# Incertezza

## Capitolo 13

*– docente: Alessandro Sperduti –*

*– presentazione basata sui lucidi di S. Russell –*

# Incertezza

Azione $A_t$ = partire per l'aeroporto $t$ minuti prima del volo
L'azione $A_t$ mi permetterà di arrivare in tempo?

Problemi:

1) osservabilità parziale (stato della strada, piano di altri veicoli, etc.)
2) sensori rumorosi (rapporti sul traffico di Isoradio ...)
3) incertezza nell'esito delle azioni (pneumatico forato, etc.)
4) immensa complessità nel modellare e nel predire il traffico

Quindi un approccio puramente logico o

1) rischia di dire il falso: "$A_{25}$ mi fa arrivare in tempo"

o 2) conduce a conclusioni che sono troppo deboli per prendere decisioni:

"$A_{25}$ mi fa arrivare in tempo se non c''e un incidente sul ponte
e non piove e non foro i pneumatici, etc etc."

($A_{1440}$ si può ragionevolmente dire che mi fa arrivare in tempo
ma devo passare la notte all'aeroporto ...)

# Probabilità

Asserzioni Probabilistiche *riassumono* gli effetti di

pigrizia: fallimento nell'enumerare le eccezioni, qualifica, etc.

ignoranza: mancanza di fatti rilevanti, condizioni iniziali, etc.

Probabilità Soggettiva o Bayesiana:

Le probabilità legano le proposizioni al proprio stato di conoscenza

p.e., $P(A_{25}|\text{nessun incidente riportato}) = 0.06$

*Non* indica alcuna tendenza probabilistica nella situazione corrente

Le probabilità delle proposizioni cambiano con l'arrivo di nuova evidenza:

p.e., $P(A_{25}|\text{nessun incidente riportato, 5 a.m.}) = 0.15$

# Decidere nell'incertezza

Supponiamo di credere che:

$$
\begin{aligned}
P(A_{25} \text{ mi fa arrivare in tempo} | \ldots) &= 0.04 \\
P(A_{90} \text{ mi fa arrivare in tempo} | \ldots) &= 0.70 \\
P(A_{120} \text{ mi fa arrivare in tempo} | \ldots) &= 0.95 \\
P(A_{1440} \text{ mi fa arrivare in tempo} | \ldots) &= 0.9999
\end{aligned}
$$

Quale azione scegliere ?

Dipende dalle mie preferenze: perdere l'aereo vs. la cucina dell'areoporto, etc.

Teoria dell'utilità è usata per rappresentare e inferire preferenze

Teoria delle decisioni = teoria dell'utilità + teoria delle probabilità

# Probabilità a priori

Prior o probabilità incondizionate di proposizioni

p.e., $P(\mathit{Cavità} = vero) = 0.1$ e $P(Tempo = sole) = 0.72$

corrispondono a gradi di credenza sull'arrivo di (nuova) evidenza

Distribuzione di probabilità fornisce valori per tutti i possibili assegnamenti:

p.e., se $Tempo$ può assumere i valori $\langle sole, pioggia, nuvole, neve \rangle$,

$\mathbf{P}(Tempo) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (*normalizzate*, cioè, sommano ad 1)

Distribuzione di probabilità congiunta per un insieme di v.a. fornisce la probabilità per ogni evento atomico su tali variabili

$\mathbf{P}(Tempo, \mathit{Cavità})$ = una matrice $4 \times 2$ di valori:

| $Tempo =$ | $sole$ | $pioggia$ | $nuvole$ | $neve$ |
|---|---|---|---|---|
| $\mathit{Cavità} = vero$ | 0.144 | 0.02 | 0.016 | 0.02 |
| $\mathit{Cavità} = falso$ | 0.576 | 0.08 | 0.064 | 0.08 |

*Ogni domanda che concerne un dominio trova risposta nella distribuzione congiunta perché ogni evento è la somma dei possibili eventi*

# Probabilità condizionale

Probabilità a posteriori o condizionale

p.e., $P(\textit{Cavità}|Mal\_di\_denti) = 0.8$

cioè, dato che $Mal\_di\_denti$ è tutto quello che so

Se so di più, p.e., *Cavità* è anche data, allora abbiamo

$P(\textit{Cavità}|Mal\_di\_denti, \textit{Cavità}) = 1$

Nota: la credenza meno specifica *rimane valida* dopo che nuova evidenza arriva, ma non necessariamente rimane *utile*

Nuova evidenza può essere irrilevante, permettendo semplificazioni, p.e.,

$P(\textit{Cavità}|Mal\_di\_denti, Vince\_Inter) = P(\textit{Cavità}|Mal\_di\_denti) = 0.8$

Questo tipo di inferenza, dovuta alla conoscenza del dominio, è cruciale

# Probabilità condizionale

Definizione di probabilità condizionale:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

La regola del prodotto fornisce una definizione alternativa:

$$P(a \wedge b) = P(a|b)P(b) = p(b|a)P(a)$$

Una versione generale vale sulle distribuzioni, p.e.,

$$\mathbf{P}(Tempo, \textit{Cavità}) = \mathbf{P}(Tempo|\textit{Cavità})\mathbf{P}(\textit{Cavità})$$

(Visto come un insieme $4 \times 2$ di equazioni, *no* moltiplicazione di matrici)

Chain rule è derivata dalla applicazione ripetuta della regola del prodotto:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})\end{aligned}$$

# Inferenza tramite enumerazione

Si inizia con la distribuzione congiunta:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | **.108** | **.012** | **.072** | **.008** |
| ¬ *cavity* | **.016** | **.064** | **.144** | **.576** |

Per ogni proposizione $\phi$, si sommano gli eventi atomici dove essa è vera:
$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

# Inferenza tramite enumerazione

Si inizia con la distribuzione congiunta:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | **.108** | **.012** | **.072** | **.008** |
| ¬ *cavity* | **.016** | **.064** | **.144** | **.576** |

Per ogni proposizione $\phi$, si sommano gli eventi atomici dove essa è vera:
$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

# Inferenza tramite enumerazione

Si inizia con la distribuzione congiunta:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | **.108** | **.012** | **.072** | **.008** |
| ¬ *cavity* | **.016** | **.064** | **.144** | **.576** |

Per ogni proposizione $\phi$, si sommano gli eventi atomici dove essa è vera:
$P(\phi) = \sum_{\omega:\omega \models \phi} P(\omega)$

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

# Inferenza tramite enumerazione

Si inizia con la distribuzione congiunta:

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | **.108** | **.012** | **.072** | **.008** |
| ¬ *cavity* | **.016** | **.064** | **.144** | **.576** |

Si possono calcolare anche le probabilità condizionali:

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

# Normalizzazione

| | *toothache* | | ¬ *toothache* | |
|---|---|---|---|---|
| | *catch* | ¬ *catch* | *catch* | ¬ *catch* |
| *cavity* | **.108** | **.012** | **.072** | **.008** |
| ¬ *cavity* | **.016** | **.064** | **.144** | **.576** |

Il denominatore può essere visto come una *costante di normalizzazione* $\alpha$

$$\begin{aligned}
\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache) \\
&= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

Idea generale: calcolare la distribuzione sulla variabile della query fissando le variabili di evidenza e sommando sulle variabili nascoste

# Inferenza tramite enumerazione

Tipicamente siamo interessati a
la distribuzione congiunta a posteriori delle variabili di query $\mathbf{Y}$
dati specifici valori $\mathbf{e}$ per le variabili di evidenza $\mathbf{E}$

Poniamo le variabili nascoste essere $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Allora la somma desiderata di entrate congiunte è ottenuta sommando sulle variabili nascoste:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

I termini nella sommatoria sono entrate congiunte perché $\mathbf{Y}$, $\mathbf{E}$, e $\mathbf{H}$ insieme esauriscono l'insieme delle variabili aleatorie

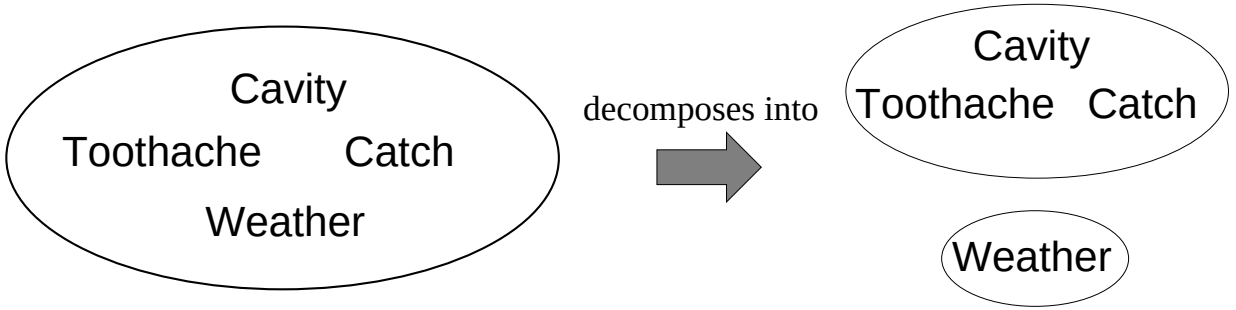
Problemi ovvi:

1) Complessità caso pessimo in tempo $O(d^n)$ dove $d$ è l'arietà più grande
2) Complessità in spazio $O(d^n)$ per memorizzare la distribuzione congiunta
3) Come stabilire i valori per $O(d^n)$ entrate???

# Indipendenza

$A$ e $B$ sono indipendenti sse

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$




$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$
$$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

32 entrate ridotte a 12; per $n$ monete "truccate" indipendenti, $2^n \rightarrow n$

Indipendenza assoluta potente ma rara

Nei problemi reali sono coinvolte centinaia di variabili,
nessuna delle quali è indipendente. Che fare?

# Indipendenza condizionale

$\mathbf{P}(Toothache, Cavity, Catch)$ ha $2^3 - 1 = 7$ entrate indipendenti

Se si ha una cavità, la probabilità che la sonda si fermi su in essa non dipende dal fatto di avere il mal di denti:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

La stessa indipendenza vale se non c'è la cavità:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ è *condizionalmente indipendente* da $Toothache$ dato $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Affermazione equivalente:

$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$

$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

# Indipendenza condizionale

Scrivere la distribuzione congiunta completa usando la chain rule:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity) \\
&= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

Cioè, $2 + 2 + 1 = 5$ numeri indipendenti

*In molti casi, l'uso di indipendenza condizionale riduce la dimensione della rappresentazione della probabilità congiunta da essere esponenziale in $n$ a lineare $n$.*

*L'indipendenza condizionale è la forma più basilare e robusta di conoscenza sugli ambienti incerti.*

# Regola di Bayes

Regola del prodotto $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Rightarrow \text{ Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

o in forma di distribuzione

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Utile per ottenere probabilità diagnostica a partire da probabilità causale:

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

P.e., sia $M$ la rappresentazione di Meningite, e $S$ di collo rigido:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Nota: la probabilità a posteriori della Meningite ancora piccola!

# Regola di Bayes e indipendenza condizionale

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache \wedge catch) \\
&= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity) \\
&= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

Questo è un esempio di modello *naive Bayes*:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$


Il numero totale di parametri è *lineare* in $n$

# Mondo dei Wumpus

| | | | |
|---|---|---|---|
| 1,4 | 2,4 | 3,4 | 4,4 |
| 1,3 | 2,3 | 3,3 | 4,3 |
| 1,2 **B** **OK** | 2,2 | 3,2 | 4,2 |
| 1,1 **OK** | 2,1 **B** **OK** | 3,1 | 4,1 |

$P_{ij} = vero$ sse $[i, j]$ contiene una trappola

$B_{ij} = vero$ sse $[i, j]$ è ventilato
Includiamo solo $B_{1,1}, B_{1,2}, B_{2,1}$ nel modello probabilistico

# Specifica del modello probabilistico

La distribuzione congiunta completa è $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$

Applicare la regola del prodotto: $\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$

(facciamo così per ottenere $P(Effect|Cause)$.)

Primo termine: 1 se le trappole sono adiacenti a "brezze", 0 altrimenti

Secondo termine: le trappole sono posizionate a caso, con probabilità 0.2 per quadrato:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \Pi_{i,j=1,1}^{4,4}\mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

per $n$ trappole.

# Osservazioni e query

Noi conosciamo i seguenti fatti:

$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

La query è $\mathbf{P}(P_{1,3}|known, b)$

Definiamo $unkown =$ i $P_{ij}$ diversi da $P_{1,3}$ e $known$

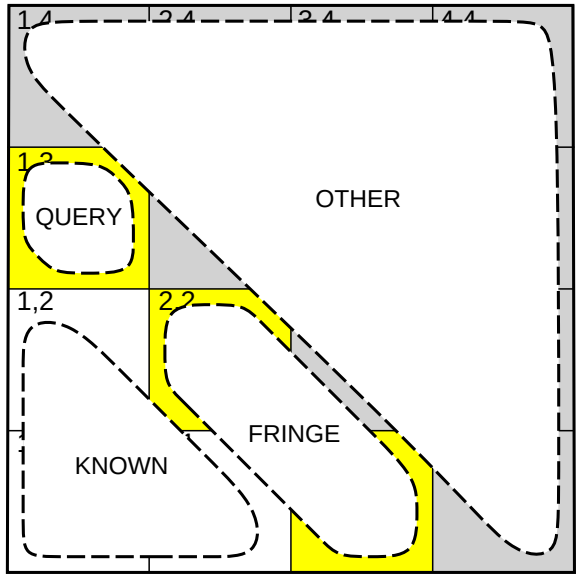
Per effettuare inferenza per enumerazione, abbiamo

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unkown} \mathbf{P}(P_{1,3}, unkown, known, b)$$

Cresce esponenzialmente con il numero di quadrati!

