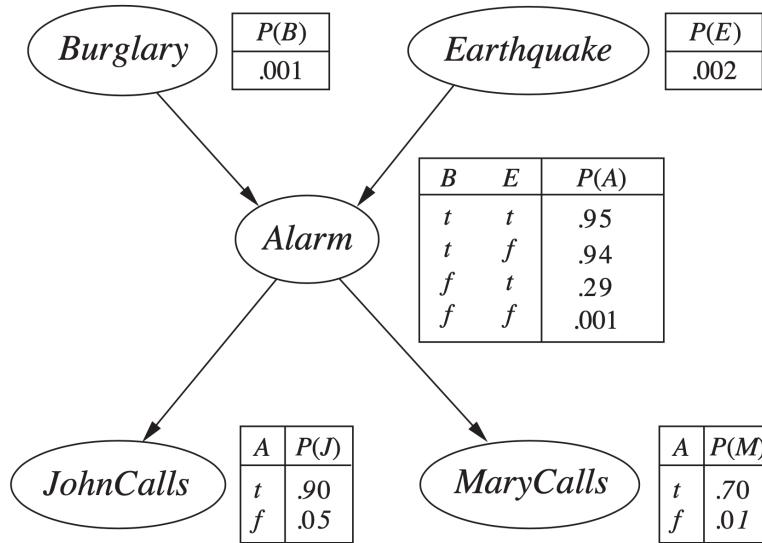
$$P(b \mid j, m) = \alpha \sum_e \sum_a P(b) P(j|a) P(m|a) P(e) P(a|b, e)$$

Devo sommare 4 termini (2*2 combinazioni di valori di *e* ed *a*)

Ogni termine è un prodotto di 5 numeri. Con *n* variabili booleane $O(n2^n)$

Inferenza esatta su reti bayesiane

$$P(b | j, m) = \alpha P(b) \sum_e P(e) \sum_a P(j|a) P(m|a) P(a|b, e)$$



$$P(B | j, m) = \alpha <0,00059224, 0,0014919>$$

Normalizzando a 1 con α : $<0,284, 0,716>$

Enumerazione ricorsiva in profondità: $O(n)$ spazio, $O(d^n)$ tempo

Inferenza esatta su reti bayesiane

function ENUMERATION-ASK(X, \mathbf{e}, bn) **returns** a distribution over X

inputs: X , the query variable

\mathbf{e} , observed values for variables E

bn , a Bayesian network with variables $\{X\} \cup E \cup Y$

$Q(X) \leftarrow$ a distribution over X , initially empty

for each value x_i of X **do**

 extend \mathbf{e} with value x_i for X

$Q(x_i) \leftarrow$ ENUMERATE-ALL(VARS[bn], \mathbf{e})

return NORMALIZE($Q(X)$)

function ENUMERATE-ALL($vars, \mathbf{e}$) **returns** a real number

if EMPTY?($vars$) **then return** 1.0

$Y \leftarrow$ FIRST($vars$)

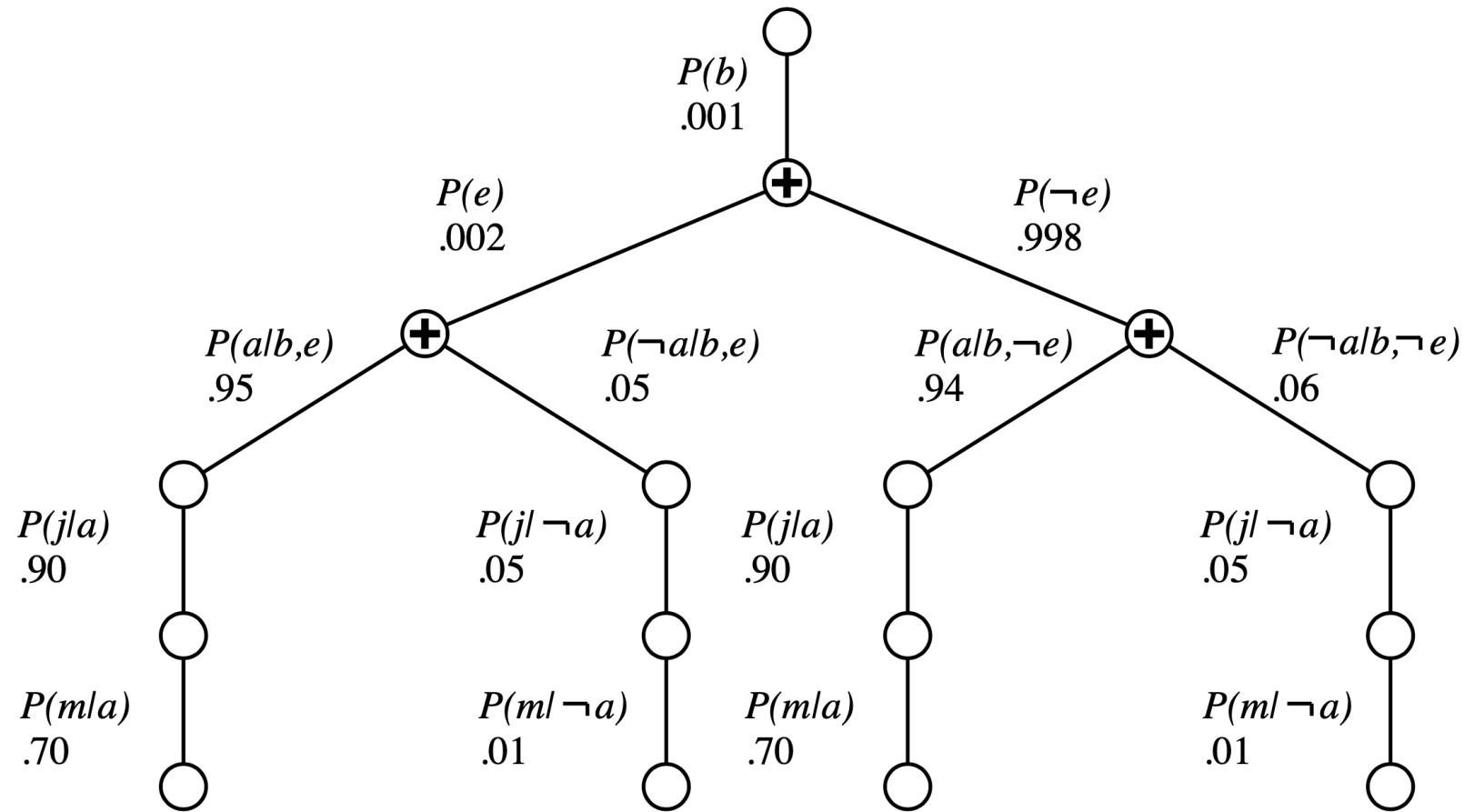
if Y has value y in \mathbf{e}

then return $P(y | Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), \mathbf{e})

else return $\sum_y P(y | Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), \mathbf{e}_y)

 where \mathbf{e}_y is \mathbf{e} extended with $Y = y$

Inferenza esatta su reti bayesiane



Inferenza attraverso l'eliminazione di variabili

- Eliminazione variabile: eseguire somme da destra a sinistra, memorizzando risultati intermedi (**fattori**) per evitare la ricomputazione

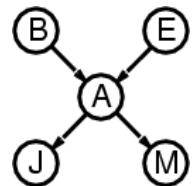
$$\begin{aligned}\mathbf{P}(B|j, m) &= \alpha \underbrace{\mathbf{P}(B)}_B \underbrace{\sum_e P(e)}_E \underbrace{\sum_a \mathbf{P}(a|B, e)}_A \underbrace{P(j|a)}_J \underbrace{P(m|a)}_M \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B, e) P(j|a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B, e) f_J(a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a f_A(a, b, e) f_J(a) f_M(a) \\ &= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b, e) \text{ (sum out } A\text{)} \\ &= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\ &= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)\end{aligned}$$

Vettore di due elementi
 $P(m|a)$ e $P(m|\text{not } a)$

Variabili irrilevanti

- ▶ Consideriamo la query $P(JohnCalls | Burglary = \text{true})$

$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b, e) P(J|a) \sum_m P(m|a)$$



- ▶ La somma su m è 1; M è irrilevante per la query
- ▶ Teorema 1: Y è irrilevante a meno che $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Here, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, and
 $Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
so $MaryCalls$ is irrelevant

- ▶ Un algoritmo di eliminazione variabili può quindi eliminare tutte queste variabili prima di valutare la query

Complessità dell'inferenza esatta

- ▶ Reti singolarmente connesse (o polialberi) :
 - ▶ due nodi qualsiasi sono collegati al massimo da un percorso non orientato
 - ▶ il costo in tempo e spazio dell'eliminazione variabili è $O(d^k n)$
- ▶ Reti a connessioni multiple:
 - ▶ Può ridursi a 3SAT => NP-Hard
 - ▶ Equivalente a contare modelli 3SAT => #P-hard



Inferenza esatta su reti bayesiane

- Complessità dell'inferenza esatta
 - In generale l'inferenza su reti Bayesiane è NP-Hard
 - Riduzione della complessità delle connessioni nella rete (Polialbero – al massimo un percorso indiretto tra due nodi della rete – lineare nel numero n dei nodi)
 - Metodi approssimati di inferenza

Inferenza mediante simulazione stocastica

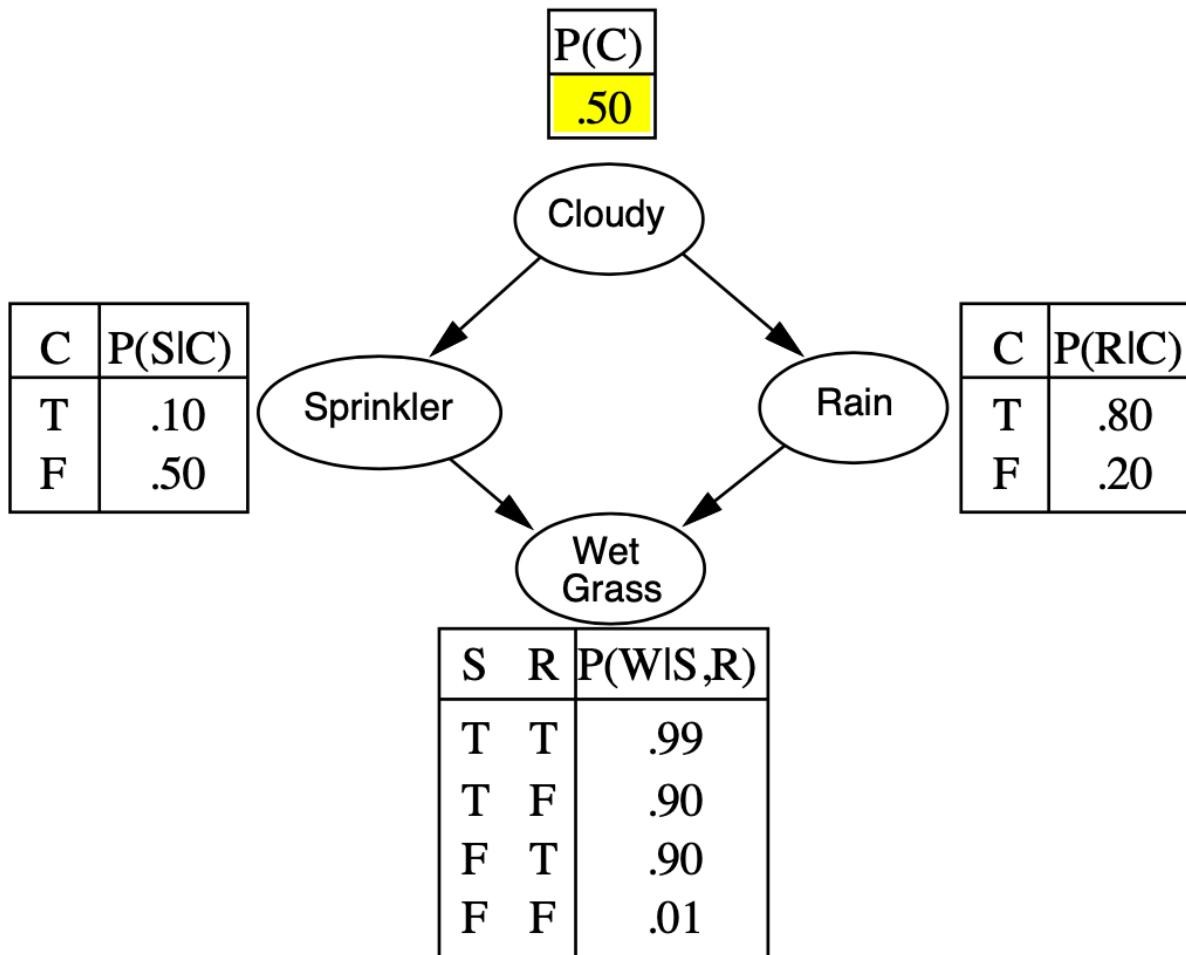
- ▶ Idea base:
 - 1) Genera N campioni da una distribuzione di campionamento S
 - 2) Calcola una probabilità a posteriori approssimativa \hat{P}
 - 3) Mostra che converge alla probabilità reale P
- ▶ Tecniche che vedremo:
 - ▶ Campionamento da una rete vuota
 - ▶ Campionamento di rigetto: respingere i campioni in disaccordo con le evidenze
 - ▶ Pesatura di verosimiglianza: utilizzare le evidenze per ponderare i campioni
 - ▶ Catena di Markov Monte Carlo (MCMC): campiona da un processo stocastico la cui distribuzione stazionaria è la vera prob. a posteriori

