

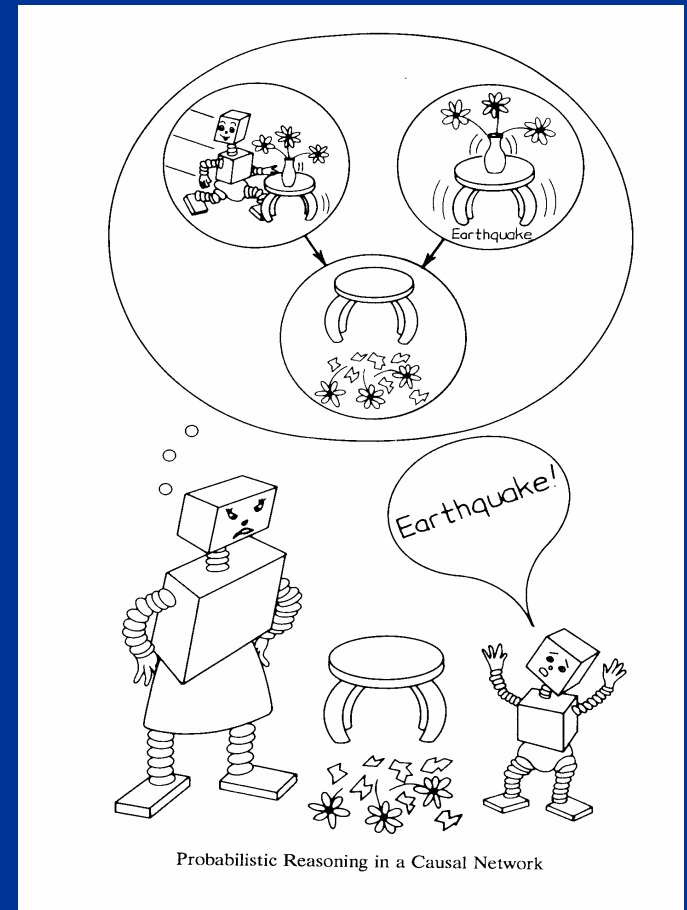
Modelos Gráficos Probabilistas

L. Enrique Sucar

INAOE

Sesión 7: Redes Bayesianas – Inferencia:

2da Parte



[Neapolitan 90]

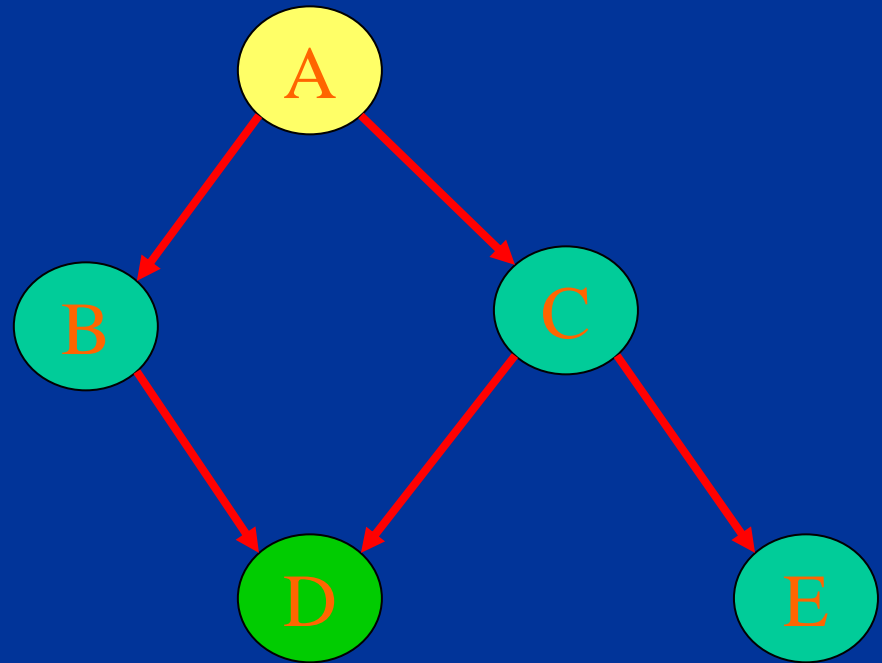
Otros métodos de inferencia

- Algoritmo de Eliminación
- Redes multiconectadas
 - condicionamiento
 - simulación estocástica
 - agrupamiento
- Abducción