# Usando l'indipendenza condizionale

Idea base: le osservazioni sono condizionalmente indipendenti da altri quadrati nascosti dati i quadrati nascosti adiacenti



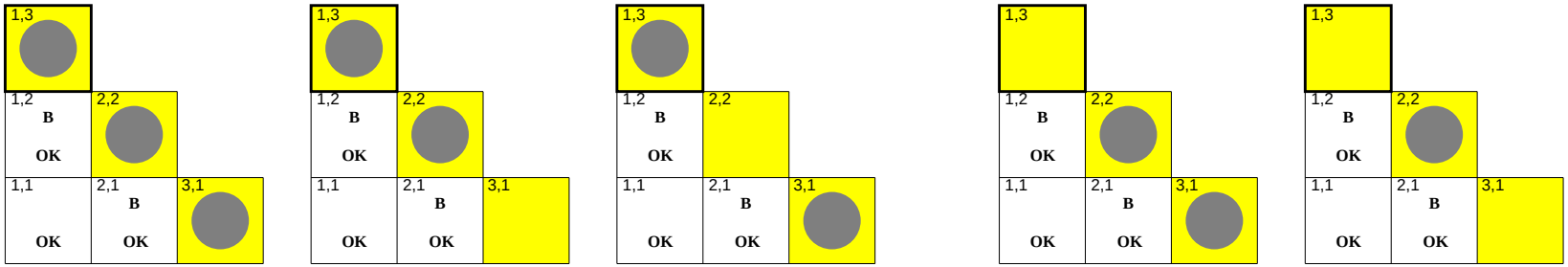

Definiamo $Unkown = Fringe \cup Other$
$\mathbf{P}(b|P_{1,3}, Known, Unkown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$

Poniamo la query in una forma dove si possa usare quanto sopra!

# Usando l'indipendenza condizionale

$$
\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unkown} \mathbf{P}(P_{1,3}, unkown, known, b) \\
&= \alpha \sum_{unkown} \mathbf{P}(b|P_{1,3}, known, unkown)\mathbf{P}(P_{1,3}, known, unkown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\
&= \alpha'\, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)
\end{aligned}
$$

# Usando l'indipendenza condizionale




$$\mathbf{P}(P_{1,3}|known, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle$$
$$\approx \langle 0.31, 0.69 \rangle$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

# Riassunto

Il calcolo delle probabilità costituisce un formalismo rigoroso per la conoscenza incerta

La distribuzione congiunta di probabilità specifica la probabilità di ogni evento atomico

Si può rispondere alle query sommando sugli eventi atomici

Per domini non banali, bisogna trovare un modo per ridurre la dimensione della rappresentazione della probabilità congiunta

Independenza ed indipendenza condizionale forniscono tale modo

# Reti Bayesiane

## Capitolo 14

*– docente: Alessandro Sperduti –*

*– presentazione basata sui lucidi di S. Russell –*

# Outline

- ♢ Sintassi
- ♢ Semantica
- ♢ Inferenza esatta tramite enumerazione
- ♢ Inferenza esatta tramite eliminazione di variabile
- ♢ Inferenza approssimata tramite simulazione stocastica

# Reti Bayesiane (Bayesian networks)

Una semplice notazione grafica per asserzioni condizionalmente indipendenti e quindi per specifiche di distribuzioni condizionali complete

Sintassi:

un insieme di nodi, uno per variabile

un grafo diretto aciclico (link $\approx$ "influenza direttamente")

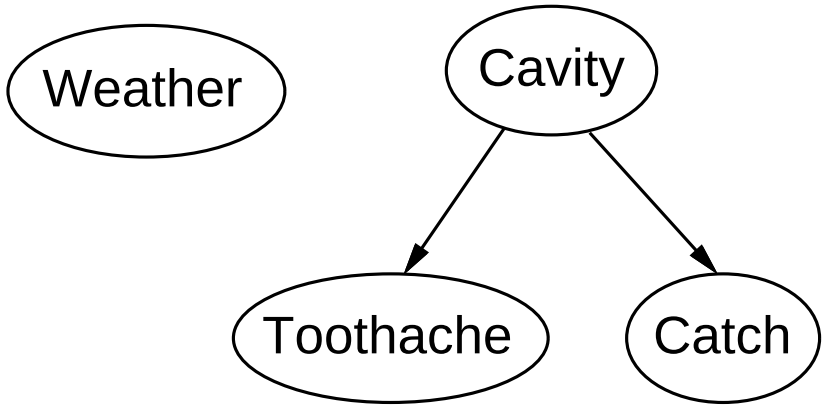
una distribuzione condizionale per ogni nodo dati i suoi genitori:

$\mathbf{P}(X_i|Parents(X_i))$

Nel caso più semplice, distribuzione condizionale rappresentata come una tabella della probabilità condizionale (CPT) data la distribuzione su $X_i$ per ogni combinazione di valori assunti dai genitori

# Esempio

La topologia della rete codifica asserzioni di indipendenza condizionale:




$Weather$ è indipendente dalle altre variabili

$Toothache$ e $Catch$ sono condizionalmente indipendenti data $Cavity$

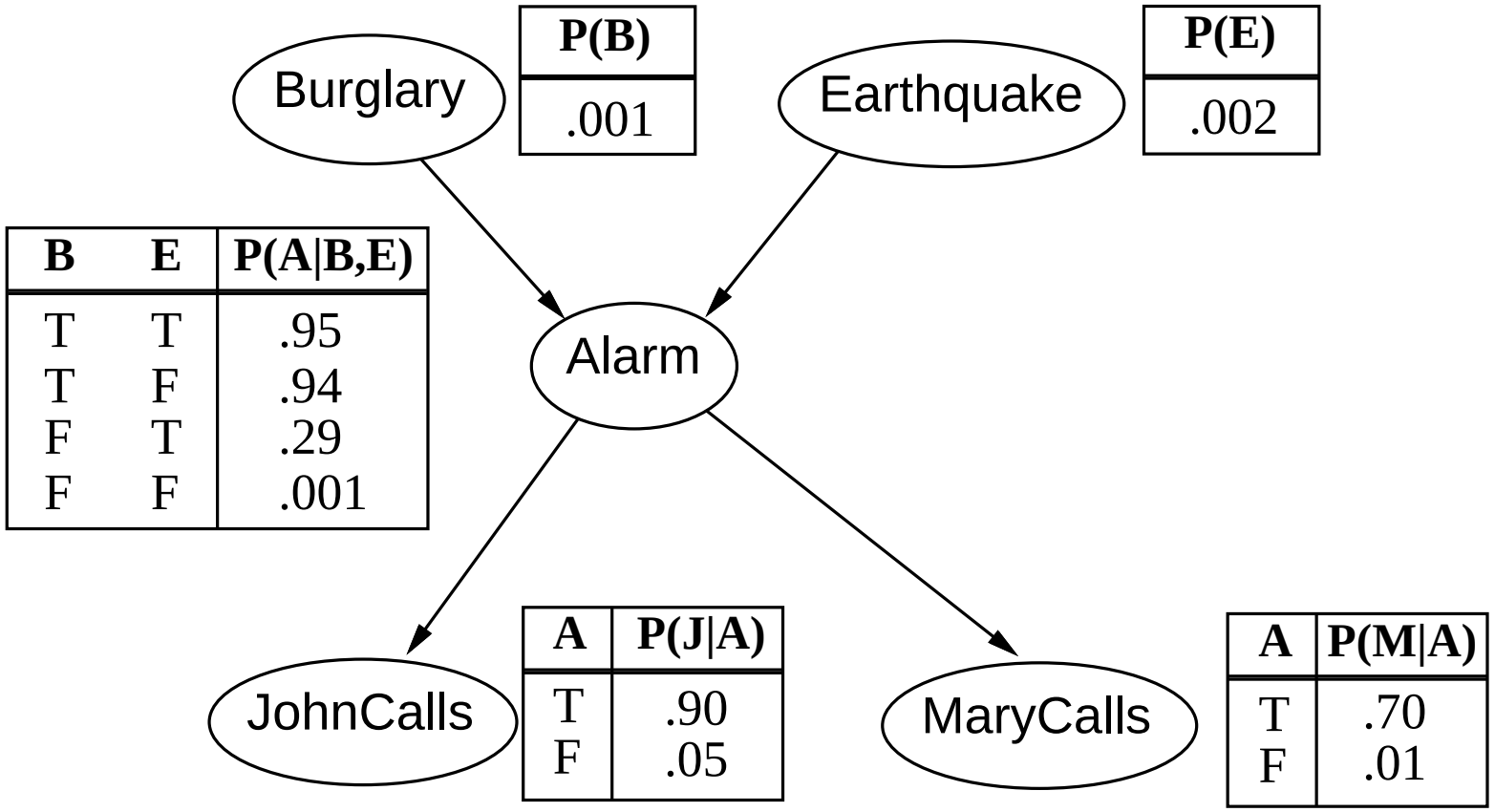
# Esempio

Sono al lavoro, il vicino John chiama per dire che il mio allarme $Alarm$ è entrato in funzione, ma la vicina Mary non chiama. Alcune volte l'allarme è attivato da piccole scosse di terremoto. C'è un ladro in casa ?

Variabili: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
La topologia della rete riflette conoscenza "causale":

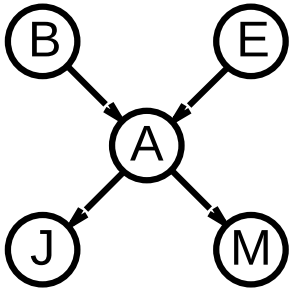
- Un ladro può attivare l'allarme
- Un terremoto può attivare l'allarme
- L'attivazione dell'allarme può indurre Mary a chiamare
- L'attivazione dell'allarme può indurre John a chiamare

# Esempio


Burglary
Earthquake
Alarm
JohnCalls
MaryCalls
P(B)
.001
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
A P(J|A)
T .90
F .05
A P(M|A)
T .70
F .01

# Compattezza

Una CPT per variabili Booleane $X_i$ con $k$ genitori Booleani ha $2^k$ righe per le combinazioni di valori dei genitori




Ogni riga richiede un numero $p$ per $X_i = vero$
(il numero per $X_i = falso$ è $1 - p$)

Se ogni variabile non ha più di $k$ genitori,
la rete completa richiede $O(n \cdot 2^k)$ numeri

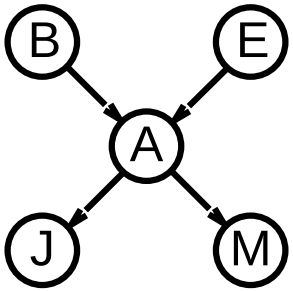
Cioè, cresce linearmente con $n$, vs. $O(2^n)$ per la distribuzione congiunta completa

Per la rete precedente, $1 + 1 + 4 + 2 + 2 = 10$ numeri (vs. $2^5 - 1 = 31$)

# Semantica globale

La semantica globale definisce la distribuzione congiunta completa come il prodotto delle distribuzioni condizionali locali:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$




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

$=$

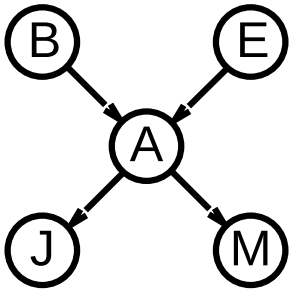
# Semantica globale

La semantica globale definisce la distribuzione congiunta completa come il prodotto delle distribuzioni condizionali locali:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

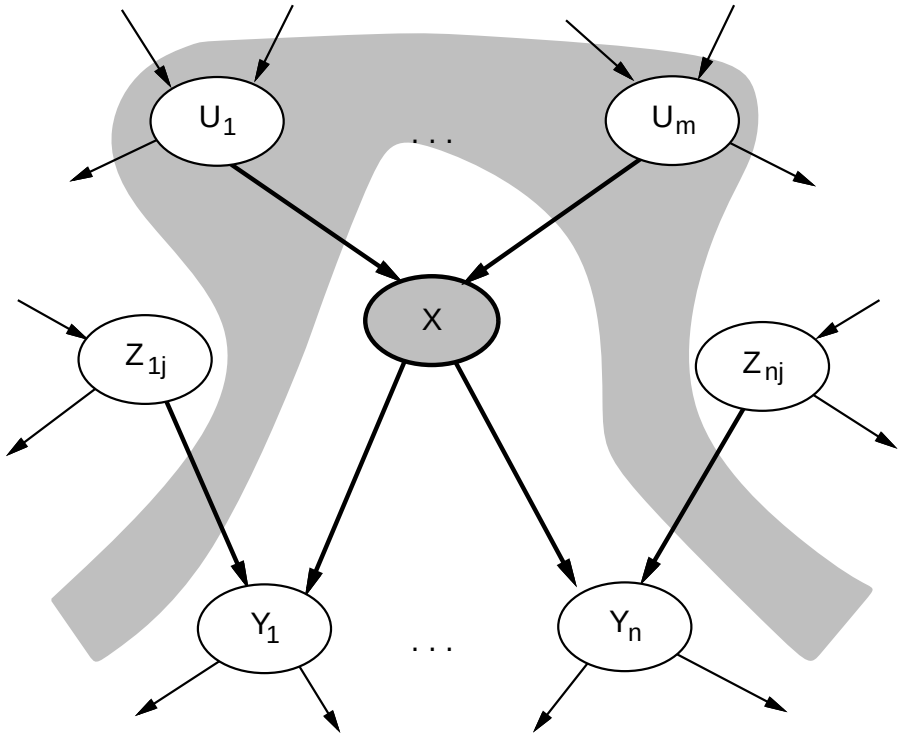
p.e., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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