Campionamento da una rete vuota

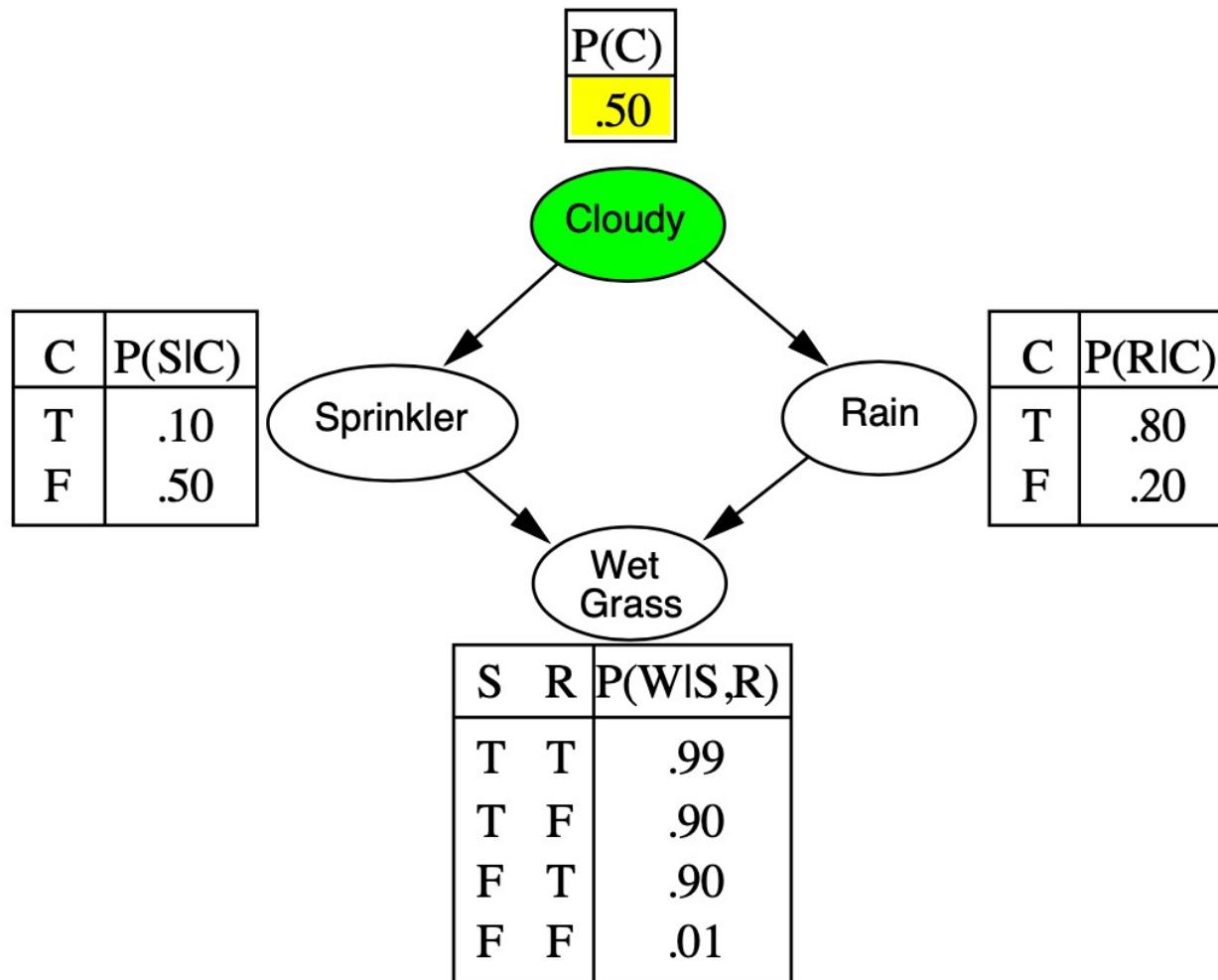
```
function PRIOR-SAMPLE(bn) returns an event sampled from bn
  inputs: bn, a belief network specifying joint distribution  $\mathbf{P}(X_1, \dots, X_n)$ 
  x  $\leftarrow$  an event with  $n$  elements
  for  $i = 1$  to  $n$  do
     $x_i \leftarrow$  a random sample from  $\mathbf{P}(X_i \mid \text{parents}(X_i))$ 
    given the values of  $\text{Parents}(X_i)$  in x
  return x
```

- ▶ Idea base:
 - 1) Campionare a turno ogni variabile, in ordine topologico.
 - 2) La distribuzione da cui si campiona è condizionata ai valori già assegnati ai genitori della variabile