# Semantica locale

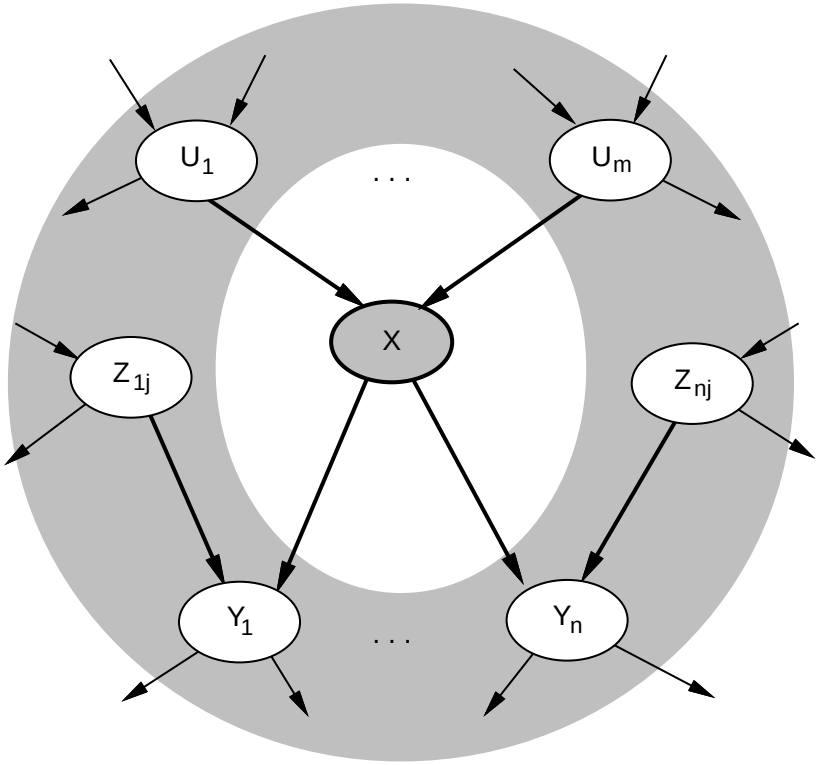
Semantica locale: ogni nodo è condizionalmente indipendente dai suoi non discendenti dati i genitori




Teorema: Semantica locale $\Leftrightarrow$ semantica globale

# Markov blanket

Ogni nodo è condizionalmente indipendente da tutti gli altri dato il suo Markov blanket: genitori + figli + genitori dei figli



# Costruzione di Reti Bayesiane

Necessità di un metodo tale che data una serie di asserzioni di indipendenza condizionale localmente controllabili, garantisca la semantica globale desiderata

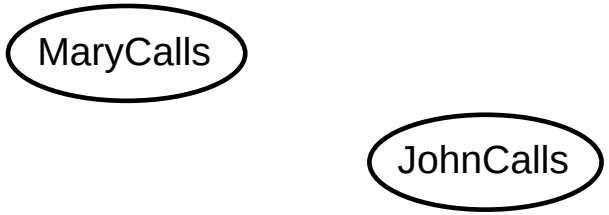
1. Scegliere un ordinamento di variabili $X_1, \ldots, X_n$
2. For $i = 1$ to $n$
   aggiungi $X_i$ alla rete
   seleziona genitori da $X_1, \ldots, X_{i-1}$ tali che
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

Questa scelta di genitori garantisce la semantica globale:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i)) \quad \text{(per costruzione)}\end{aligned}$$

# Esempio

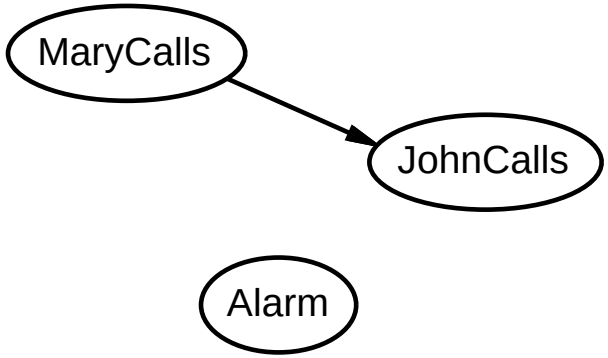
Supponiamo di scegliere l'ordine $M$, $J$, $A$, $B$, $E$




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

# Esempio

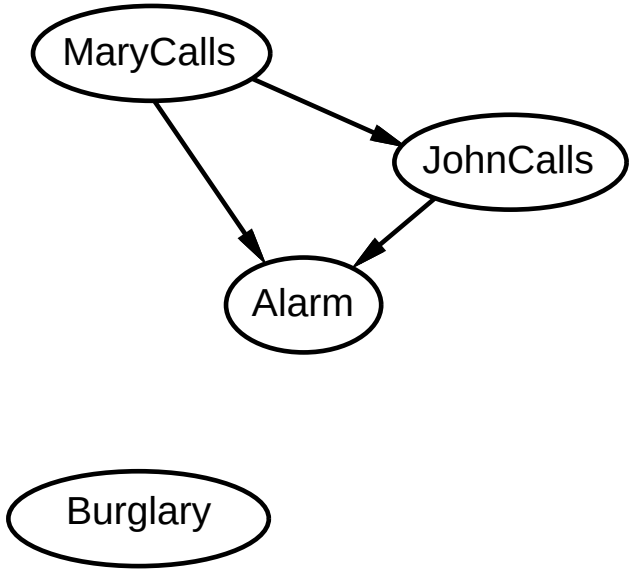
Supponiamo di scegliere l'ordine $M$, $J$, $A$, $B$, $E$




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

# Esempio

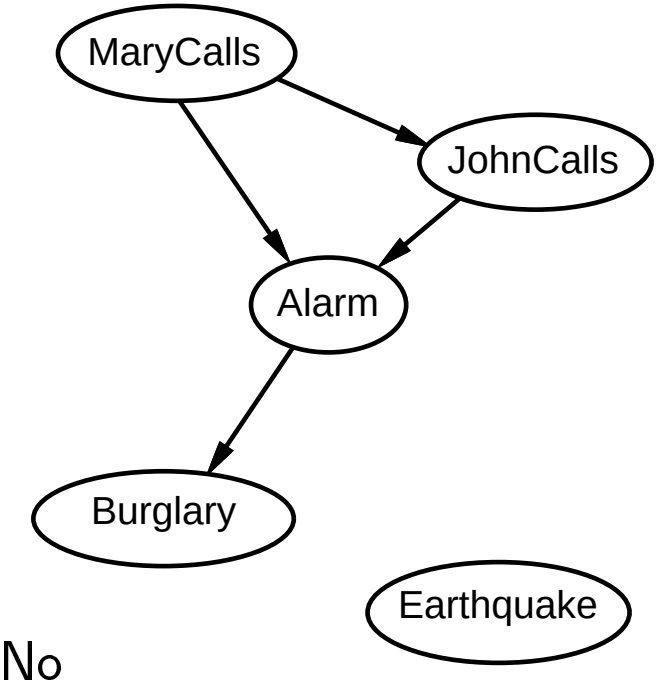
Supponiamo di scegliere l'ordine $M$, $J$, $A$, $B$, $E$




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

# Esempio

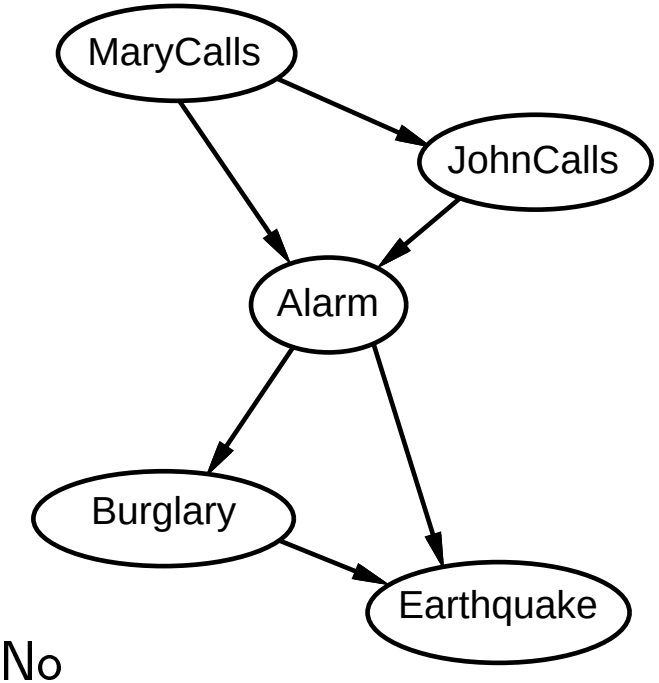
Supponiamo di scegliere l'ordine $M$, $J$, $A$, $B$, $E$




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

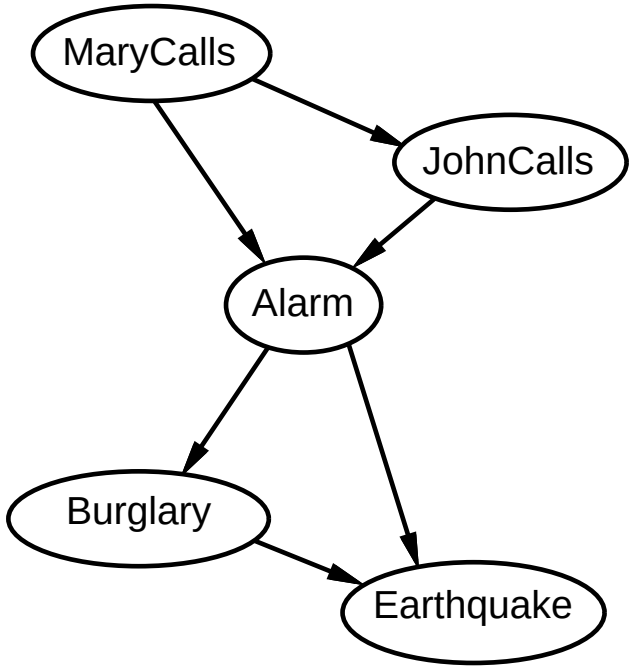
# Esempio

Supponiamo di scegliere l'ordine $M$, $J$, $A$, $B$, $E$




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

# Esempio




Decidere l'indipendenza condizionale è difficile nelle direzioni non causali

Valutare le probabilità condizionali è difficile in direzioni non causali

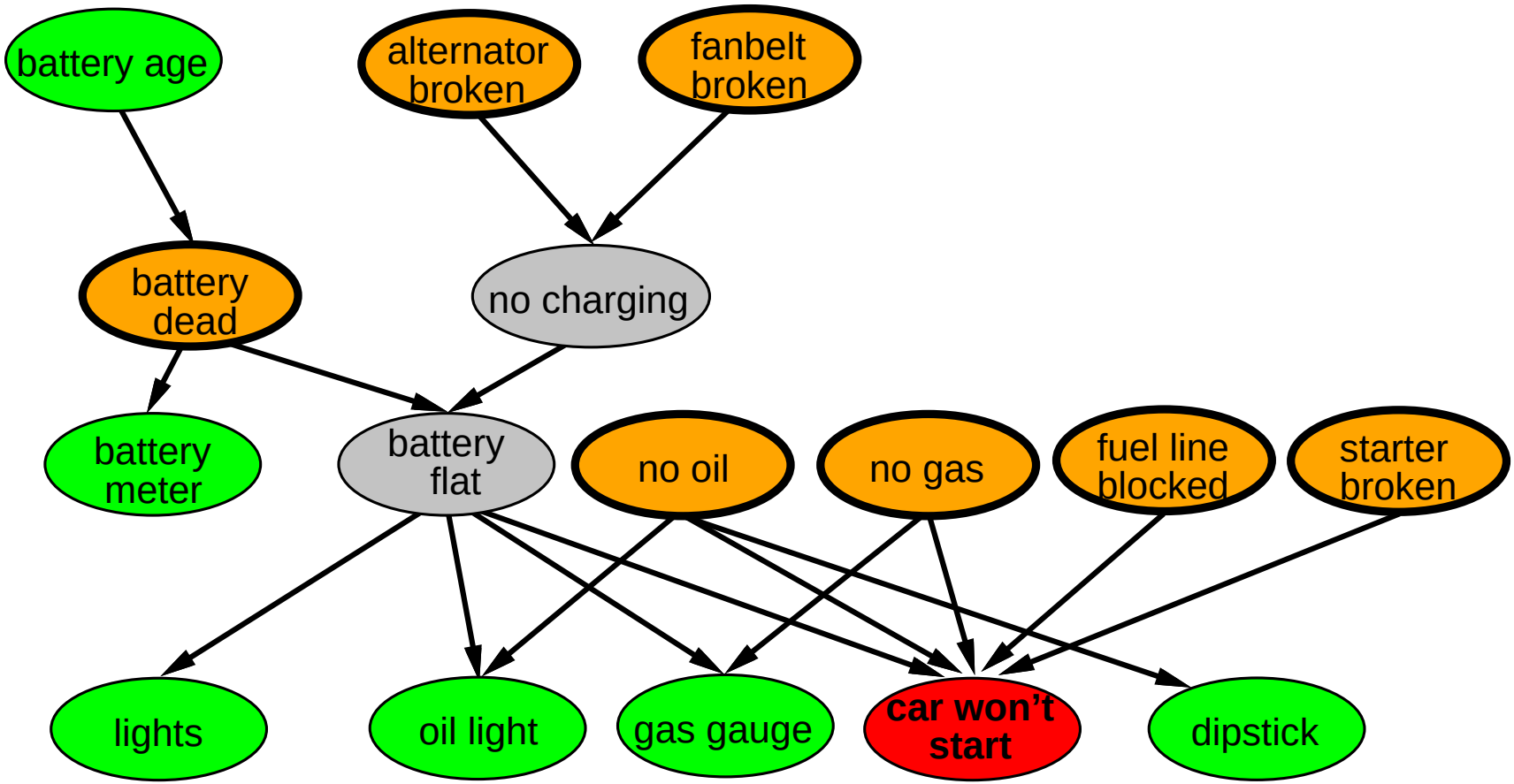
La rete è meno compatta: $1+2+4+2+4=13$ numeri necessari