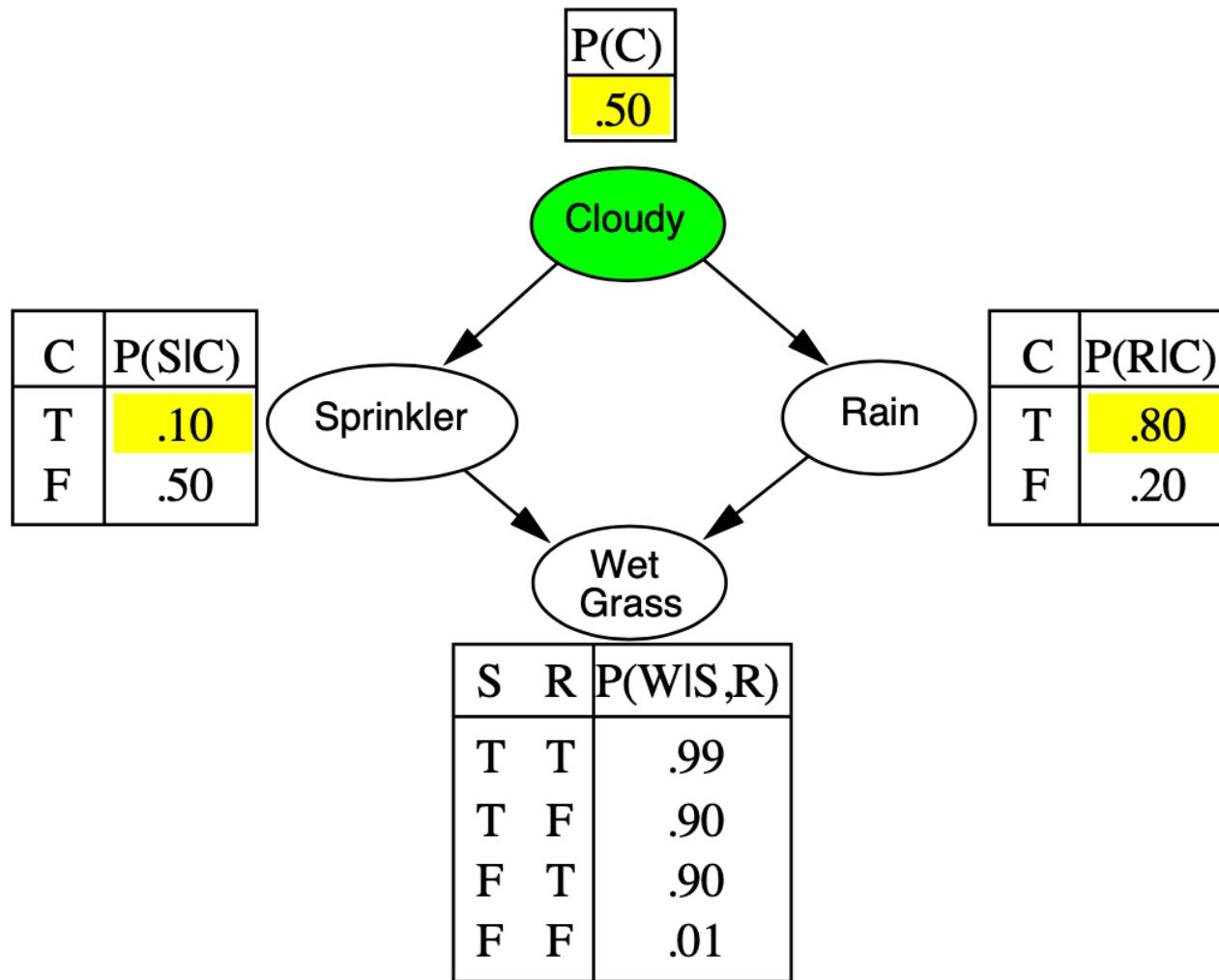
Esempio



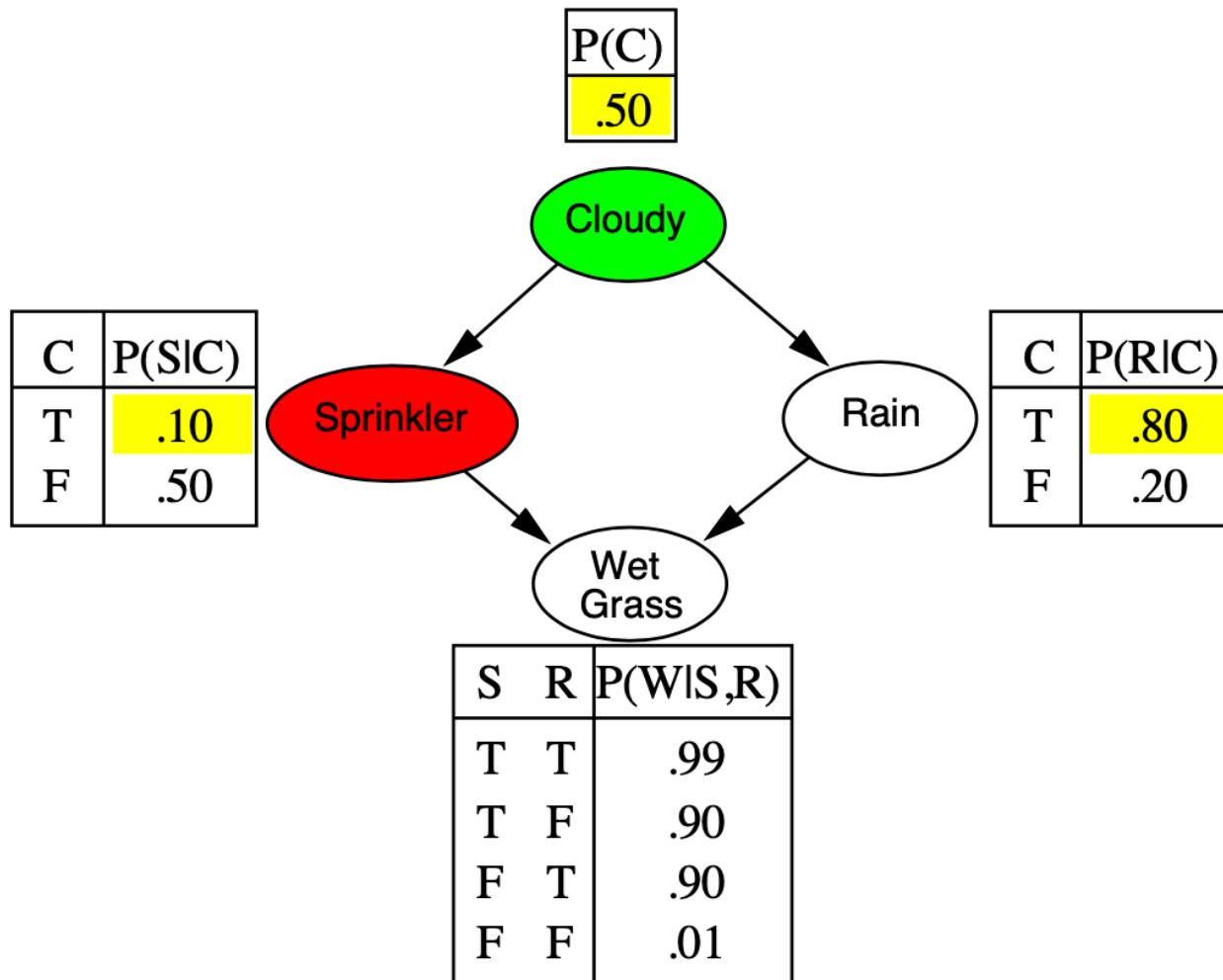
Esempio



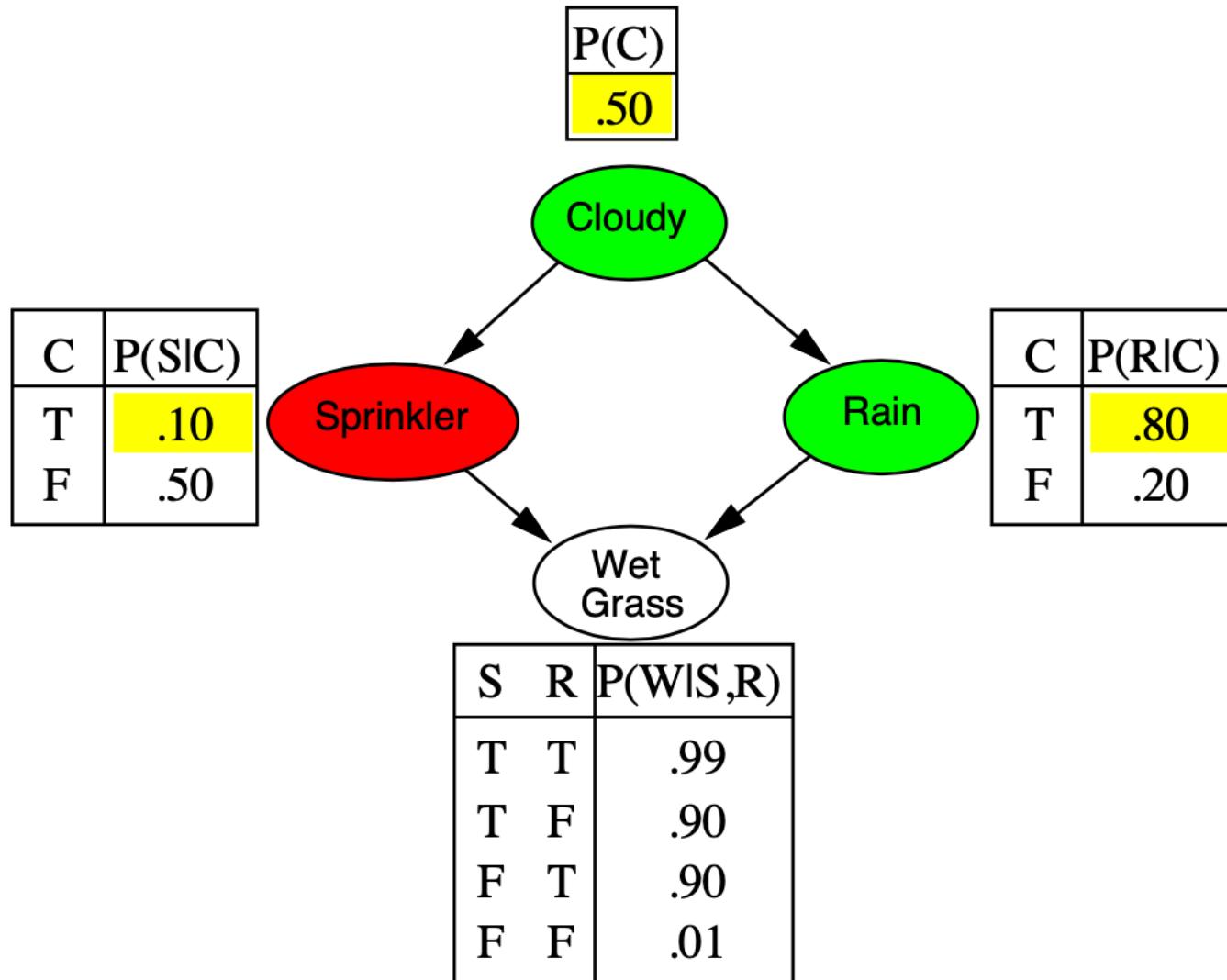
Esempio



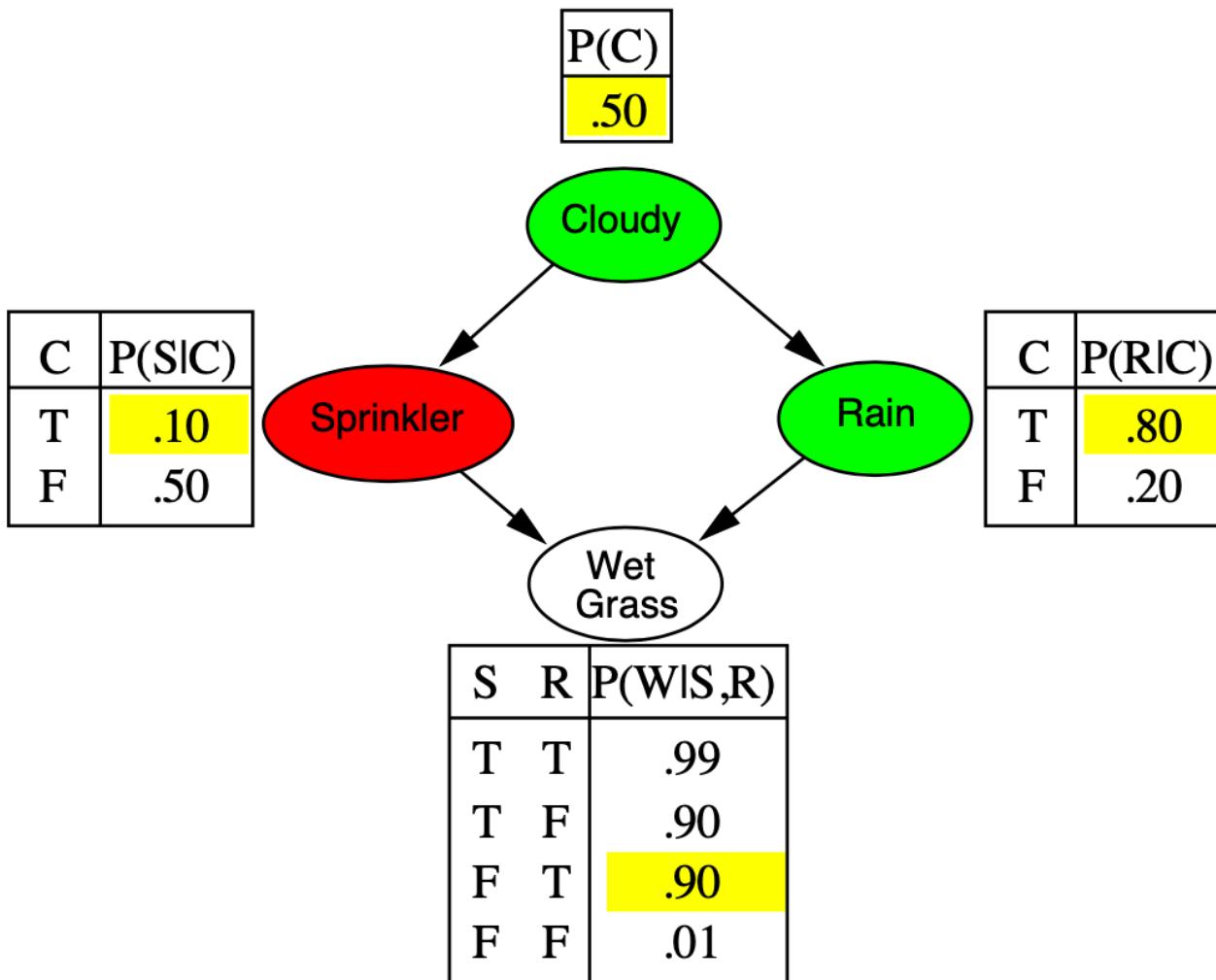
Esempio



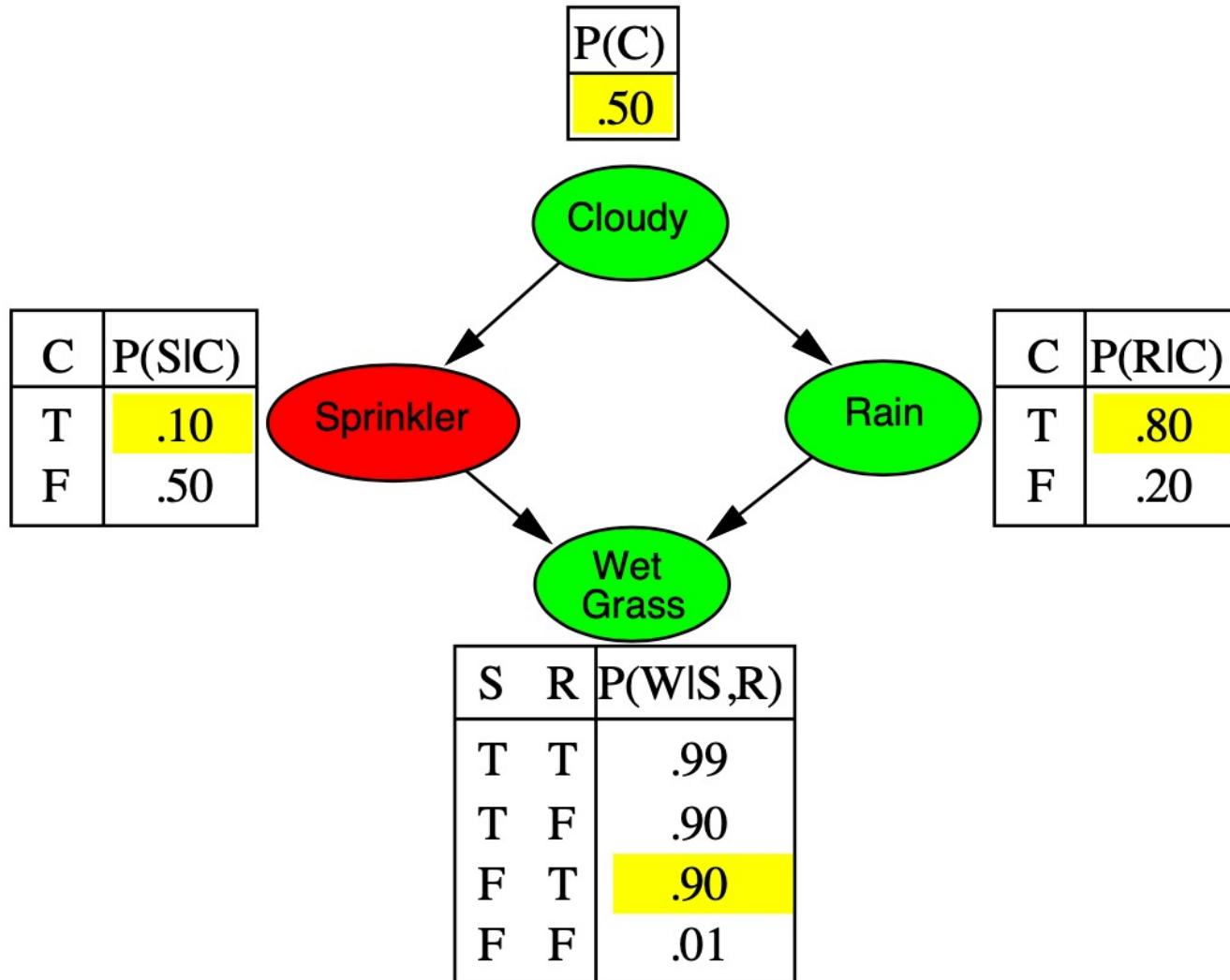
Esempio



Esempio



Esempio



Campionamento

- ▶ Per simulare una moneta con D facce:
- ▶ Step 1: ottieni il campione u dalla distribuzione uniforme su $[0, 1]$
 - ▶ Per esempio. `random()` in Python
- ▶ Step 2: converti questo campione u in un risultato per la data distribuzione associando ciascun risultato x con un sottointervallo di dimensioni $P(x)$ di $[0,1]$



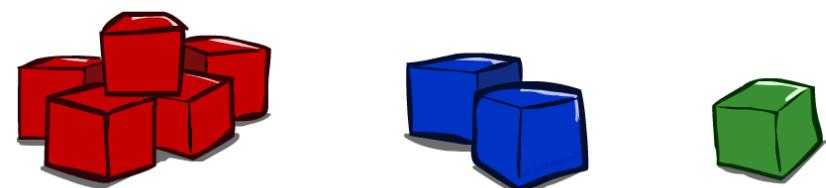
0.6 0.1 0.3

■ Esempio

C	$P(C)$
red	0.6
green	0.1
blue	0.3

$0.0 \leq u < 0.6, \rightarrow C=\text{red}$
 $0.6 \leq u < 0.7, \rightarrow C=\text{green}$
 $0.7 \leq u < 1.0, \rightarrow C=\text{blue}$

- Se `random()` restituisce $u = 0.83$, allora il campione è $C = \text{blue}$
- E.g, dopo aver campionato 8 volte:



Campionamento da una rete vuota

- ▶ Probabilità che PriorSample generi un evento particolare

$$S_{PS}(x_1 \dots x_n) = \prod_{i=1}^n P(x_i | parents(X_i)) = P(x_1 \dots x_n)$$

- ▶ cioè la vera probabilità a priori (distr. congiunta rete Bayesiana)
- ▶ Ad esempio $S_{PS}(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t)$

- ▶ Sia $N_{PS}(x_1 \dots x_n)$ numeri di eventi in cui x_1, \dots, x_n è generato

- ▶ Esempio: se ho 1000 campioni per la rete WetGrass, e in 511 Rain=true, allora $P(\text{Rain}=true) = 0,511$.

- ▶ Allora abbiamo
$$\begin{aligned} \lim_{N \rightarrow \infty} \hat{P}(x_1, \dots, x_n) &= \lim_{N \rightarrow \infty} N_{PS}(x_1, \dots, x_n)/N \\ &= S_{PS}(x_1, \dots, x_n) \\ &= P(x_1 \dots x_n) \end{aligned}$$

- ▶ In altre parole, le stime derivate da PriorSample sono CONSISTENTI

$$\hat{P}(x_1, \dots, x_n) \approx P(x_1 \dots x_n)$$

Campionamento di rigetto

- $\hat{P}(X|e)$ stimato da campioni concordanti con l'evidenza e

```
function REJECTION-SAMPLING( $X, e, bn, N$ ) returns an estimate of  $P(X|e)$ 
local variables:  $N$ , a vector of counts over  $X$ , initially zero
```

```
for  $j = 1$  to  $N$  do
     $x \leftarrow$  PRIOR-SAMPLE( $bn$ )
    if  $x$  is consistent with  $e$  then
         $N[x] \leftarrow N[x] + 1$  where  $x$  is the value of  $X$  in  $x$ 
return NORMALIZE( $N[X]$ )
```

- Ad esempio, stima $P(Rain|Sprinkler=true)$ usando 100 campioni
 - 27 campioni hanno $Sprinkler=true$
 - di questi 8 hanno $Rain=true$ e 19 hanno $Rain=false$
- $\hat{P}(Rain|Sprinkler=true) = \text{NORMALIZE}(\langle 8, 19 \rangle) = \langle 0.296, 0.704 \rangle$
- Simile a una procedura di stima empirica nel mondo reale

Campionamento di rigetto

$$\begin{aligned}\hat{\mathbf{P}}(X|\mathbf{e}) &= \alpha \mathbf{N}_{PS}(X, \mathbf{e}) && \text{(algorithm defn.)} \\ &= \mathbf{N}_{PS}(X, \mathbf{e}) / N_{PS}(\mathbf{e}) && \text{(normalized by } N_{PS}(\mathbf{e})\text{)} \\ &\approx \mathbf{P}(X, \mathbf{e}) / P(\mathbf{e}) && \text{(property of PRIORSAMPLE)} \\ &= \mathbf{P}(X|\mathbf{e}) && \text{(defn. of conditional probability)}\end{aligned}$$

- ▶ Quindi il campionamento di rigetto restituisce stime a posteriori CONSISTENTI
- ▶ Problema: irrimediabilmente costoso se $P(\mathbf{e})$ è piccolo (rifiuterà moltissimi campioni!)
- ▶ $P(\mathbf{e})$ diminuisce esponenzialmente al crescere del numero di variabili evidenza!

Pesatura di verosimiglianza (LW)

- ▶ Idea: fissa le variabili evidenza, campiona solo le variabili non di evidenza e pesa ogni campione in base alla probabilità che si accordi con le evidenze (**gli eventi in cui è improbabile che le prove siano verificate devono pesare di meno nel conteggio**)

```
function LIKELIHOOD-WEIGHTING( $X, \mathbf{e}, bn, N$ ) returns an estimate of  $P(X|\mathbf{e})$ 
local variables:  $\mathbf{W}$ , a vector of weighted counts over  $X$ , initially zero
for  $j = 1$  to  $N$  do
     $\mathbf{x}, w \leftarrow$  WEIGHTED-SAMPLE( $bn$ )
     $\mathbf{W}[\mathbf{x}] \leftarrow \mathbf{W}[\mathbf{x}] + w$  where  $x$  is the value of  $X$  in  $\mathbf{x}$ 
return NORMALIZE( $\mathbf{W}[X]$ )
```

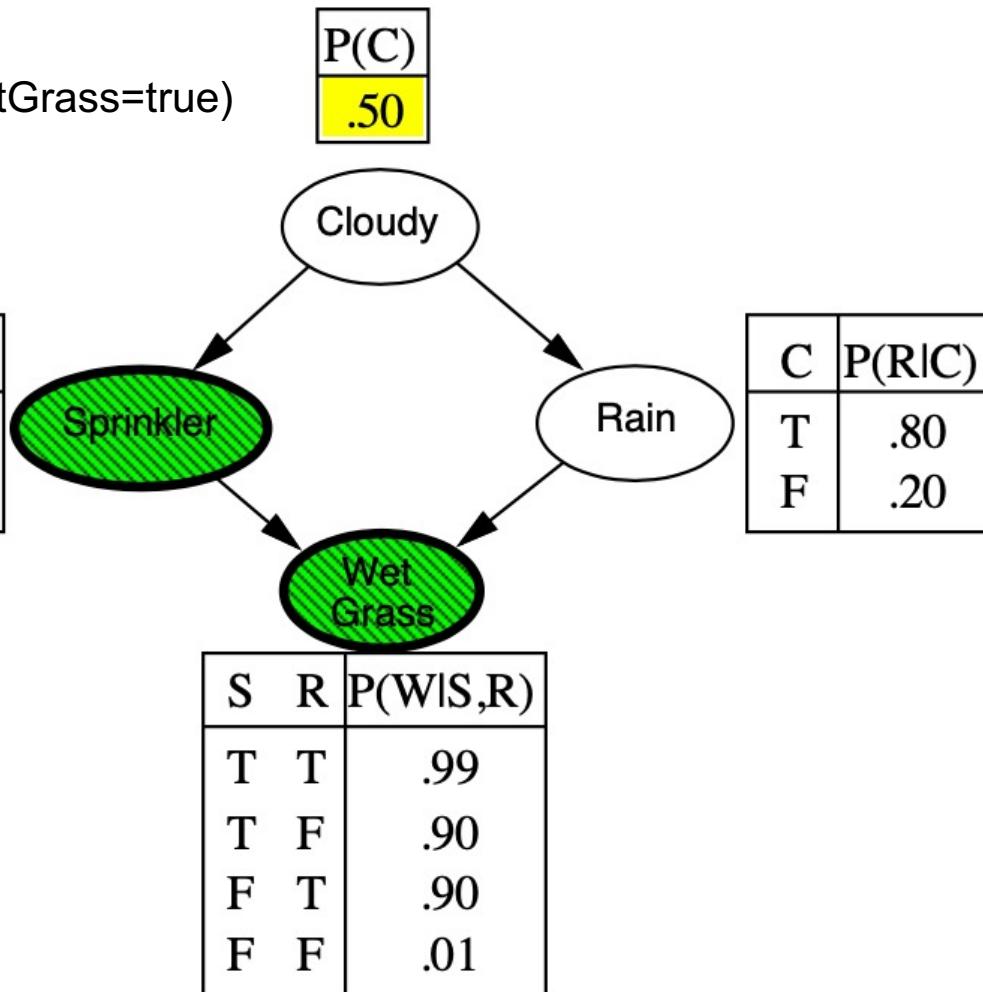
```
function WEIGHTED-SAMPLE( $bn, \mathbf{e}$ ) returns an event and a weight
 $\mathbf{x} \leftarrow$  an event with  $n$  elements;  $w \leftarrow 1$ 
for  $i = 1$  to  $n$  do
    if  $X_i$  has a value  $x_i$  in  $\mathbf{e}$ 
        then  $w \leftarrow w \times P(X_i = x_i | parents(X_i))$ 
        else  $x_i \leftarrow$  a random sample from  $P(X_i | parents(X_i))$ 
return  $\mathbf{x}, w$ 
```

Pesatura di verosimiglianza

Query

$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$

C	$P(S C)$
T	.10
F	.50



$$w = 1.0$$

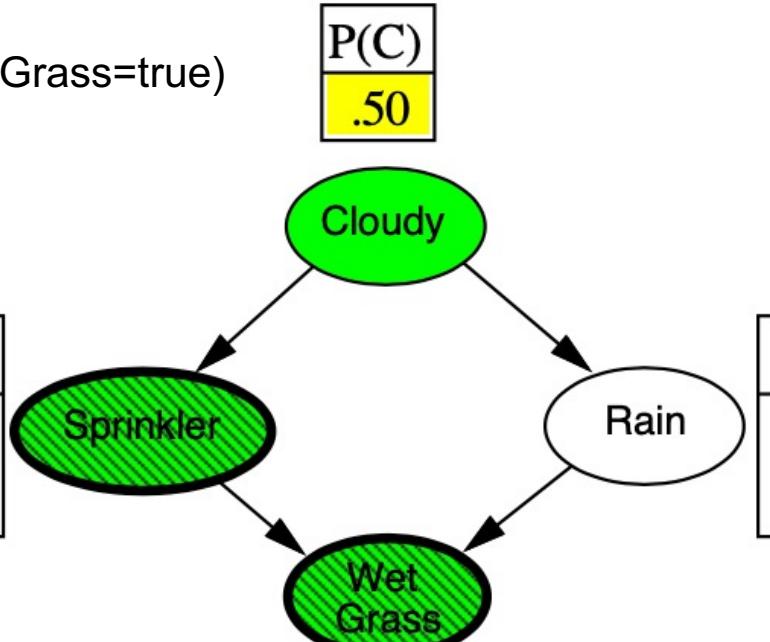
Pesatura di verosimiglianza

Query

$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$

P(C)	
.50	

C	P(S C)
T	.10
F	.50



C	P(R C)
T	.80
F	.20

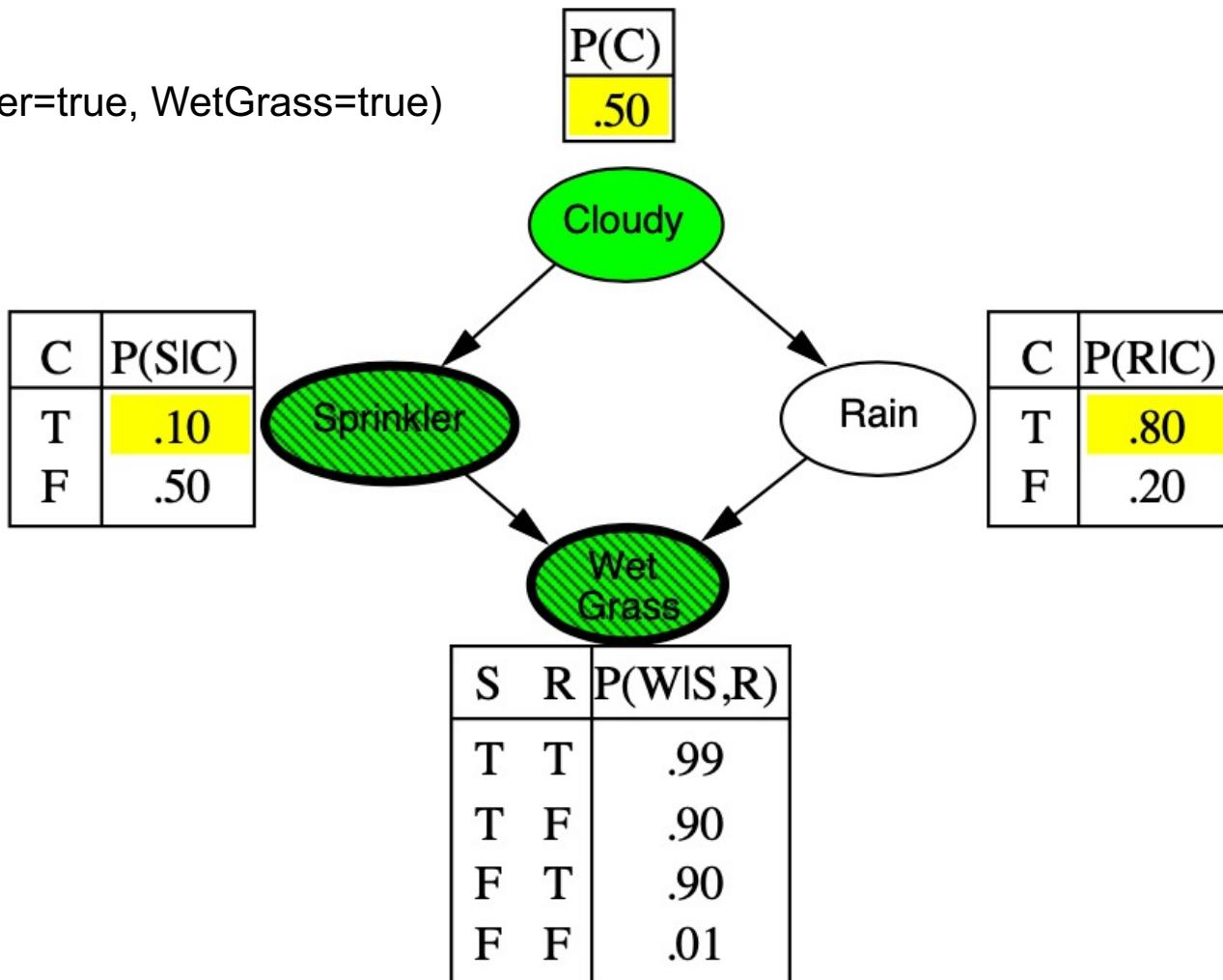
S	R	P(W S,R)
T	T	.99
T	F	.90
F	T	.90
F	F	.01