# Esempio: diagnosi per automobile

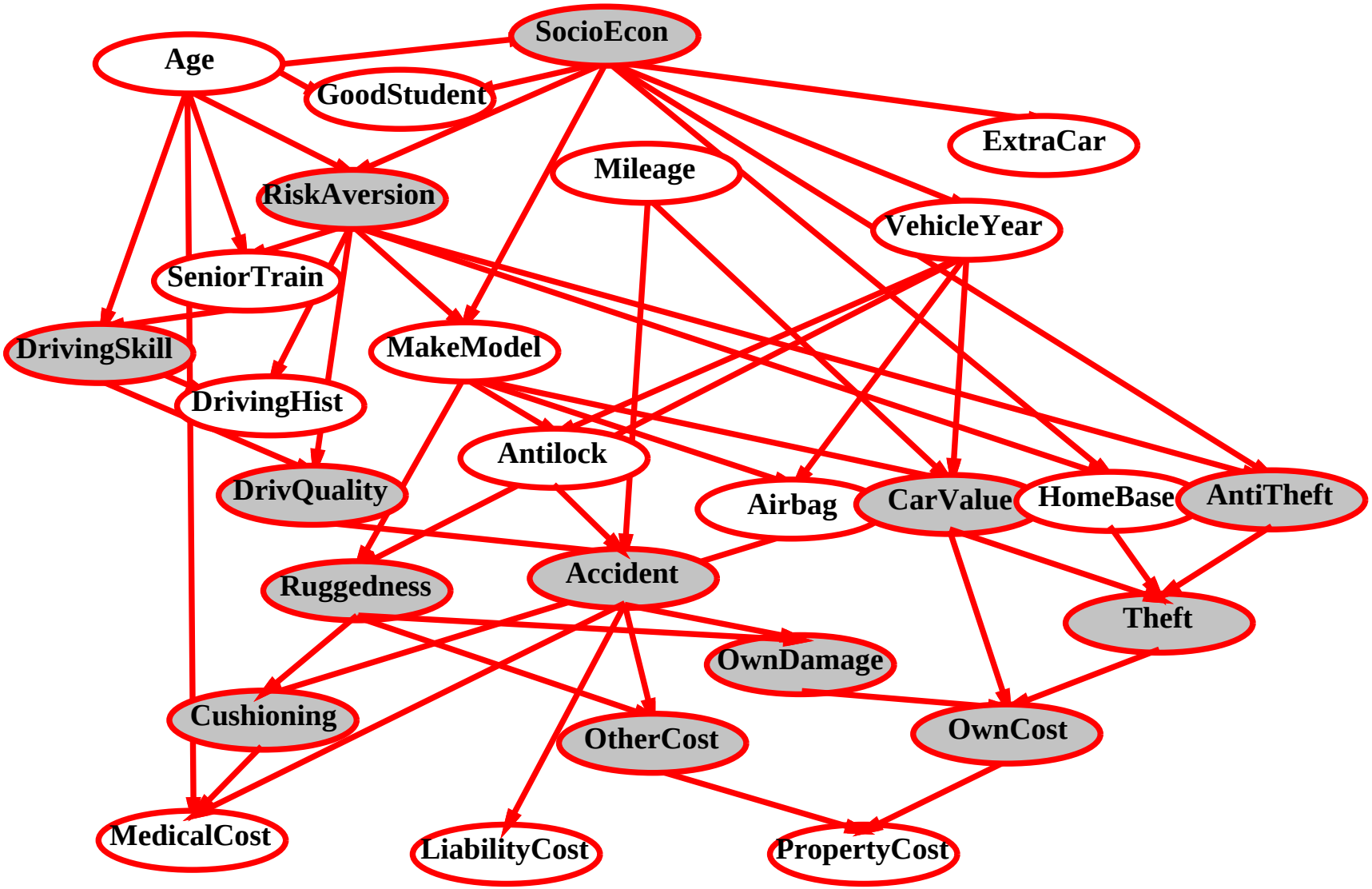
Evidenza iniziale: auto non parte

Variabili controllabili (in verde), variabili "rotto, da aggiustare" (in arancio)

Variabili nascoste (in grigio) assicurano struttura sparsa, riducono i parametri



# Esempio: assicurazione dell'automobile



# Compiti di inferenza

Query semplici: calcolare la probabilità a posteriori marginale $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
p.e., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$

Query congiuntive: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$

Decisioni ottimali: reti di decisioni includono informazioni di utilità;
inferenza probabilistica richiesta per $P(outcome|action, evidence)$

Recupero informazione: quale evidenza si deve cercare?

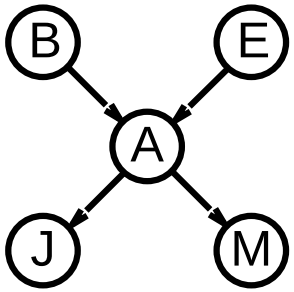
Analisi della sensitività: quali valori di probabilità sono i più critici?

Spiegazione: perché ho bisogno di un nuovo motore di avviamento?

# Inferenza tramite enumerazione

Modo un pò più furbo per marginalizzare alcune variabili dalla distribuzione congiunta senza costruire esplicitamente la sua rappresentazione

Query semplice sulla rete dell'allarme:




$$\mathbf{P}(B|j,m)$$
$$= \mathbf{P}(B,j,m)/P(j,m)$$
$$= \alpha\mathbf{P}(B,j,m)$$
$$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$$

Riscrittura di entrate della distribuzione congiunta usando il prodotto di entrate di CPT:

$$\mathbf{P}(B|j,m)$$
$$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$$
$$= \alpha\mathbf{P}(B)\Sigma_e P(e)\Sigma_a\mathbf{P}(a|B,e)P(j|a)P(m|a)$$

Enumerazione ricorsiva depth-first: $O(n)$ in spazio, $O(d^n)$ in tempo

# Algoritmo di enumerazione

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

**inputs**: $X$, the query variable
$\mathbf{e}$, observed values for variables $\mathbf{E}$
$bn$, a Bayesian network with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$

$\mathbf{Q}(X) \leftarrow$ a distribution over $X$, initially empty
**for each** value $x_i$ of $X$ **do**
extend $\mathbf{e}$ with value $x_i$ for $X$
$\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[$bn$], $\mathbf{e}$)
**return** NORMALIZE($\mathbf{Q}(X)$)

---

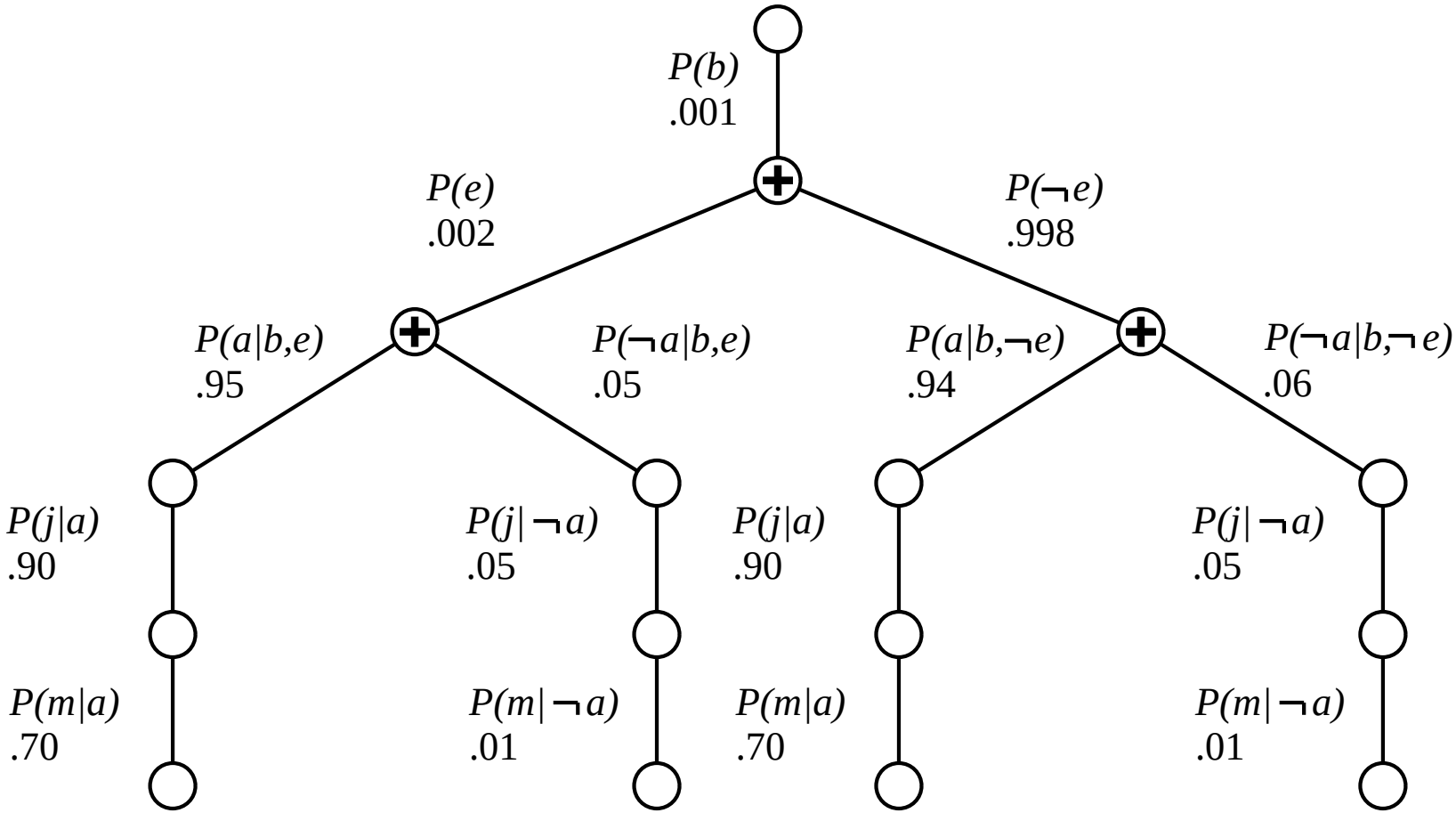
**function** ENUMERATE-ALL($vars, \mathbf{e}$) **returns** a real number

**if** EMPTY?($vars$) **then return** 1.0
$Y \leftarrow$ FIRST($vars$)
**if** $Y$ has value $y$ in $\mathbf{e}$
**then return** $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}$)
**else return** $\Sigma_y \; P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)
where $\mathbf{e}_y$ is $\mathbf{e}$ extended with $Y = y$

# Albero di valutazione

L'enumerazione è inefficiente: calcoli ripetuti

p.e., calcola $P(j|a)P(m|a)$ per ogni valore di $e$



# Inferenza tramite eliminazione di variabile

Eliminazione di variabile: effettuare le somme da destra a sinista, memorizzare i risultati intermedi (fattori) per evitare di ricalcolarli

$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&\quad = \alpha \underbrace{\mathbf{P}(B)}_{B} \Sigma_e \underbrace{P(e)}_{E} \Sigma_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a f_A(a,b,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) f_{\bar{A}JM}(b,e) \text{ (elimina } A\text{)} \\
&\quad = \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (elimina } E\text{)} \\
&\quad = \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

# Eliminazione di variabile: operazioni base

Eliminare una variabile da un prodotto di fattori:

1. muovere i fattori costanti al di fuori della somma
2. aggiungere le sottomatrici al prodotto "pointwise" dei fattori rimanenti

$$\sum_x f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \sum_x f_{i+1} \times \cdots \times f_k = f_1 \times \cdots \times f_i \times f_{\bar{X}}$$

assumendo che $f_1, \ldots, f_i$ non dipendano da $X$

Prodotto pointwise di fattori $f_1$ e $f_2$:

$$f_1(x_1, \ldots, x_j, y_1, \ldots, y_k) \times f_2(y_1, \ldots, y_k, z_1, \ldots, z_l) = f(x_1, \ldots, x_j, y_1, \ldots, y_k, z_1, \ldots, z_l)$$

P.e., $f_1(a, b) \times f_2(b, c) = f(a, b, c)$

# Algoritmo di eliminazione di variabile

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

**inputs**: $X$, the query variable

$\mathbf{e}$, evidence specified as an event

$bn$, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

$factors \leftarrow [\,]$; $vars \leftarrow$ REVERSE(VARS[$bn$])

**for each** $var$ **in** $vars$ **do**

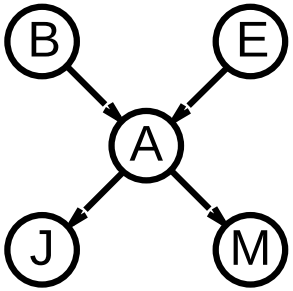
$factors \leftarrow [\text{MAKE-FACTOR}(var, \mathbf{e})|factors]$

**if** $var$ is a hidden variable **then** $factors \leftarrow \text{SUM-OUT}(var, factors)$

**return** NORMALIZE(POINTWISE-PRODUCT($factors$))

# Variabili irrilevanti

Consideriamo la query $P(JohnCalls|Burglary = true)$




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

La somma su $m$ è uguale a 1; $M$ è **irrilevante** per la query

Thm 1: $Y$ è irrilevante a meno che $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Qui, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, e
$Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
quindi $M$ è irrilevante

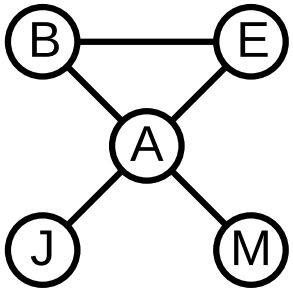
(Confrontare con backward chaining a partire dalla query in KB con clausole di Horn)

# Variabili irrilevanti

Defn: grafo moralizzato di una rete bayesiana: sposare tutti i genitori ed eliminare la direzione degli archi

Defn: $\mathbf{F}$ è m-separato da $\mathbf{G}$ tramite $\mathbf{H}$ sse è separato tramite $\mathbf{H}$ nel grafo moralizzato

Thm 2: $Y$ è irrilevante se m-separato da $X$ tramite $\mathbf{E}$




Per $P(JohnCalls|Alarm = true)$, sia
$Burglary$ che $Earthquake$ sono irrilevanti

# Complessità dell'inferenza esatta

Reti singolarmente connesse (o polytree):

– ogni coppia di nodi è connessa da al più un cammino (non diretto)

– il costo in tempo e spazio della eliminazione di variabile è $O(d^k n)$

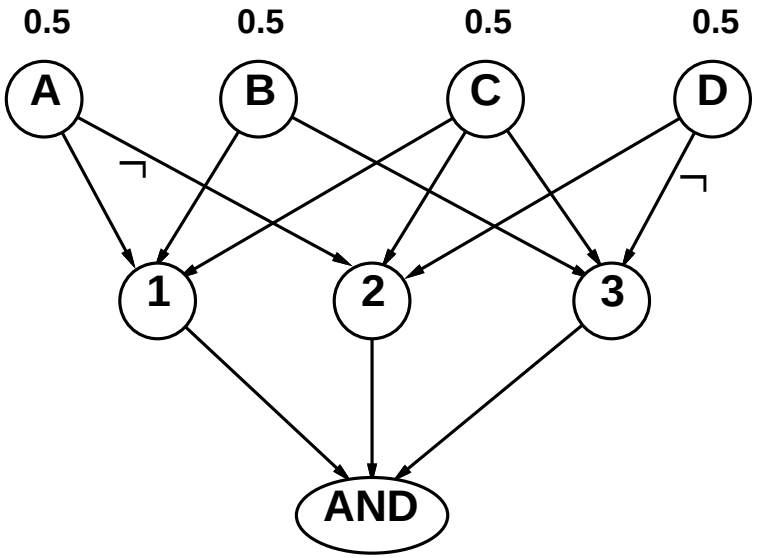
Reti connesse più che singolarmente:

– possibile ridurre 3SAT alla inferenza esatta $\Rightarrow$ NP-hard

– equivalente a modelli 3SAT con **conteggio** (del numero di soluzioni)

$\Rightarrow$ #P-complete

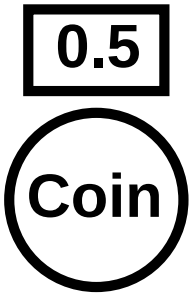
1. A v B v C
2. C v D v ¬A
3. B v C v ¬D



# Inferenza tramite simulazione stocastica

Idea base:

1) Estrarre $N$ campioni da una distribuzione di campionamento $S$
2) Calcolare la probabilità a posteriori approssimata $\hat{P}$
3) Mostrare che converge alla vera probabilità $P$




Outline:

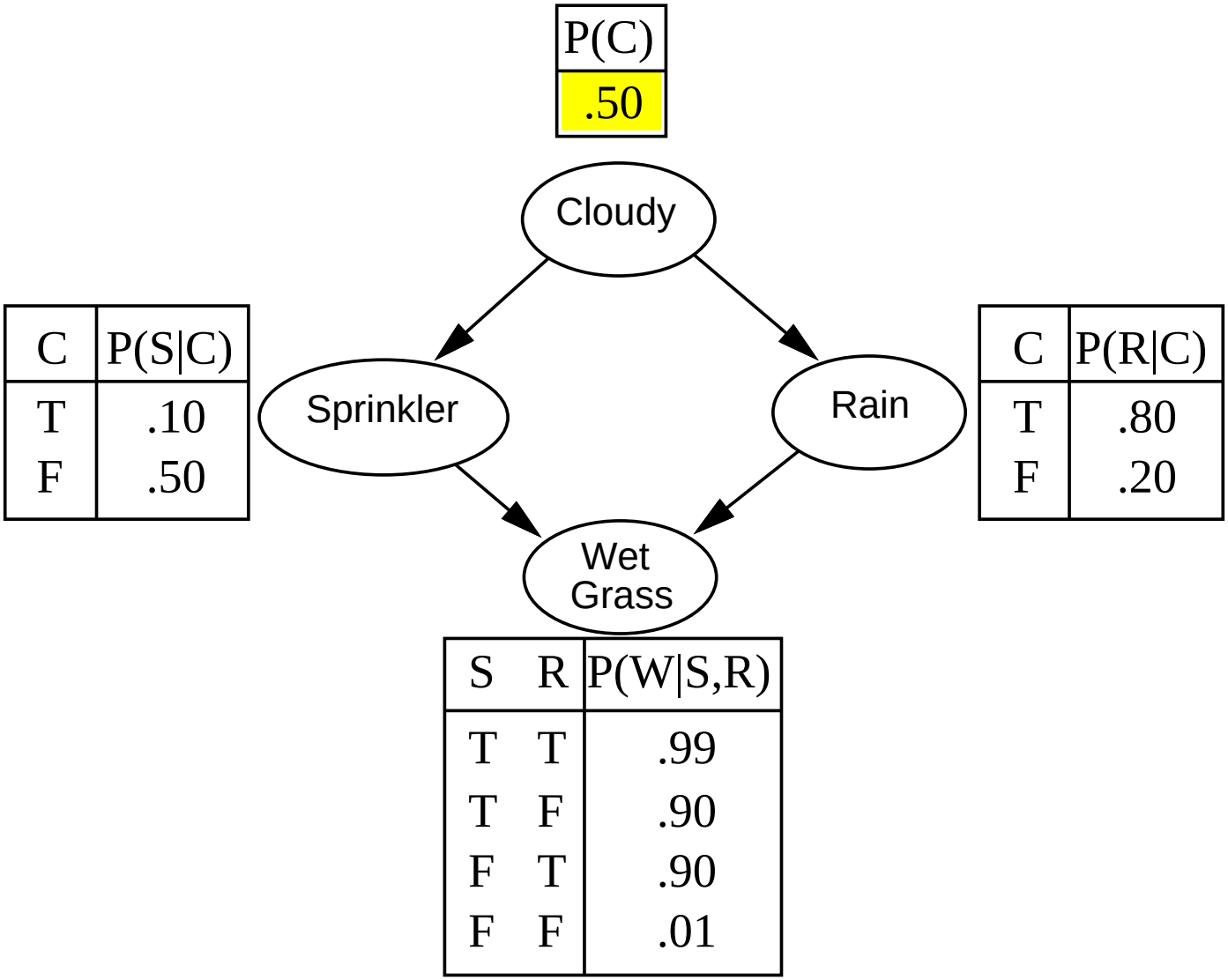
- Campionamento da una rete vuota
- Rejection sampling: rigettare i campioni in disaccordo con l'evidenza
- Likelihood weighting: usare l'evidenza per pesare i campioni
- Markov chain Monte Carlo (MCMC): campiona in accordo ad un processo stocastico la cui distribuzione stazionaria è la vera probabilità

# Campionamento da una rete vuota

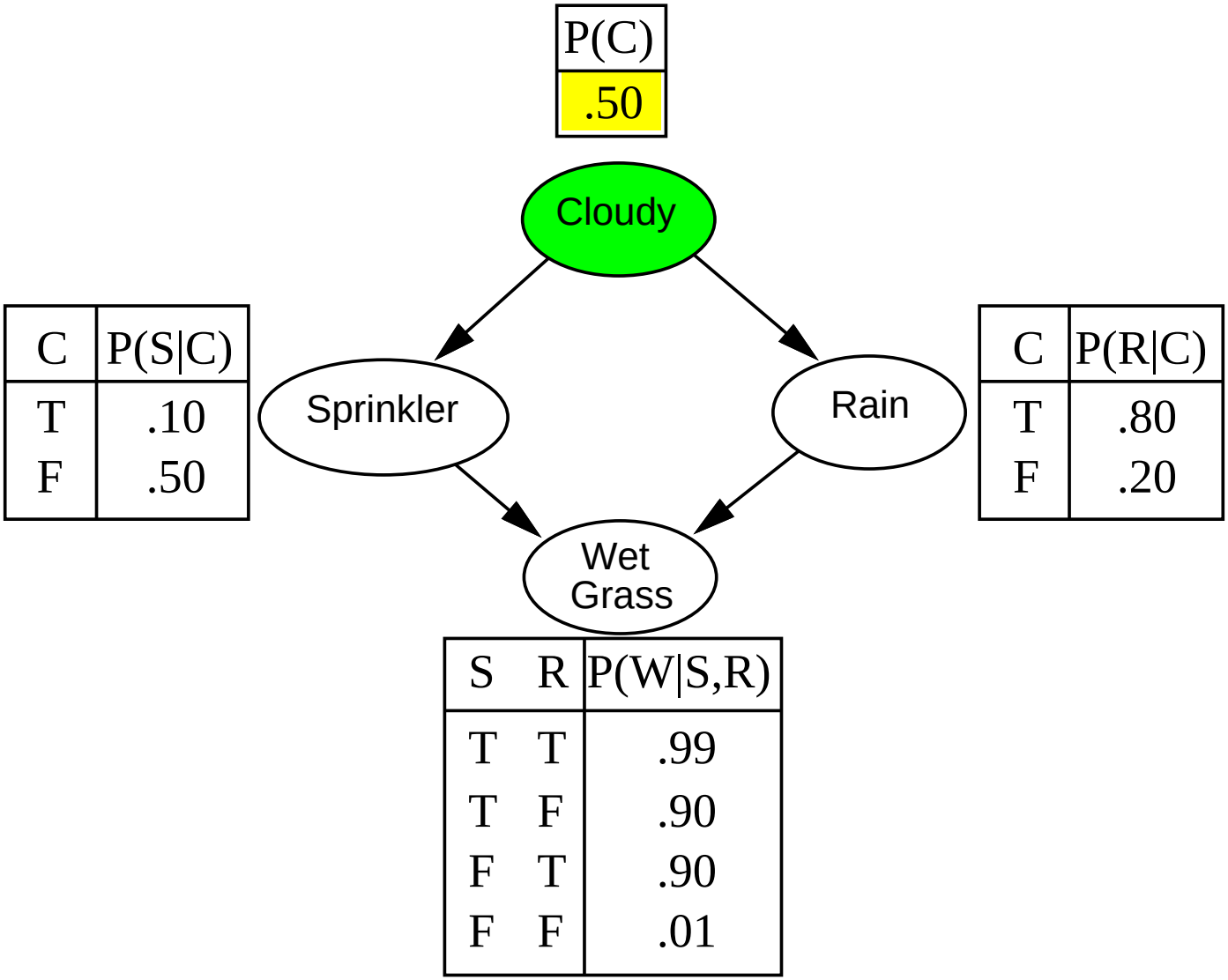
**function** PRIOR-SAMPLE($bn$) **returns** an event sampled from $bn$
  **inputs**: $bn$, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  $\mathbf{x} \leftarrow$ an event with $n$ elements
  **for** $i = 1$ **to** $n$ **do**
    $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid Parents(X_i))$
  **return** $\mathbf{x}$

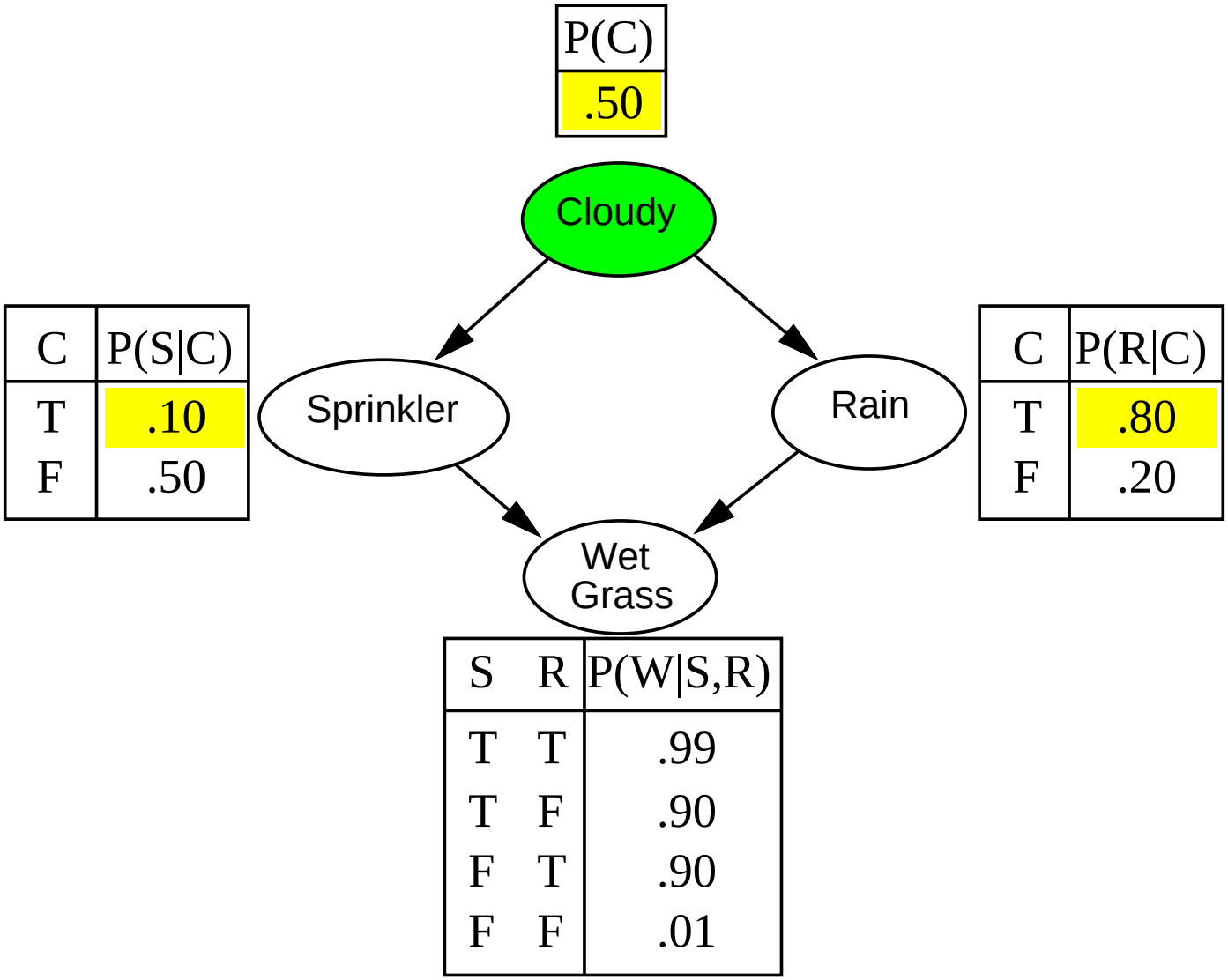
# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