$$w = 1.0$$

Pesatura di verosimiglianza

Query

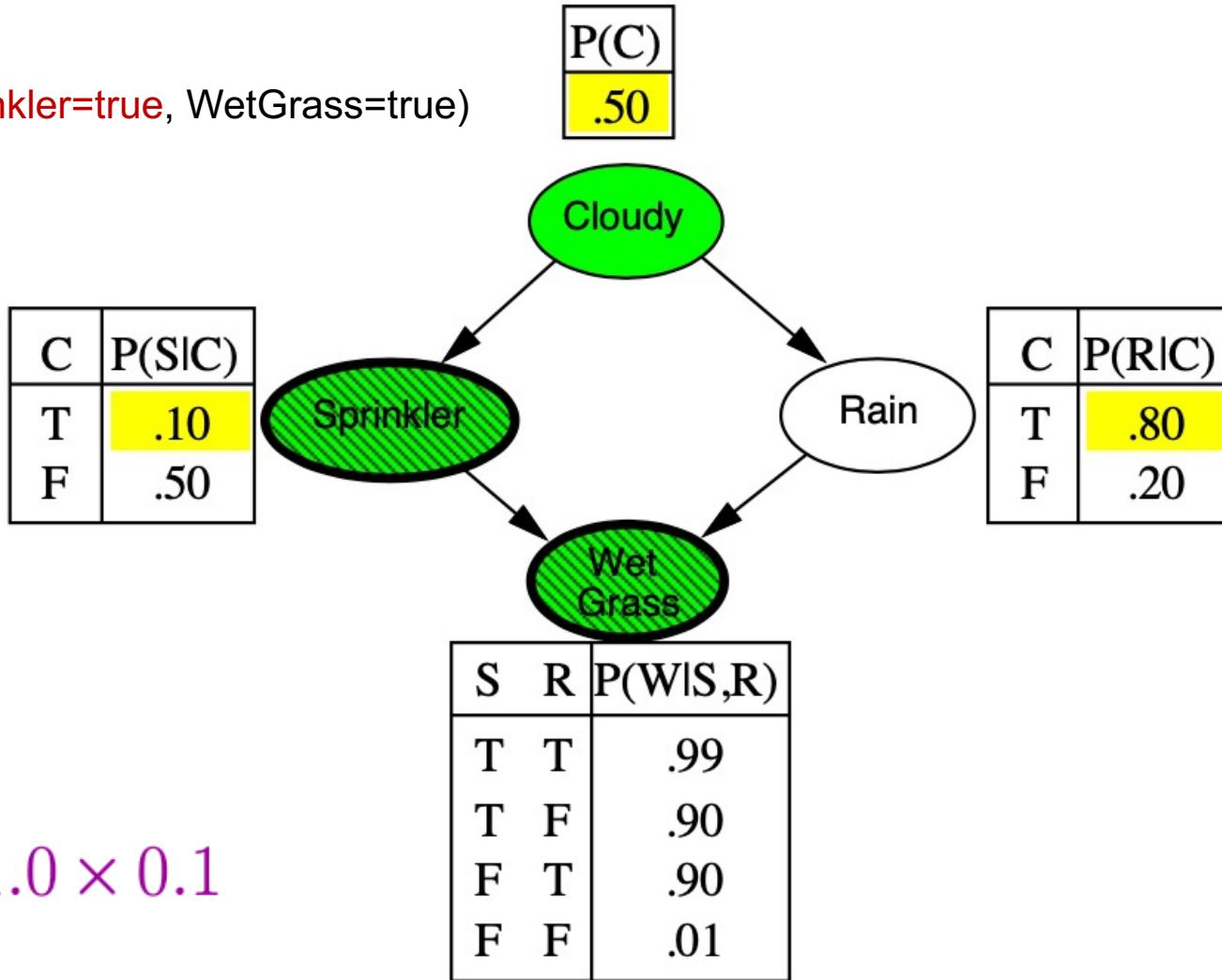
$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$



Pesatura di verosimiglianza

Query

$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$

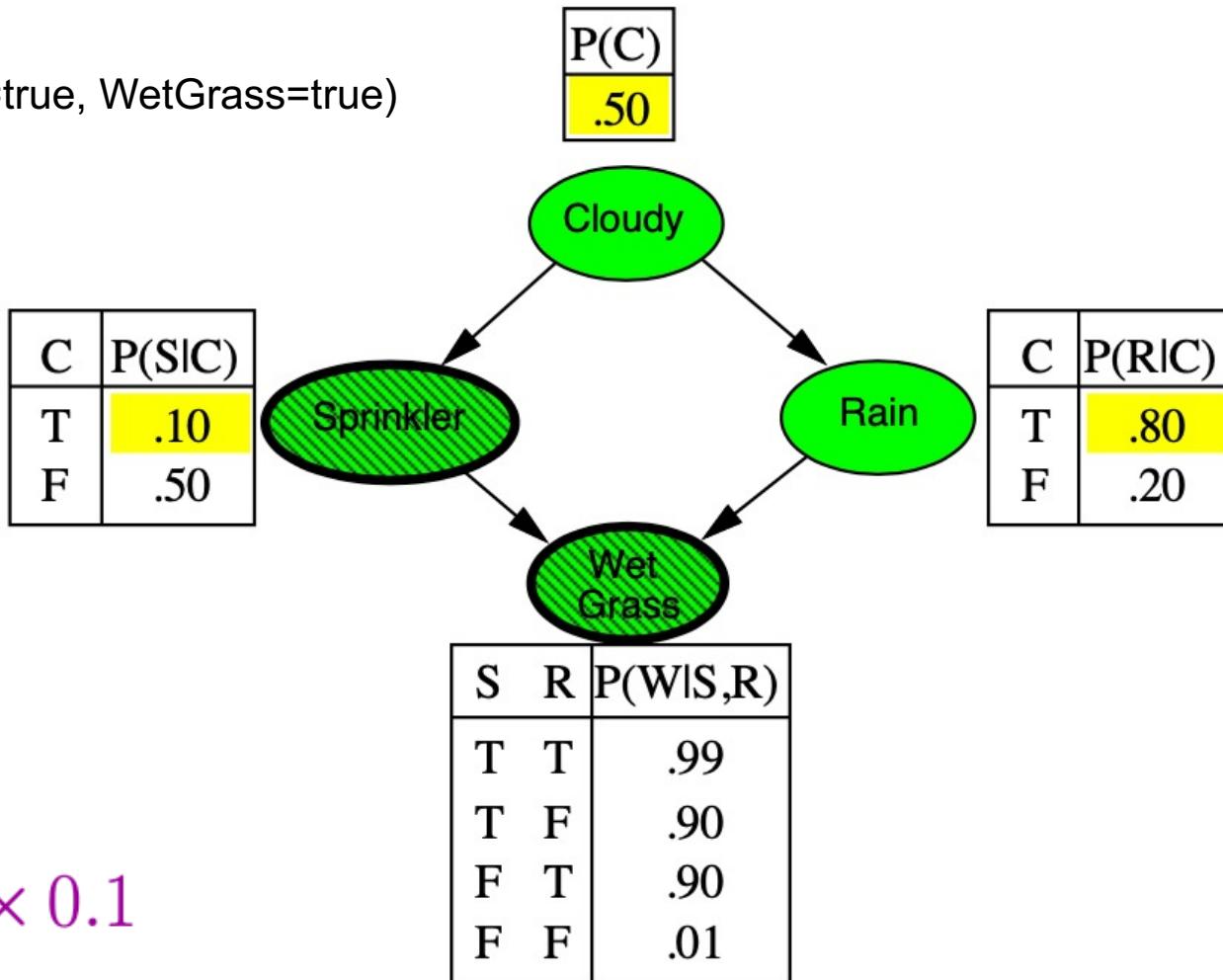


$$w = 1.0 \times 0.1$$

Pesatura di verosimiglianza

Query

$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$

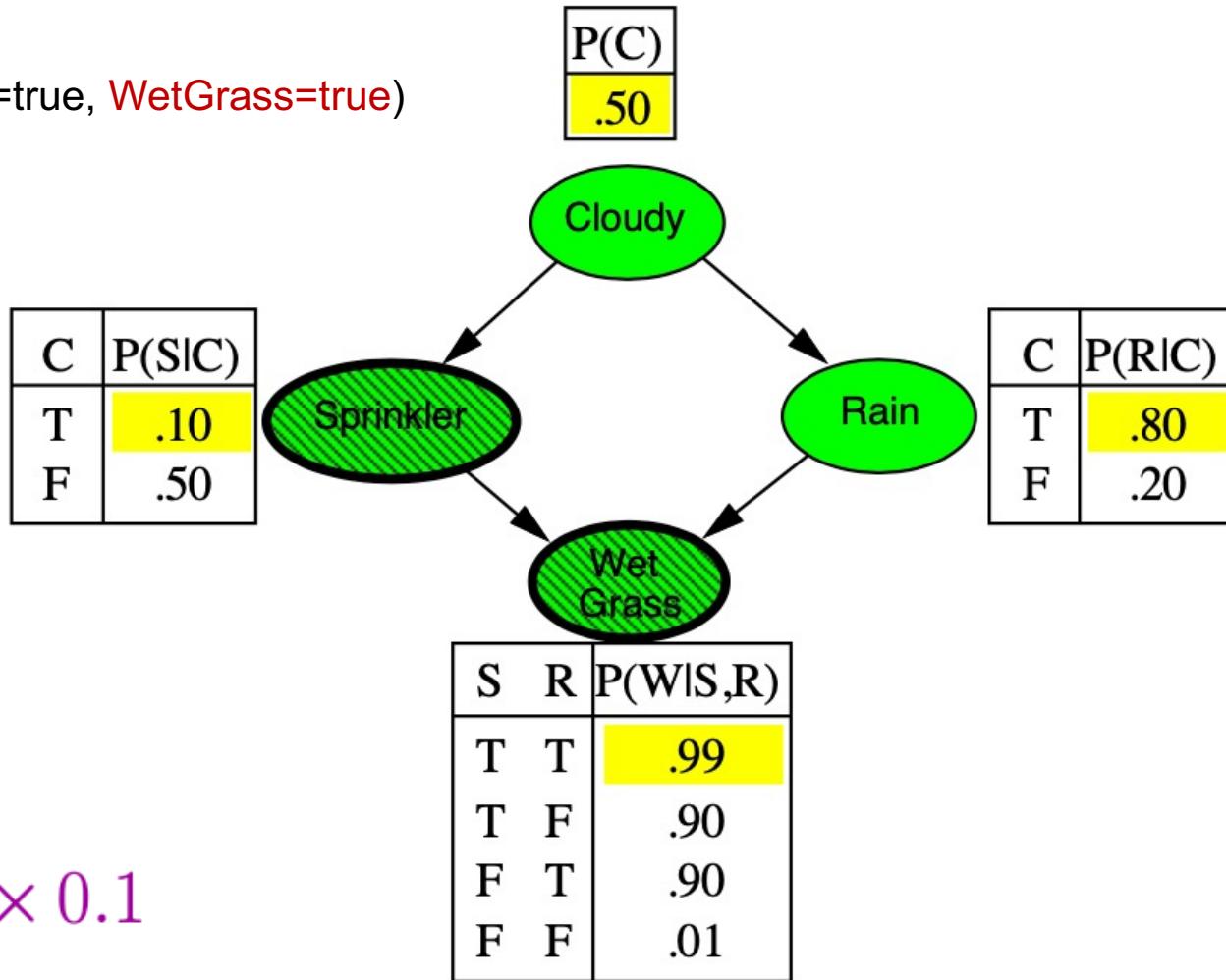


$$w = 1.0 \times 0.1$$

Pesatura di verosimiglianza

Query

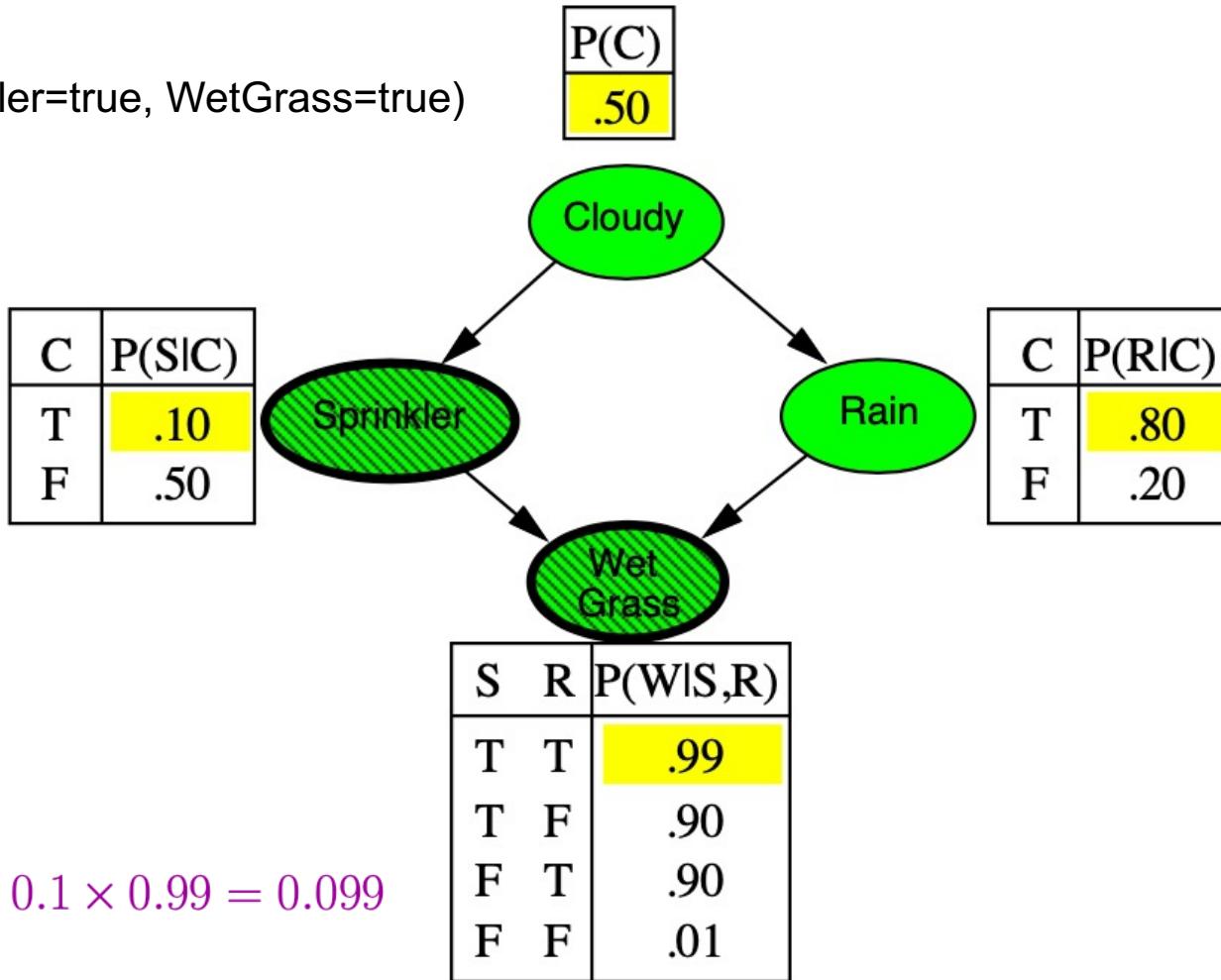
$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$



Pesatura di verosimiglianza

Query

$P(\text{Rain} \mid \text{Sprinkler}=\text{true}, \text{WetGrass}=\text{true})$



Analisi Pesatura di verosimiglianza

- ▶ La probabilità di campionamento per WeightedSample è

$$S_{WS}(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^l P(z_i | parents(Z_i))$$

- ▶ dove \mathbf{z} sono tutte le variabili tranne quelle di evidenza, mentre $parents(Z_i)$ può includere sia variabili nascoste che variabili di evidenza.
- ▶ Nota: presta attenzione alle evidenze solo negli **antenati** \Rightarrow ignorando le evidenze che non sono presenti tra gli antenati di Z_i

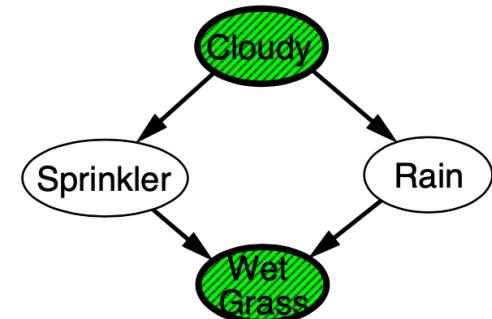
- ▶ Peso per un dato campione \mathbf{z}, \mathbf{e} è

$$w(\mathbf{z}, \mathbf{e}) = \prod_{i=1}^m P(e_i | parents(E_i))$$

- ▶ La probabilità pesata di un campione è

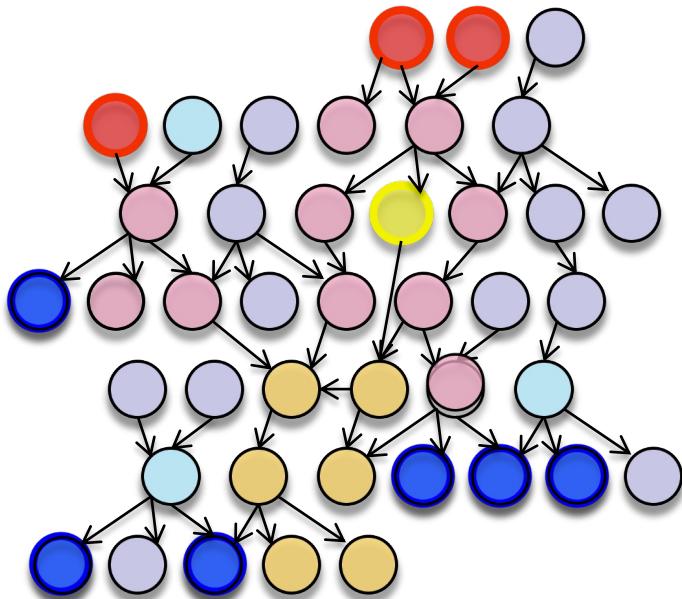
$$\begin{aligned} S_{WS}(\mathbf{z}, \mathbf{e}) w(\mathbf{z}, \mathbf{e}) \\ &= \prod_{i=1}^l P(z_i | parents(Z_i)) \prod_{i=1}^m P(e_i | parents(E_i)) \\ &= P(\mathbf{z}, \mathbf{e}) \text{ (by standard global semantics of network)} \end{aligned}$$

- ▶ Quindi la probabilità pesata restituisce stime CONSISTENTI ma le prestazioni peggiorano ancora con molte variabili di evidenza perché alcuni campioni hanno quasi tutto il peso totale

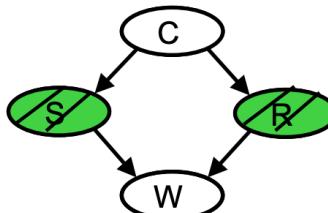


Analisi Pesatura di verosimiglianza

- ▶ Likelihood weighting è buona
 - ▶ Vengono utilizzati tutti i campioni
 - ▶ I valori delle variabili **a valle** sono influenzati dalle evidenze **a monte**



- Likelihood weighting presenta ancora punti deboli
 - I valori delle variabili **a monte** non sono influenzati dall'evidenza **a valle**
 - Ad esempio, supponiamo che la prova sia un video di un incidente stradale
 - Con evidenza in k nodi foglia, i pesi saranno $O(2^k)$
 - Con alta probabilità, un campione fortunato avrà un peso molto maggiore degli altri, dominando il risultato
- Vorremmo che ogni variabile "vedesse" **tutte** le prove!
 - C non ha maggiori probabilità di ottenere un valore corrispondente all'evidenza



Inferenza approssimativa usando MCMC

- ▶ "Stato corrente" della rete = assegnazione corrente a tutte le variabili.
- ▶ Genera lo stato successivo campionando una variabile data la coperta di Markov
- ▶ Campiona ciascuna variabile a turno, mantenendo le evidenze prefissate

```
function MCMC-Ask( $X, e, bn, N$ ) returns an estimate of  $P(X|e)$ 
  local variables:  $\mathbf{N}[X]$ , a vector of counts over  $X$ , initially zero
     $\mathbf{Z}$ , the nonevidence variables in  $bn$ 
     $\mathbf{x}$ , the current state of the network, initially copied from  $e$ 

  initialize  $\mathbf{x}$  with random values for the variables in  $\mathbf{Y}$ 
  for  $j = 1$  to  $N$  do
    for each  $Z_i$  in  $\mathbf{Z}$  do
      sample the value of  $Z_i$  in  $\mathbf{x}$  from  $P(Z_i|mb(Z_i))$ 
      given the values of  $MB(Z_i)$  in  $\mathbf{x}$ 
       $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$  where  $x$  is the value of  $X$  in  $\mathbf{x}$ 
  return NORMALIZE( $\mathbf{N}[X]$ )
```

- ▶ Puoi anche scegliere una variabile da campionare a caso ogni volta

Inferenza approssimativa usando MCMC

- ▶ "Stato corrente" della rete = assegnazione corrente a tutte le variabili.
- ▶ Genera lo stato successivo campionando una variabile data la coperta di Markov
- ▶ Campiona ciascuna variabile a turno, mantenendo le evidenze prefissate

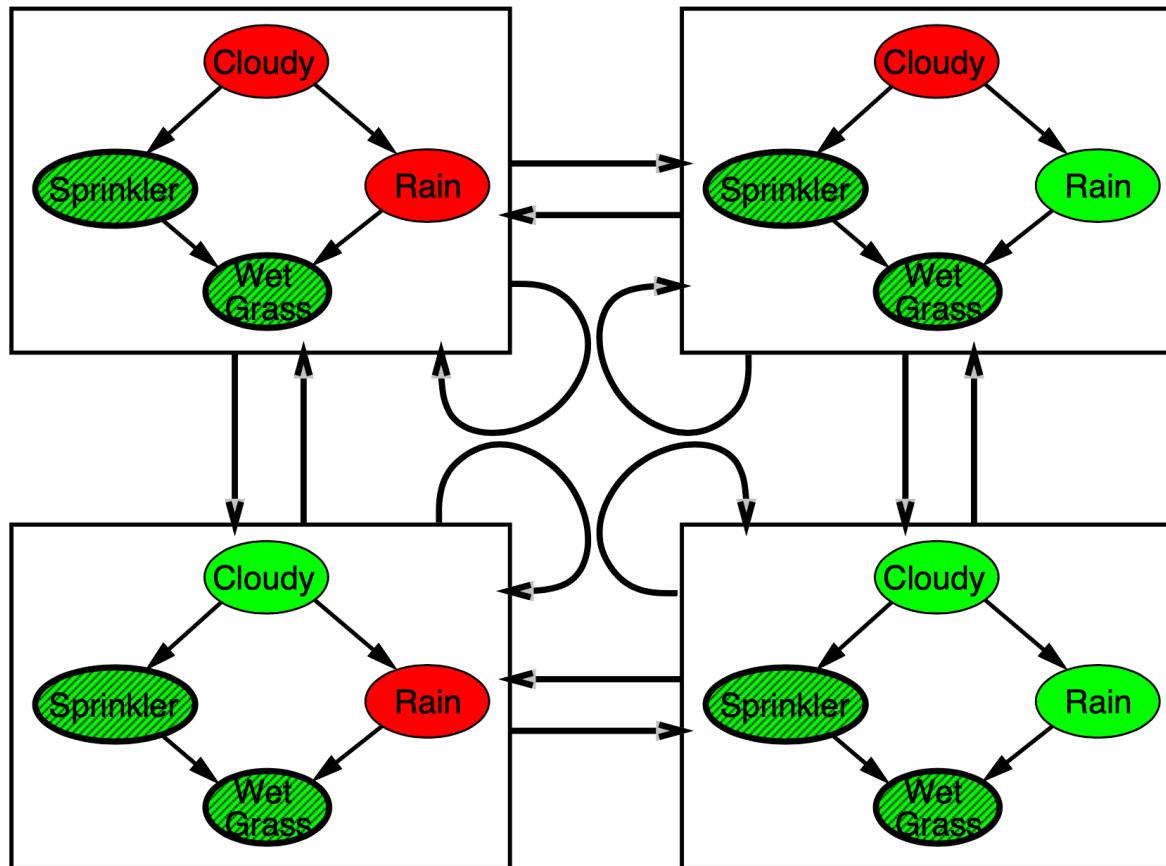
```
function MCMC-Ask( $X, e, bn, N$ ) returns an estimate of  $P(X|e)$ 
  local variables:  $\mathbf{N}[X]$ , a vector of counts over  $X$ , initially zero
     $\mathbf{Z}$ , the nonevidence variables in  $bn$ 
     $\mathbf{x}$ , the current state of the network, initially copied from  $e$ 

  initialize  $\mathbf{x}$  with random values for the variables in  $\mathbf{Y}$ 
  for  $j = 1$  to  $N$  do
    for each  $Z_i$  in  $\mathbf{Z}$  do
      sample the value of  $Z_i$  in  $\mathbf{x}$  from  $P(Z_i|mb(Z_i))$ 
      given the values of  $MB(Z_i)$  in  $\mathbf{x}$ 
       $\mathbf{N}[x] \leftarrow \mathbf{N}[x] + 1$  where  $x$  is the value of  $X$  in  $\mathbf{x}$ 
  return NORMALIZE( $\mathbf{N}[X]$ )
```