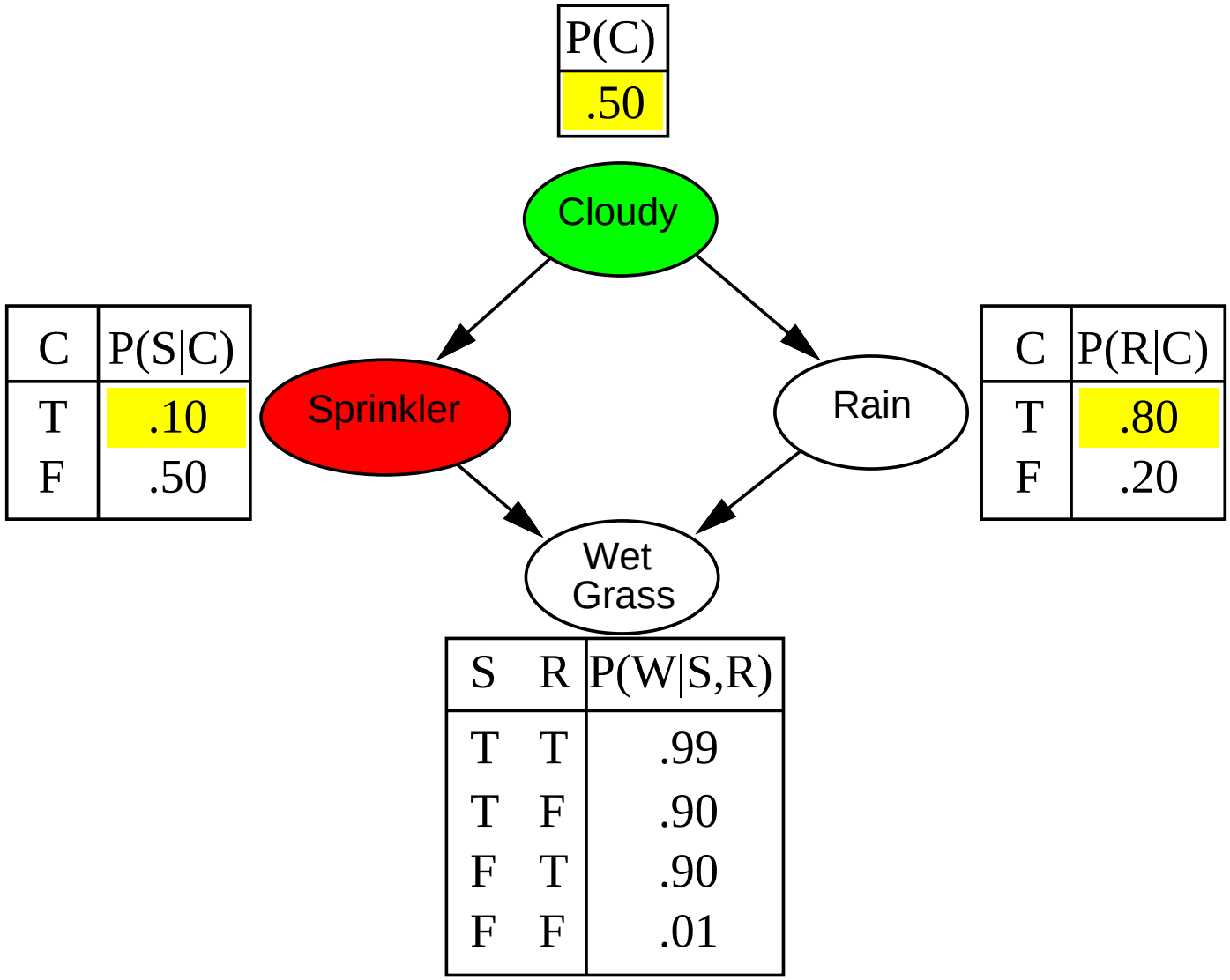
# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

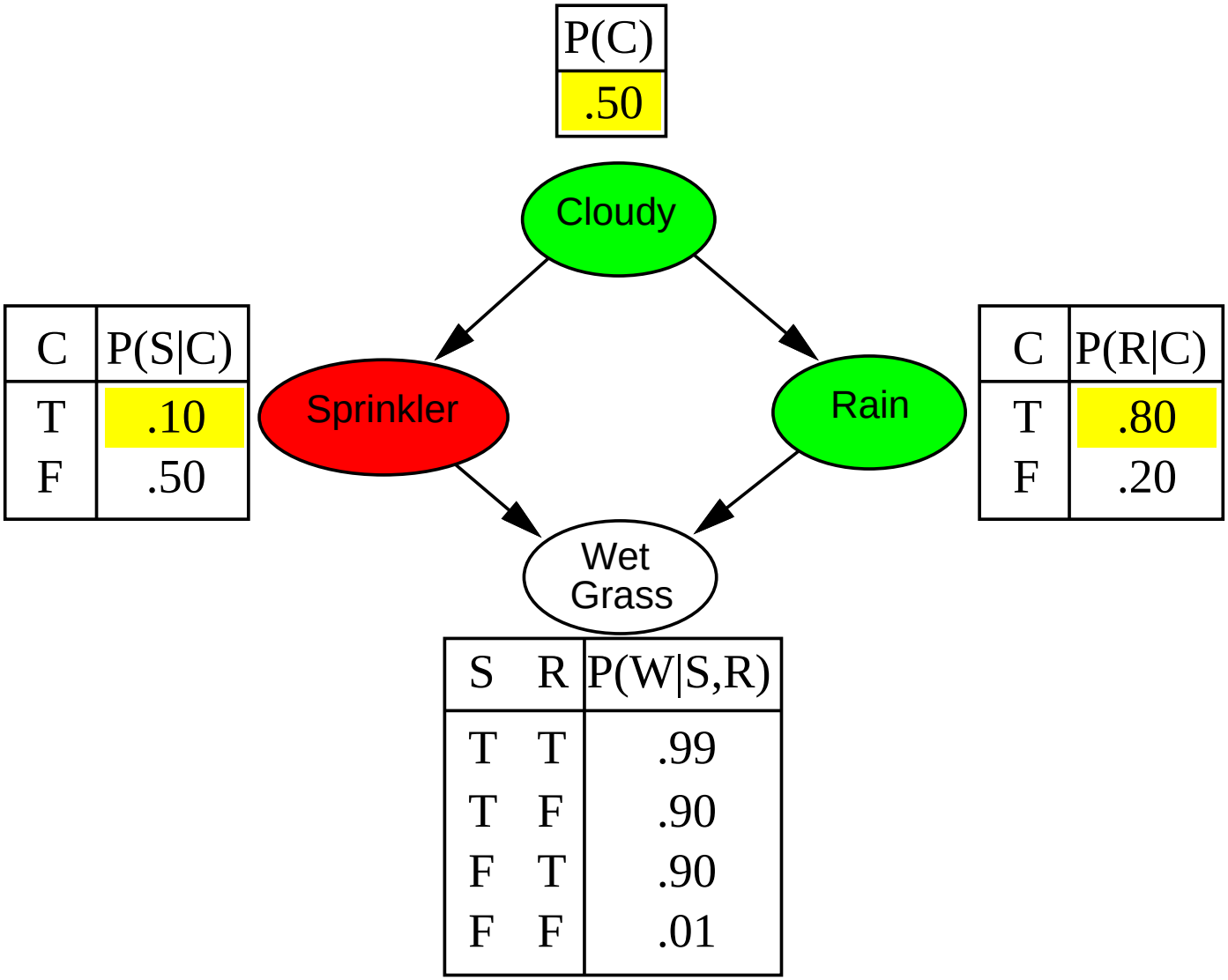
# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

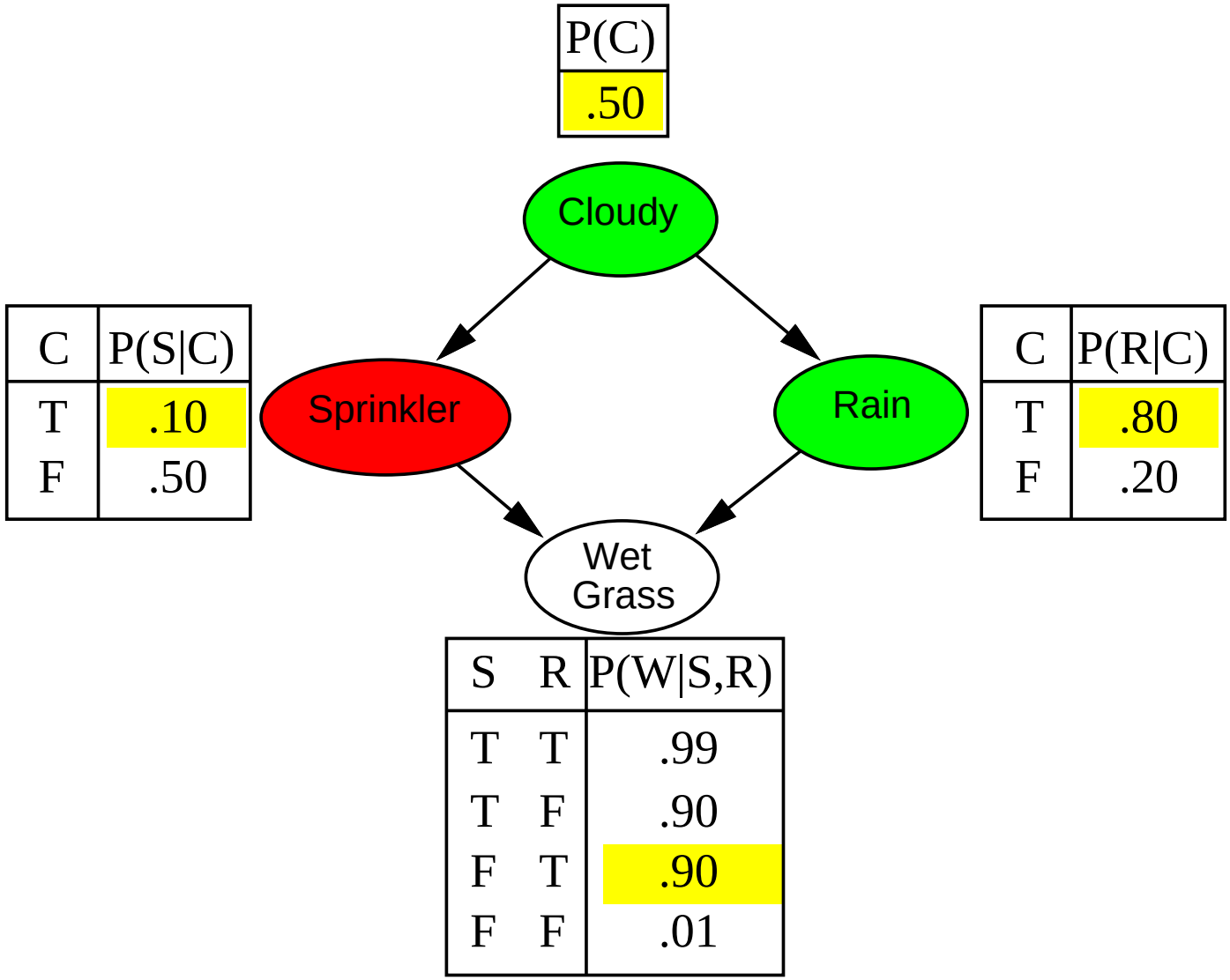
# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

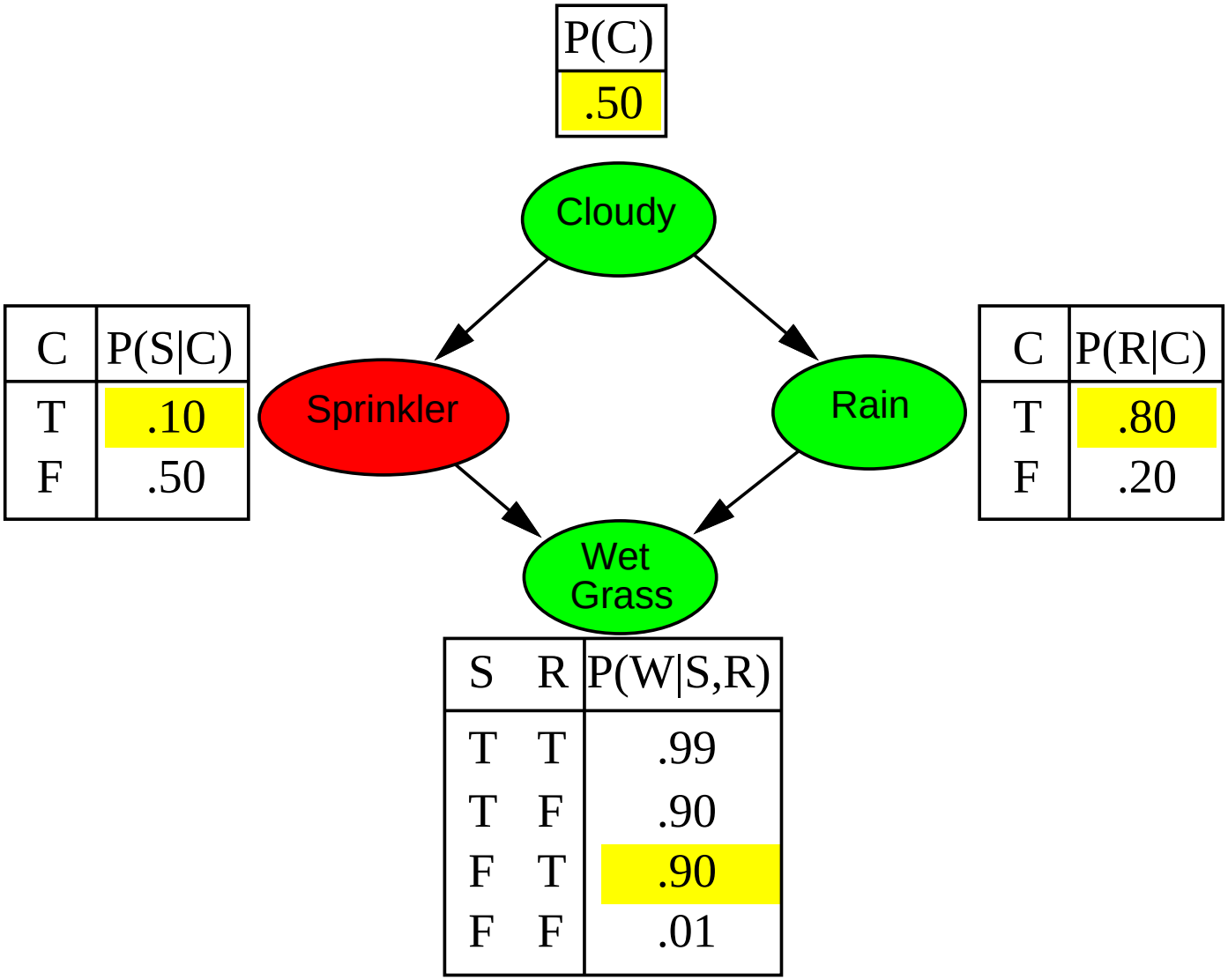
# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

# Esempio


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

# Campionamento da una rete vuota

Probabilità che PriorSample generi un evento particolare

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

cioè, la vera probabilità a priori

P.e., $S_{PS}(t, f, t, t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t, f, t, t)$

Posto $N_{PS}(x_1 \dots x_n)$ essere il numero di campioni generati per l'evento $x_1, \dots, x_n$

Allora abbiamo

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

Cioè, stime derivate da PriorSample sono consistenti

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

# Rejection sampling

$\hat{\mathbf{P}}(X|\mathbf{e})$ stimate da campioni in accordo con $\mathbf{e}$

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

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

P.e., stimare $\mathbf{P}(Rain|Sprinkler = true)$ usando 100 campioni
  27 campioni hanno $Sprinkler = true$
    Di questi, 8 hanno $Rain = true$ e 19 hanno $Rain = false$.