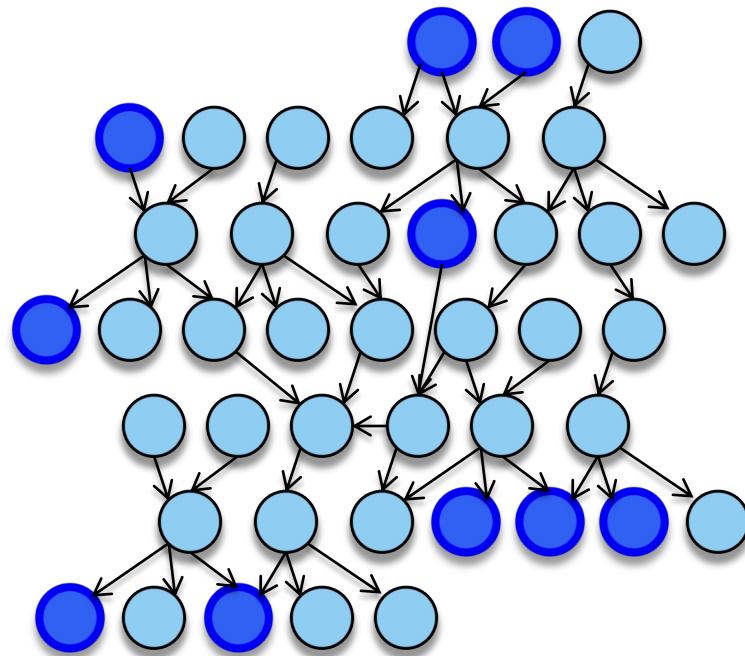
- ▶ Puoi anche scegliere una variabile da campionare a caso ogni volta

Le Catene di Markov

- Con $\text{Sprinkler} = \text{true}$, $\text{WetGrass} = \text{true}$ ci sono quattro stati



Perché qualcuno dovrebbe farlo?



Sia le variabili a monte che quelle a valle condizionano l'evidenza!

Al contrario: la pesatura di verosimiglianza condiziona solo sull'evidenza a monte, e quindi i pesi ottenuti nella pesatura di verosimiglianza possono a volte essere molto piccoli.

Esempio MCMC

- ▶ Stima $\mathbf{P}(\text{Rain} | \text{Sprinkler} = \text{true}, \text{WetGrass} = \text{true})$
- ▶ Campiona *Cloudy* o *Rain* data la sua coperta Markov, ripeti.
- ▶ Conta il numero di volte *Rain* è vera e falsa nei campioni.
- ▶ Ad esempio, 100 stati
 - ▶ 31 hanno *Rain* = true, 69 hanno *Rain* = false

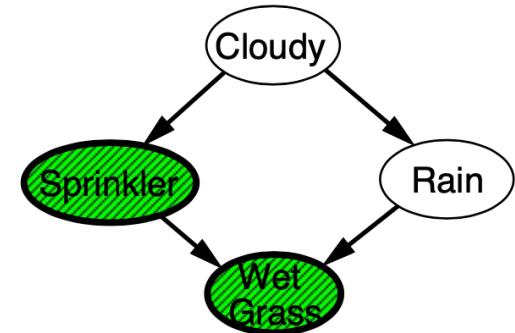
$$\hat{\mathbf{P}}(\text{Rain} | \text{Sprinkler} = \text{true}, \text{WetGrass} = \text{true}) \\ = \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$$

- ▶ Teorema: la catena si avvicina alla **distribuzione stazionaria**: la frazione di lungo periodo del tempo trascorso in ogni stato è esattamente proporzionale alla sua probabilità posteriore

Campionamento di coperte di Markov

- ▶ La coperta di Markov di *Cloudy* è *Sprinkler* e *Rain*
- ▶ La coperta di Markov di *Rain* è *Cloudy*, *Sprinkler* e *WetGrass*
- ▶ La probabilità data la coperta di Markov è calcolata come segue:

$$P(x'_i | mb(X_i)) = P(x'_i | parents(X_i)) \prod_{Z_j \in Children(X_i)} P(z_j | parents(Z_j))$$



- ▶ Principali problemi computazionali:
 - 1) Difficile dire se è stata raggiunta la convergenza
 - 2) Può essere dispendioso se la coperta di Markov è grande:
 $P(X_i | mb(X_i))$ non cambierà molto (legge dei grandi numeri)

Sommario

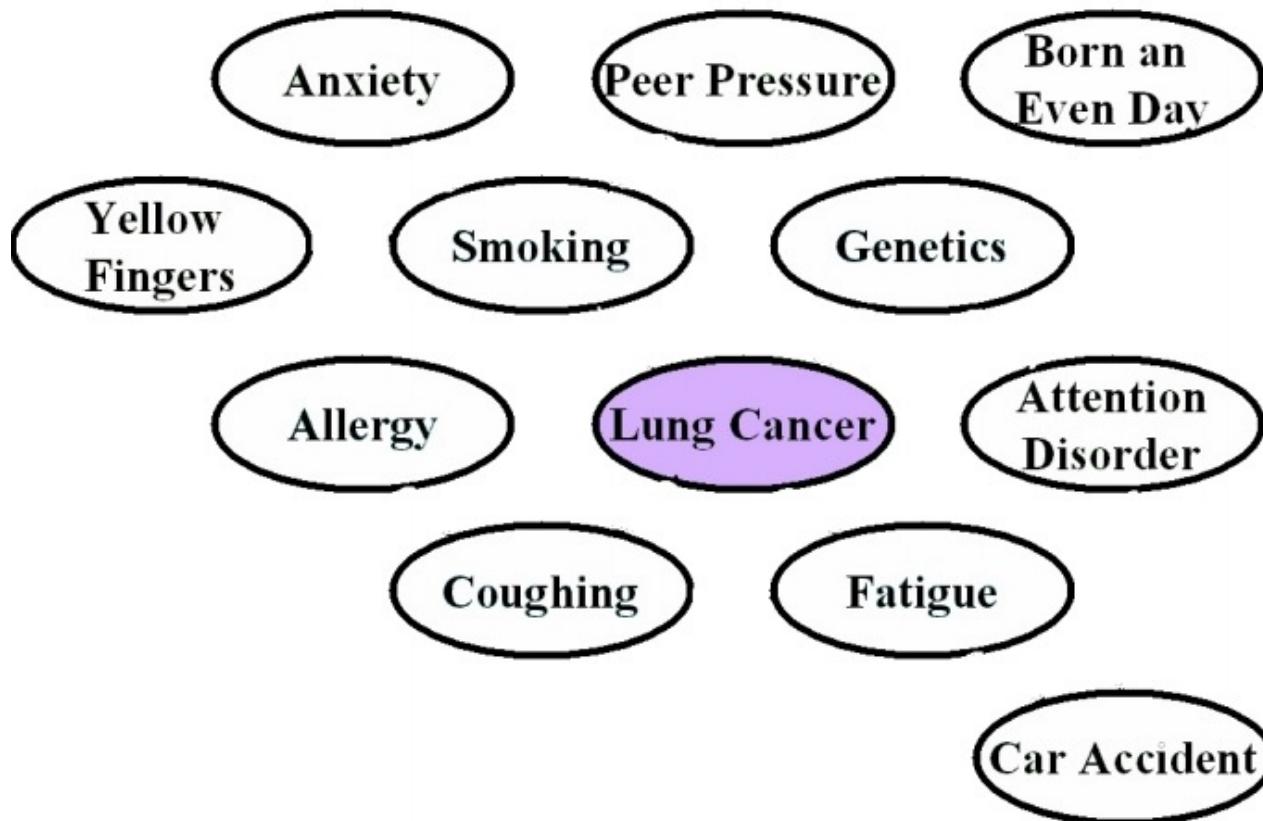
- ▶ Inferenza esatta per eliminazione variabile:
 - ▶ polinomiale su polialberi, NP-hard su grafi generici
 - ▶ spazio = tempo, molto sensibile alla topologia
- ▶ Inferenza approssimativa con LW, MCMC:
 - ▶ LW va male quando ci sono molte evidenze (a valle)
 - ▶ LW, MCMC generalmente insensibile alla topologia
 - ▶ La convergenza può essere molto lenta con probabilità vicine a 1 o 0
 - ▶ Può gestire combinazioni arbitrarie di variabili discrete e continue

Cosa hanno a che fare le BN con la Feature Selection?

- ▶ Label di classe, Lung cancer: YES/NO
- ▶ 11 features
 - ▶ X1 : patient has yellow fingers?
 - ▶ X2 : ...has anxiety?
 - ▶ X3 : ...smokes?
 - ▶ X4 : ...has experienced peer pressure?
 - ▶ X5 : ...born on an even-numbered day?
 - ▶ X6 : ...has allergies?
 - ▶ X7 : ...has genetic factors for cancer?
 - ▶ X8 : ...has concentration problems?
 - ▶ X9 : ...is coughing?
 - ▶ X10 : ... has fatigue?
 - ▶ X11 : ... recently in car accident?
- ▶ Alcune sono rilevanti altre irrilevanti

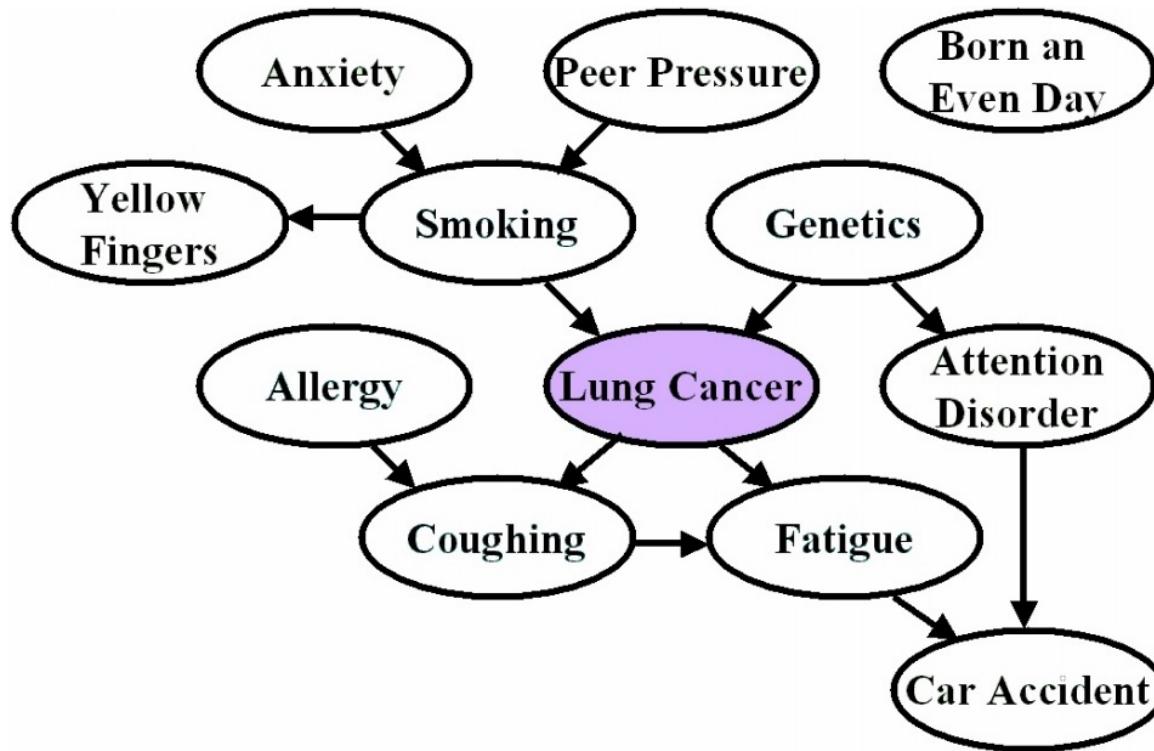
Problema di feature selection

- ▶ Di quali feature ho bisogno?