$\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19\rangle) = \langle 0.296, 0.704\rangle$

# Analisi di rejection sampling

$\hat{\mathbf{P}}(X|\mathbf{e}) = \alpha \mathbf{N}_{PS}(X,\mathbf{e})$ (def. algoritmo)
$= \mathbf{N}_{PS}(X,\mathbf{e})/N_{PS}(\mathbf{e})$ (normalizzato tramite $N_{PS}(\mathbf{e})$)
$\approx \mathbf{P}(X,\mathbf{e})/P(\mathbf{e})$ (properietà di PriorSample)
$= \mathbf{P}(X|\mathbf{e})$ (def. di probabilità condizionale)

Quindi rejection sampling restituisce stime consistenti della prob. a posteriori

Problemi: costosissimo se $P(\mathbf{e})$ è piccola

$P(\mathbf{e})$ converge esponenzialmente con il numero di variabili di evidenza!

# Likelihood weighting

Idea: fissare le variabili di evidenza, campionare solo variabili non di evidenza, e pesare ogni campione con la likelihood accordata dall'evidenza

**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

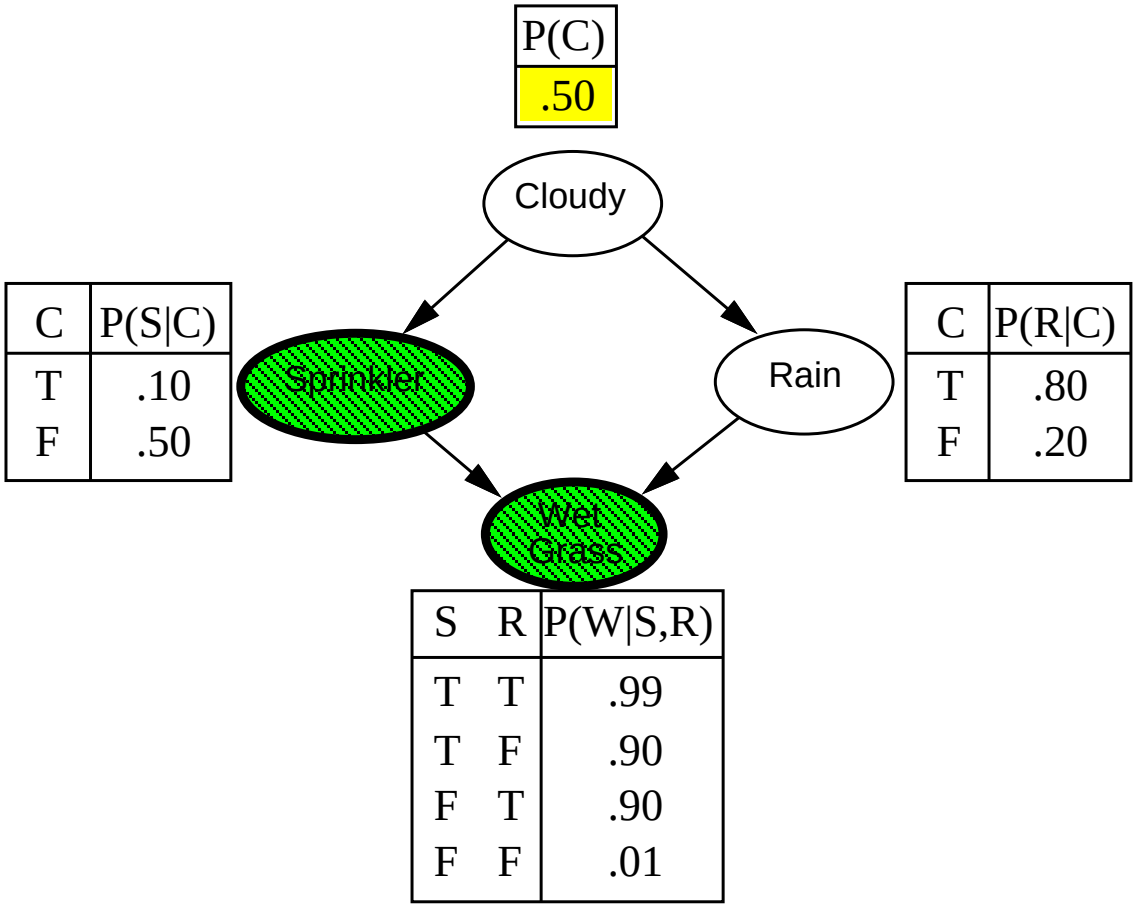
  **for** $j = 1$ to $N$ **do**
    $\mathbf{x}, w \leftarrow$ WEIGHTED-SAMPLE($bn$)
    $\mathbf{W}[x] \leftarrow \mathbf{W}[x] + w$ where $x$ is the value of $X$ in $\mathbf{x}$
  **return** NORMALIZE($\mathbf{W}[X]$)

---

**function** WEIGHTED-SAMPLE($bn, \mathbf{e}$) **returns** an event and a weight

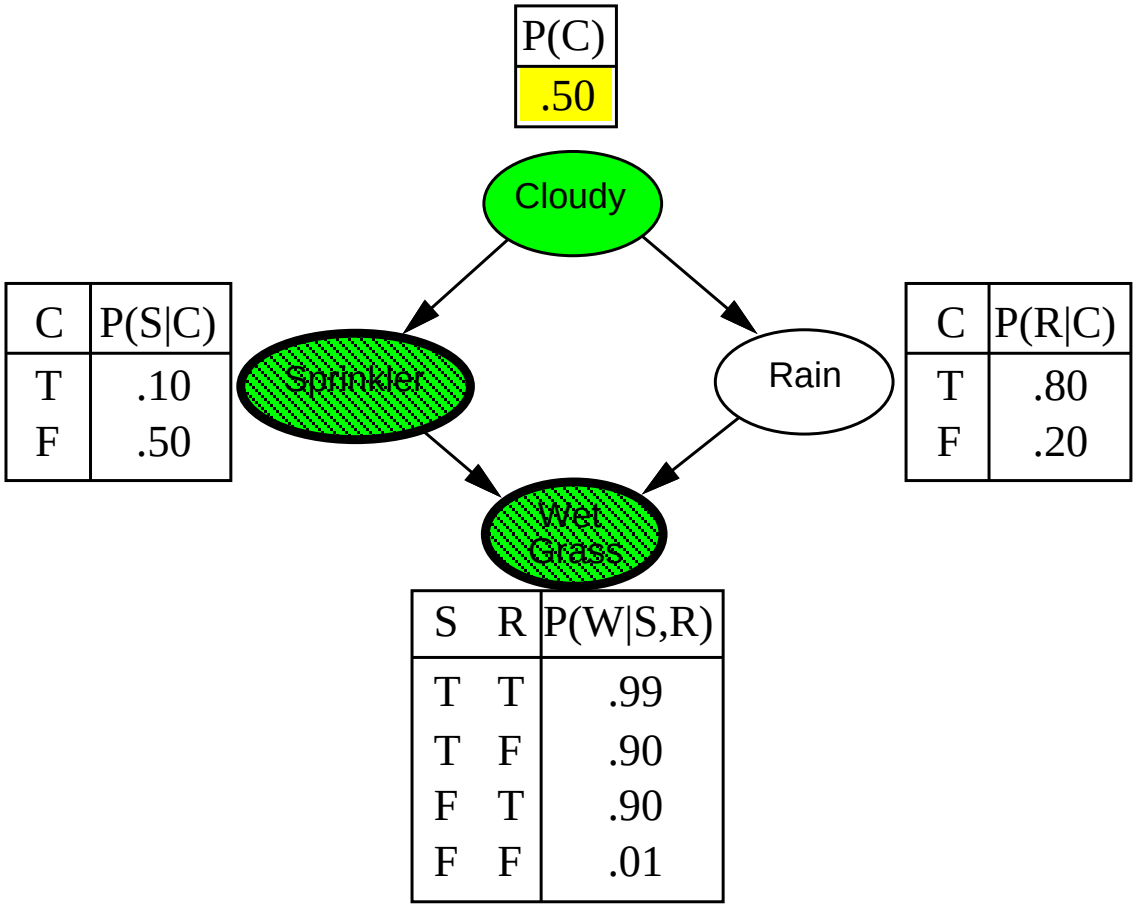
  $\mathbf{x} \leftarrow$ an event with $n$ elements; $w \leftarrow 1$
  **for** $i = 1$ **to** $n$ **do**
    **if** $X_i$ has a value $x_i$ in $\mathbf{e}$
      **then** $w \leftarrow w \times P(X_i = x_i \mid Parents(X_i))$
      **else** $x_i \leftarrow$ a random sample from $\mathbf{P}(X_i \mid Parents(X_i))$
  **return** $\mathbf{x}$, $w$

# Esempio di likelihood weighting



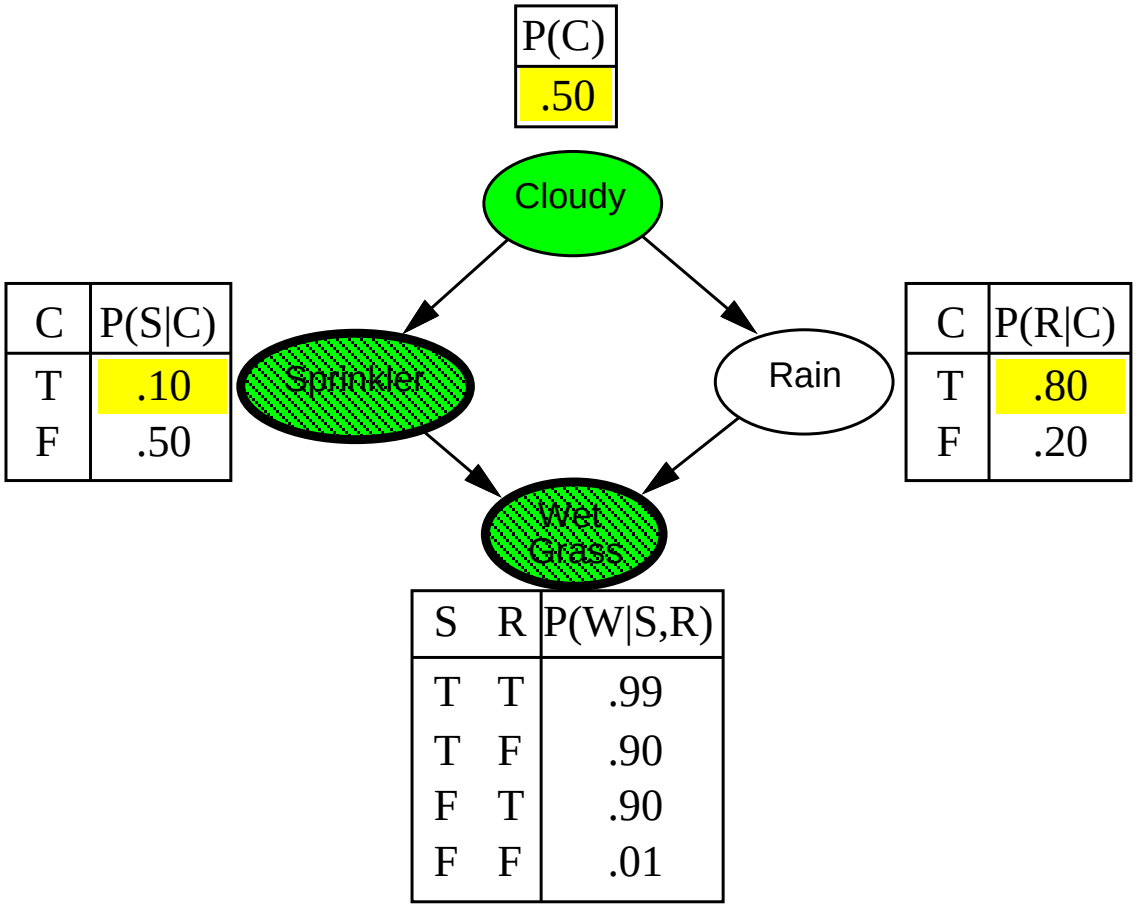

$w = 1.0$

# Esempio di likelihood weighting



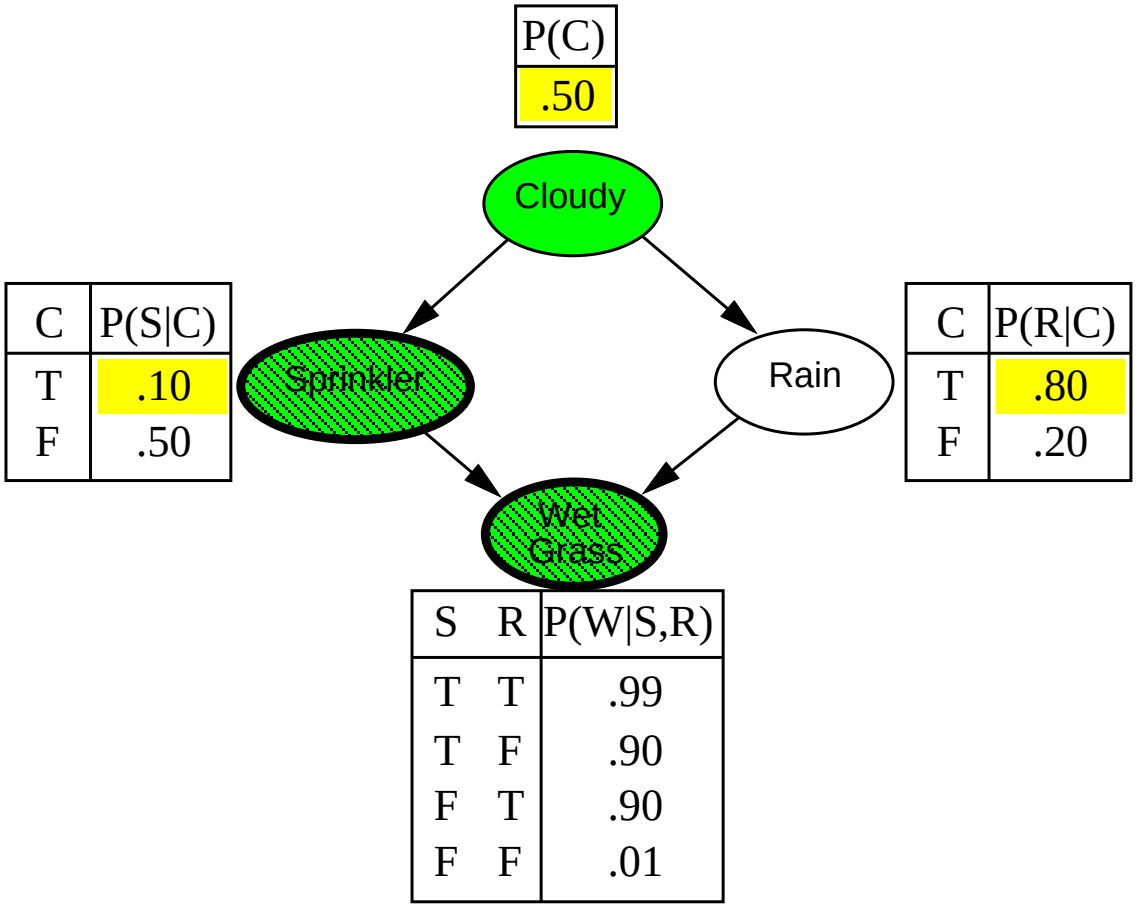

$w = 1.0$

# Esempio di likelihood weighting



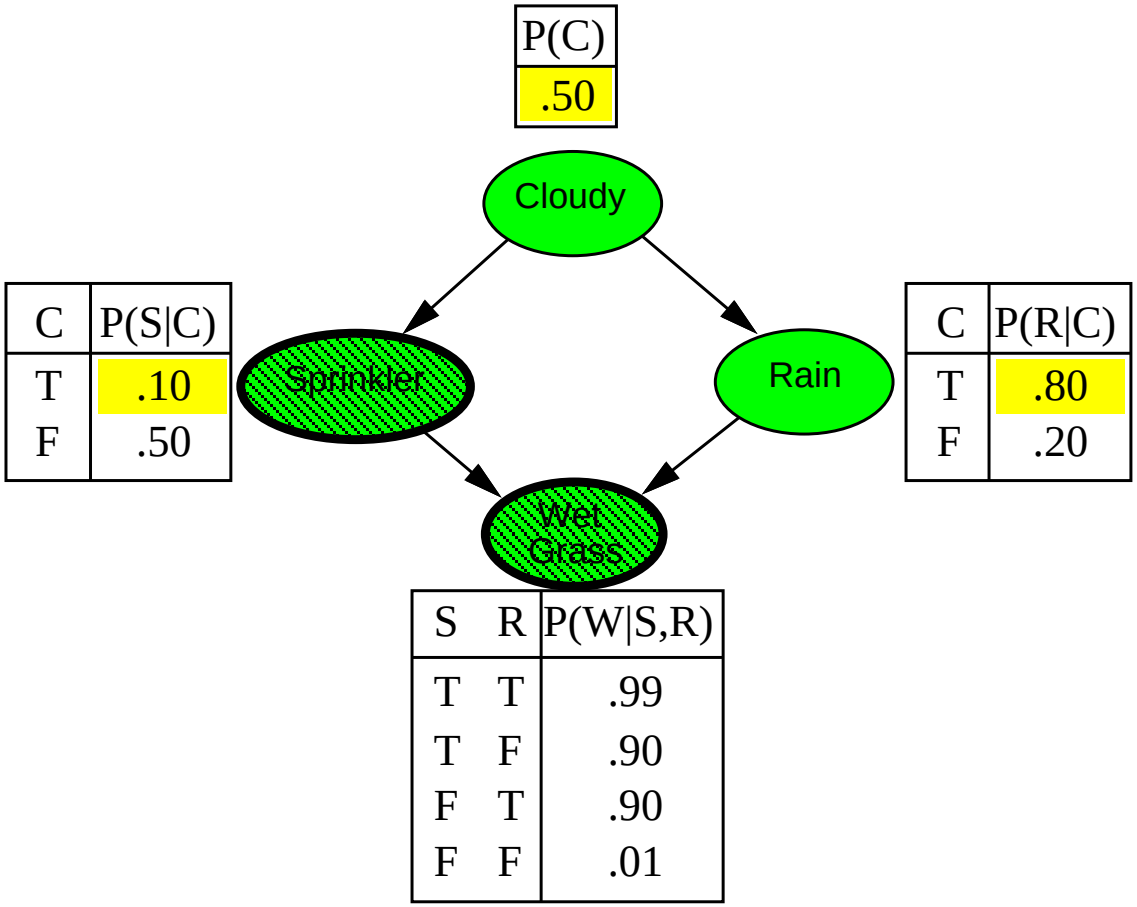

$w = 1.0$

# Esempio di likelihood weighting



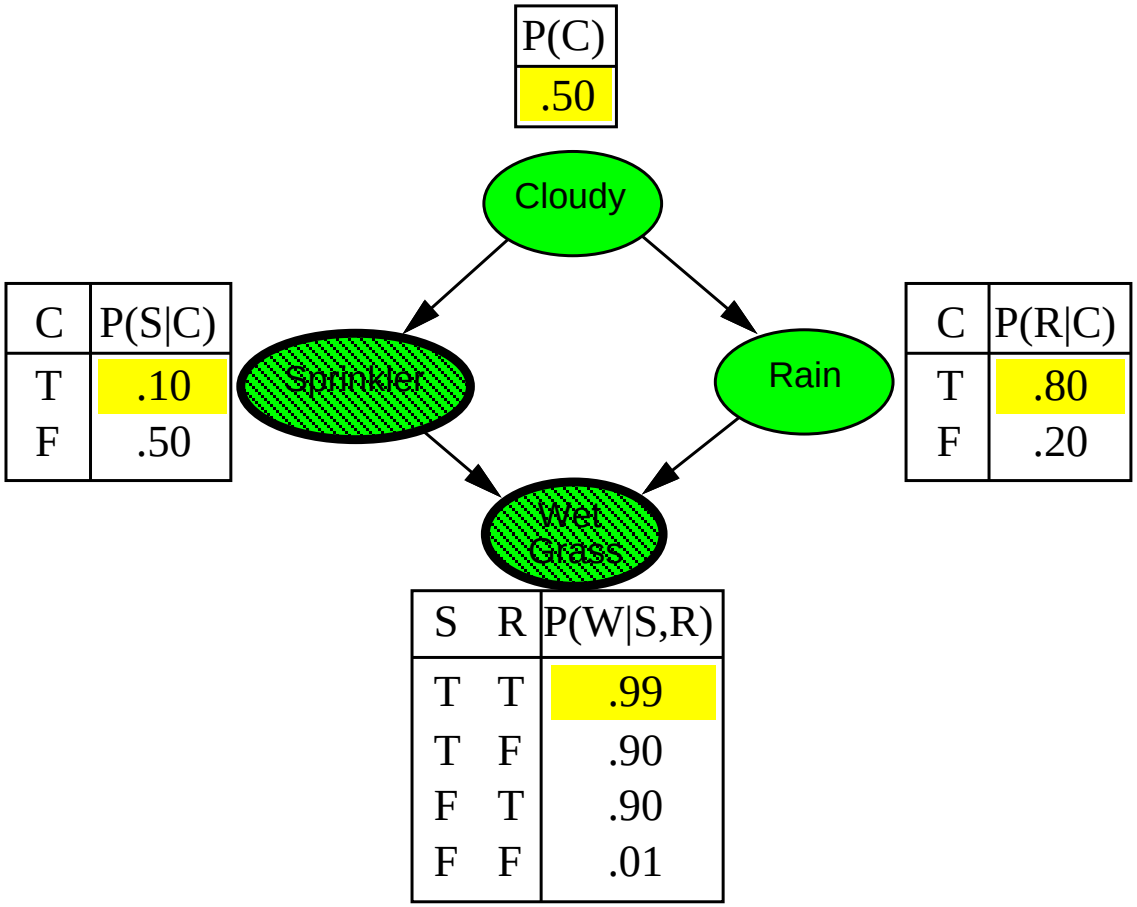

$w = 1.0 \times 0.1$

# Esempio di likelihood weighting



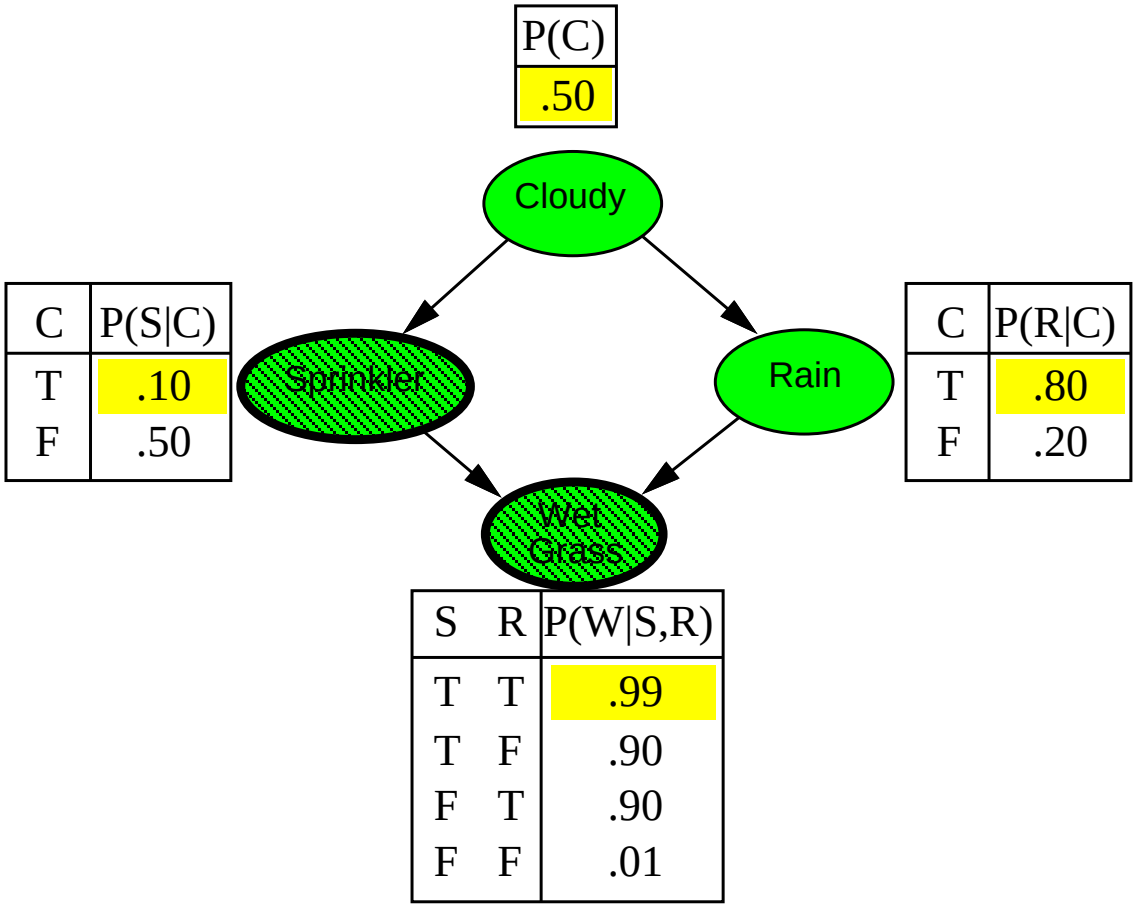

$w = 1.0 \times 0.1$

# Esempio di likelihood weighting




$w = 1.0 \times 0.1$

# Esempio di likelihood weighting




$w = 1.0 \times 0.1 \times 0.99 = 0.099$

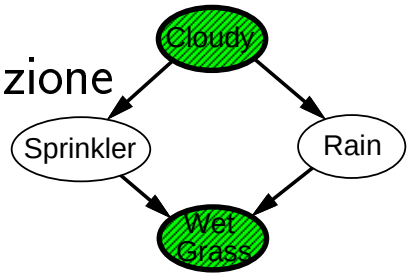
# Analisi di likelihood weighting

La probabilità di campionamento per WeightedSample è

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

Nota: pone attenzione solo all'evidenza negli **antenati**

$\Rightarrow$ da qualche parte "nel mezzo" fra la distribuzione a priori e quella a posteriori




Il peso per un dato campione $\mathbf{z}, \mathbf{e}$ è

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

La probabilità di campionamento pesata è

$$\begin{aligned} & S_{WS}(\mathbf{z}, \mathbf{e}) w(\mathbf{z}, \mathbf{e}) \\ & = \Pi_{i=1}^{l} P(z_i | Parents(Z_i)) \ \ \Pi_{i=1}^{m} P(e_i | Parents(E_i)) \\ & = P(\mathbf{z}, \mathbf{e}) \text{ (per la semantica standard globale della rete)} \end{aligned}$$

Quindi likelihood weighting restituisce stime consistenti
però le prestazioni degradano con la presenza di tante variabili di evidenza
poiché pochi esempi hanno quasi tutto il peso totale

# Riassunto

Inferenza esatta tramite l'eliminazione di variabile:

- polinomiale su polialberi, NP-hard in generale
- spazio = tempo, dipendente dalla topologia

Inferenza approssimata tramite LW:

- LW si comporta male quando c'è molta evidenza (soprattutto a "valle")
- LW in genere indipendente dalla topologia
- La convergenza può essere molto lenta per probabilità vicine a 1 o 0
- Può trattare combinazioni arbitrarie di variabili discrete e continue