Learning Bayesian Networks

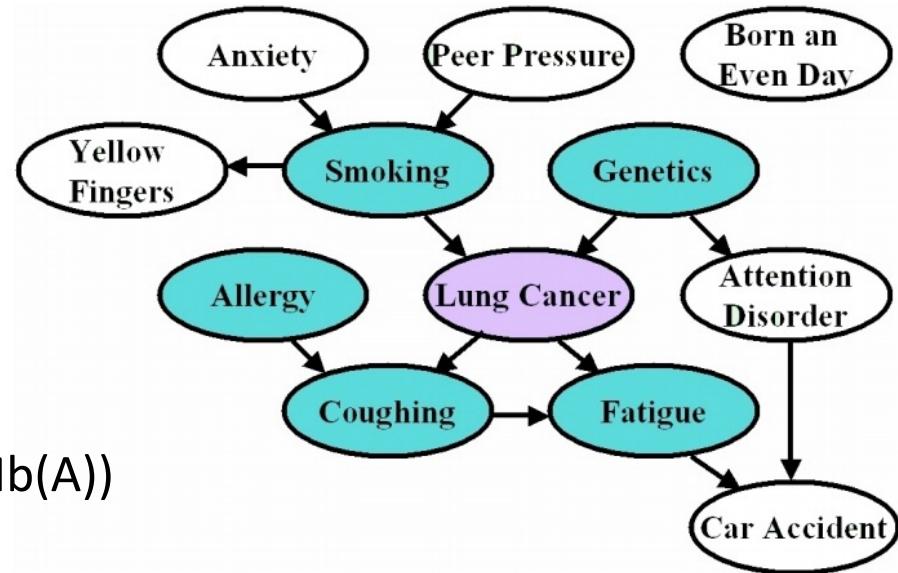
- ▶ Esistono algoritmi che apprendono dai dati la struttura della rete.



- ▶ Non tutte le distribuzioni possono essere rappresentate da una rete Bayesiana.

Coperta di Markov

- ▶ Le feature rilevanti sono la coperta di Markov del nodo Lung Cancer



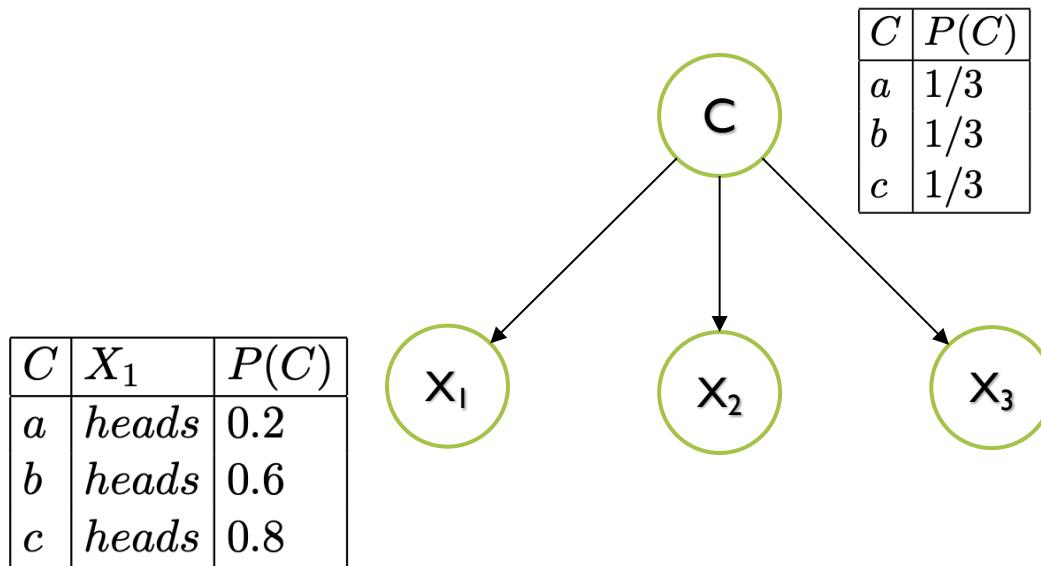
- ▶ $\Pr(A \mid (\text{Mb}(A), \text{any node})) = \Pr(A \mid \text{Mb}(A))$
- ▶ La coperta di Markov rende il nodo target condizionatamente indipendente da tutte le altre feature.
- ▶ Una volta noti i valori per la coperta di Markov, le altre feature possono cambiare senza influire sul target.

Esercizio

- ▶ Abbiamo un sacco con tre monete distorte a , b e c con probabilità di ottenere testa rispettivamente del 20%, 60% e 80%.
 - ▶ Una moneta viene estratta casualmente dal sacchetto (con la stessa probabilità di estrarre ciascuna delle tre monete), quindi la moneta viene lanciata tre volte per generare i risultati X_1 , X_2 e X_3 .
1. Disegna la rete bayesiana corrispondente a questa configurazione e definisci i CPT necessari.
 2. Calcola quale moneta era più probabile che fosse stata estratta dal sacchetto se i lanci osservati escono due volte testa e una volta croce.

Soluzione

- ▶ Abbiamo un sacco con tre monete distinte a , b e c con probabilità di ottenere testa rispettivamente del 20%, 60% e 80%.
- ▶ Una moneta viene estratta casualmente dal sacchetto (con la stessa probabilità di estrarre ciascuna delle tre monete), quindi la moneta viene lanciata tre volte per generare i risultati X_1 , X_2 e X_3 .



Soluzione

- Calcola quale moneta sia stata estratta dal sacchetto con più probabilità se i lanci osservati sono due volte testa e una volta croce.

▶ $P(C | 2 \text{ testa}, 1 \text{ croce}) = P(2 \text{ testa}, 1 \text{ croce} | C) P(C) / P(2 \text{ testa}, 1 \text{ croce})$

$$= \alpha P(2 \text{ testa}, 1 \text{ croce} | C) P(C)$$

$$= \alpha P(2 \text{ testa}, 1 \text{ croce} | C)$$

▶ $P(X_1 = \text{croce}, X_2 = \text{testa}, X_3 = \text{testa} | C=a) =$

▶ $P(X_1 = \text{croce} | C=a) P(X_2 = \text{testa} | C=a) P(X_3 = \text{testa} | C=a) =$

▶ $0.8 \times 0.2 \times 0.2 = 0.032$

▶ $P(2 \text{ testa}, 1 \text{ croce} | C=a) = 3 \times 0.032 = 0.096$

▶ $P(X_1 = \text{croce} | C=b) P(X_2 = \text{testa} | C=b) P(X_3 = \text{testa} | C=b) =$

▶ $0.4 \times 0.6 \times 0.6 = 0.144 \quad P(2 \text{ testa}, 1 \text{ croce} | C=b) = 3 \times 0.144 = 0.432$

▶ $P(X_1 = \text{croce} | C=c) P(X_2 = \text{testa} | C=c) P(X_3 = \text{testa} | C=c) = 0.2 \times 0.8 \times 0.8 = 0.128$

▶ $P(2 \text{ testa}, 1 \text{ croce} | C=c) = 3 \times 0.128 = 0.384$

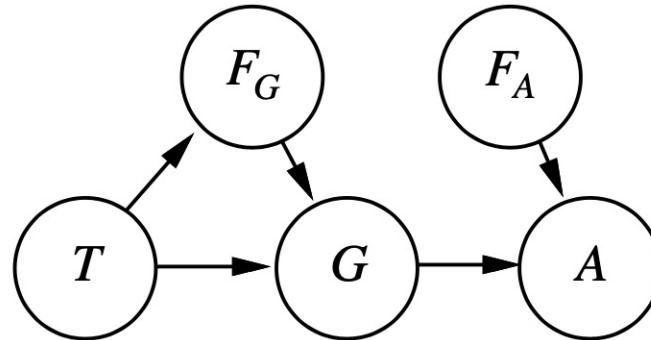
C	X_1	$P(C)$
a	<i>heads</i>	0.2
b	<i>heads</i>	0.6
c	<i>heads</i>	0.8

Esercizio

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Disegna una rete bayesiana per questo dominio, dato che è più probabile che l'indicatore fallisce quando la temperatura interna diventa troppo elevata.

Soluzione

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Disegna una rete bayesiana per questo dominio, dato che è più probabile che l'indicatore fallisce quando la temperatura interna diventa troppo elevata.



Esercizio

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Supponiamo che ci siano solo due possibili temperature effettive e misurate, *Normal* e *High*; la probabilità che l'indicatore dia la temperatura corretta è x quando sta funzionando, ma y quando è difettoso. Fornisci la tabella delle probabilità condizionale associata a G .

Soluzione

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Supponiamo che ci siano solo due possibili temperature effettive e misurate, *Normal* e *High*; la probabilità che l'indicatore dia la temperatura corretta è x quando sta funzionando, ma y quando è difettoso. Fornisci la tabella delle probabilità condizionale associata a G .

		$T = \text{Normal}$		$T = \text{High}$	
		F_G	$\neg F_G$	F_G	$\neg F_G$
$G = \text{Normal}$	y	x	$1 - y$	$1 - x$	
	$1 - y$	$1 - x$	y	x	

Esercizio

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Supponiamo che l'allarme funzioni correttamente a meno che non sia difettoso, nel qual caso non suona mai. Indica la tabella delle probabilità condizionale associata ad A .

Soluzione

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Supponiamo che l'allarme funzioni correttamente a meno che non sia difettoso, nel qual caso non suona mai. Indica la tabella delle probabilità condizionale associata ad A .

		$G = Normal$		$G = High$	
		F_A	$\neg F_A$	F_A	$\neg F_A$
A	0	0	0	1	
	1	1	1	0	

Esercizio

- ▶ In una centrale nucleare c'è un allarme che rileva quando un indicatore di temperatura supera una determinata soglia. L'indicatore misura la temperatura del nucleo. Considerare le variabili booleane A (suona l'allarme), F_A (l'allarme è difettoso) e F_G (l'indicatore è difettoso), e i nodi multivalore G (lettura dell'indicatore) e T (temperatura effettiva del nucleo).
- ▶ Supponiamo che $P(T) = p$, $P(F_G | T) = g$, and $P(F_G | \sim T) = h$.
- ▶ Supponiamo che l'allarme e l'indicatore funzionino e che l'allarme suoni. Mostrare come calcolare la probabilità che la temperatura interna sia troppo elevata, usando l'inferenza per enumerazione. (La risposta può includere termini di sommatoria e può includere le probabilità indicate sopra, ovvero x , y , p , g e h .)

$$P(T | a, \neg f_G, \neg f_A) = \alpha \sum_g P(T, a, \neg f_G, \neg f_A, g)$$

$$\begin{aligned} P(t | a, \neg f_G, \neg f_A) &= \alpha \sum_g P(t) P(a|g, \neg f_A) P(g|\neg f_G, \neg f_A) P(\neg f_A) P(\neg f_G | t) \\ &= \alpha P(t) P(\neg f_A) P(\neg f_G | t) \sum_g P(a|g, \neg f_A) P(g|\neg f_G, t) \\ &= \alpha p (1-g) P(\neg f_A) [P(a|g, \neg f_A) P(g|\neg f_G, t) + P(a|\neg g, \neg f_A) P(\neg g|\neg f_G, t)] \\ &= \alpha p (1-g) P(\neg f_A) [1(x) + 0(1-x)] \\ &= \alpha p (1-g) P(\neg f_A) x \end{aligned}$$

$$\begin{aligned} \alpha &= P(a, \neg f_G, \neg f_A) = \sum_t \sum_g P(t) P(a|g, \neg f_A) P(g|\neg f_G, \neg f_A) P(\neg f_A) P(\neg f_G | t) \\ &= P(\neg f_A) \sum_t P(t) P(\neg f_G | t) \sum_g P(a|g, \neg f_A) P(g|\neg f_G, t) \\ &= P(\neg f_A) (p(1-g) + (1-p)(1-h)) [1(x) + 0(1-x)] = x P(\neg f_A) (p (1-g) + (1-p)(1-h)) \end{aligned}$$

$$P(t | a, \neg f_G, \neg f_A) = p (1-g) / (p (1-g) + (1-p)(1-h))$$