Algoritmo de Eliminación

- Supongamos que deseamos calcular la probabilidad de un nodo dado un conjunto de nodos conocidos
- En la RB:

$$P(A | D)$$



Eliminación

- Podemos distinguir 3 grupos de nodos:
 - XE: evidencia (D)
 - XF: hipótesis – para el cual obtenemos la probabilidad (A)
 - XR: resto – se marginalizan (B,C,E)
- Podemos entonces obtener la probabilidad posterior por marginalización:

$$P(XF | XE) = P(XE, XF) / P(XE)$$

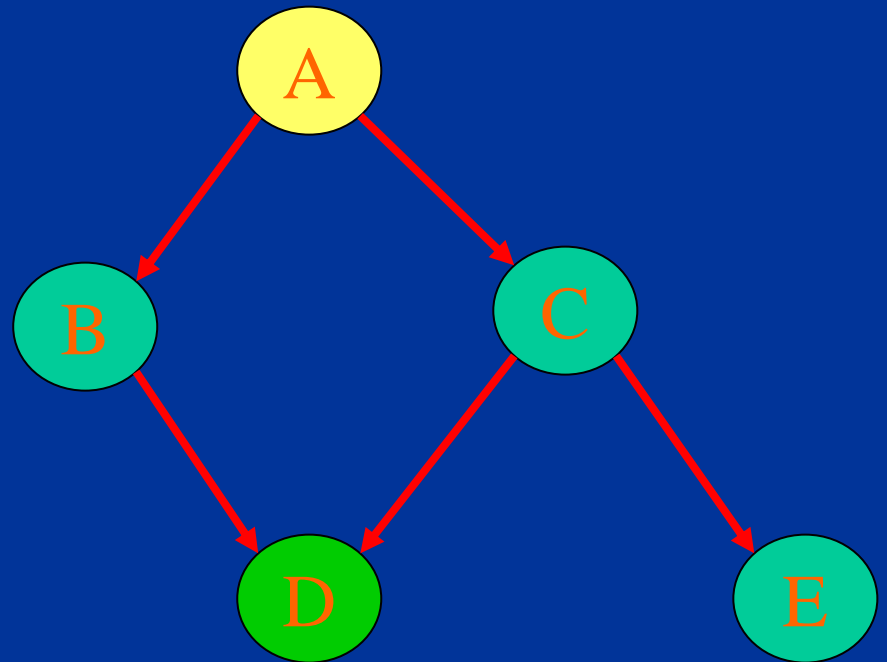
$$P(XE, XF) = \sum_{XR} P(XE, XF, XR)$$

$$P(XE) = \sum_{XF} P(XE, XF)$$

Eliminación

- El problema es que si hacemos esto directamente, el cómputo se vuelve muy complejo (número exponencial de sumas)
- Para hacer mucho más eficiente el cálculo, representamos la distribución en forma factorizada (independencias) y explotamos la ley distributiva

Ejemplo



$$P(A,B,C,E) = \sum_D P(A)P(B|A)P(C|A)P(D|B,C)P(E|C)$$

$$P(A,B,C,E) = P(A)P(B|A)P(C|A) P(E|C) \sum_D P(D|B,C)$$

Hay un ahorro de k^5 a k^3 , donde k es el # de valores por variable

Ejemplo

- Obtengamos ahora los términos necesarios para calcular $P(A|D)$ (recordar que \underline{D} es conocida, por lo que esa tabla se reduce):

$$P(A, D) = \sum_B \sum_C \sum_E P(A)P(B|A)P(C|A)P(\underline{D}|B,C)P(E|C)$$

$$P(A, D) = \sum_B \sum_C P(A)P(B|A)P(C|A)P(\underline{D}|B,C) \sum_E P(E|C)$$

$$P(A, D) = \sum_B P(A)P(B|A) \sum_C P(C|A)P(\underline{D}|B,C) \sum_E P(E|C)$$

$$P(A, D) = P(A) \sum_B P(B|A) \sum_C P(C|A)P(\underline{D}|B,C) \sum_E P(E|C)$$

Ejemplo

- Si introducimos cierta notación:

$$m_E(C) = \sum_E P(E|C), \text{ Entonces:}$$

$$P(A, D) = P(A) \sum_B P(B|A) \sum_C P(C|A) P(\underline{D}|B, C) m_E(C)$$

$$m_C(A, B) = \sum_C P(C|A) P(\underline{D}|B, C) m_E(C)$$

$$P(A, D) = P(A) \sum_B P(B|A) m_C(A, B)$$

$$m_B(A) = \sum_B P(B|A) m_C(A, B)$$

$$P(A, D) = P(A) m_B(A)$$

Ejemplo

- De aquí podemos obtener $P(D)$:

$$P(D) = \sum_A P(A) m_B(A)$$

- Y entonces:

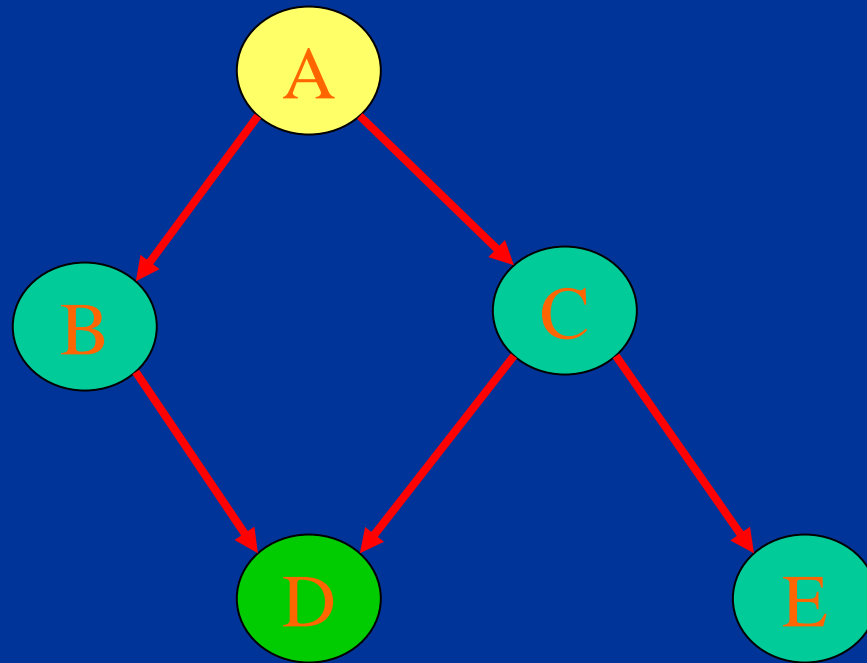
$$P(A|D) = P(A) m_B(A) / \sum_A P(A) m_B(A)$$

- A partir de estas ideas se deriva el algoritmo general de Eliminación para inferencia en redes bayesianas (ver notas Jordan – cap 3)

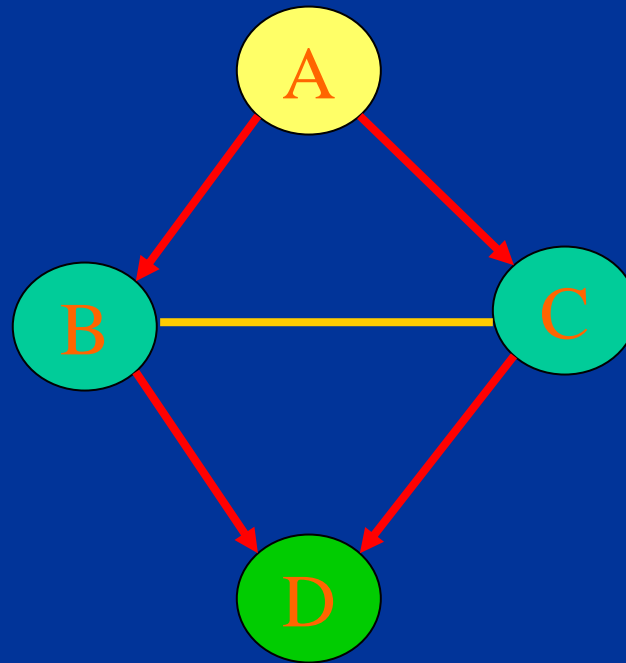
Interpretación gráfica

- La marginalización de las variables en el algoritmo, corresponde a la idea de *eliminación* de nodos en el grafo
- De acuerdo a un orden de las variables, vamos eliminando los nodos, conectando sus vecinos

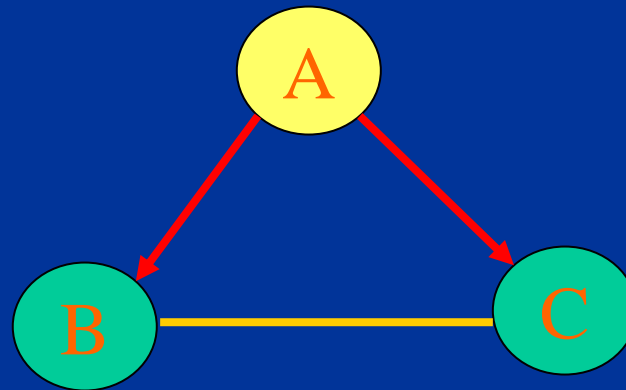
Ejemplo



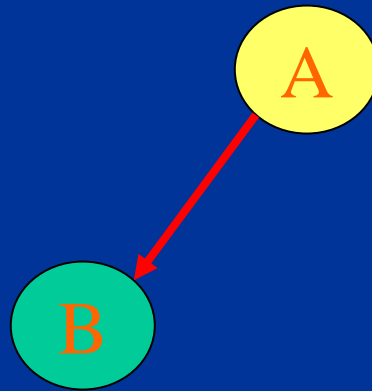
Ejemplo



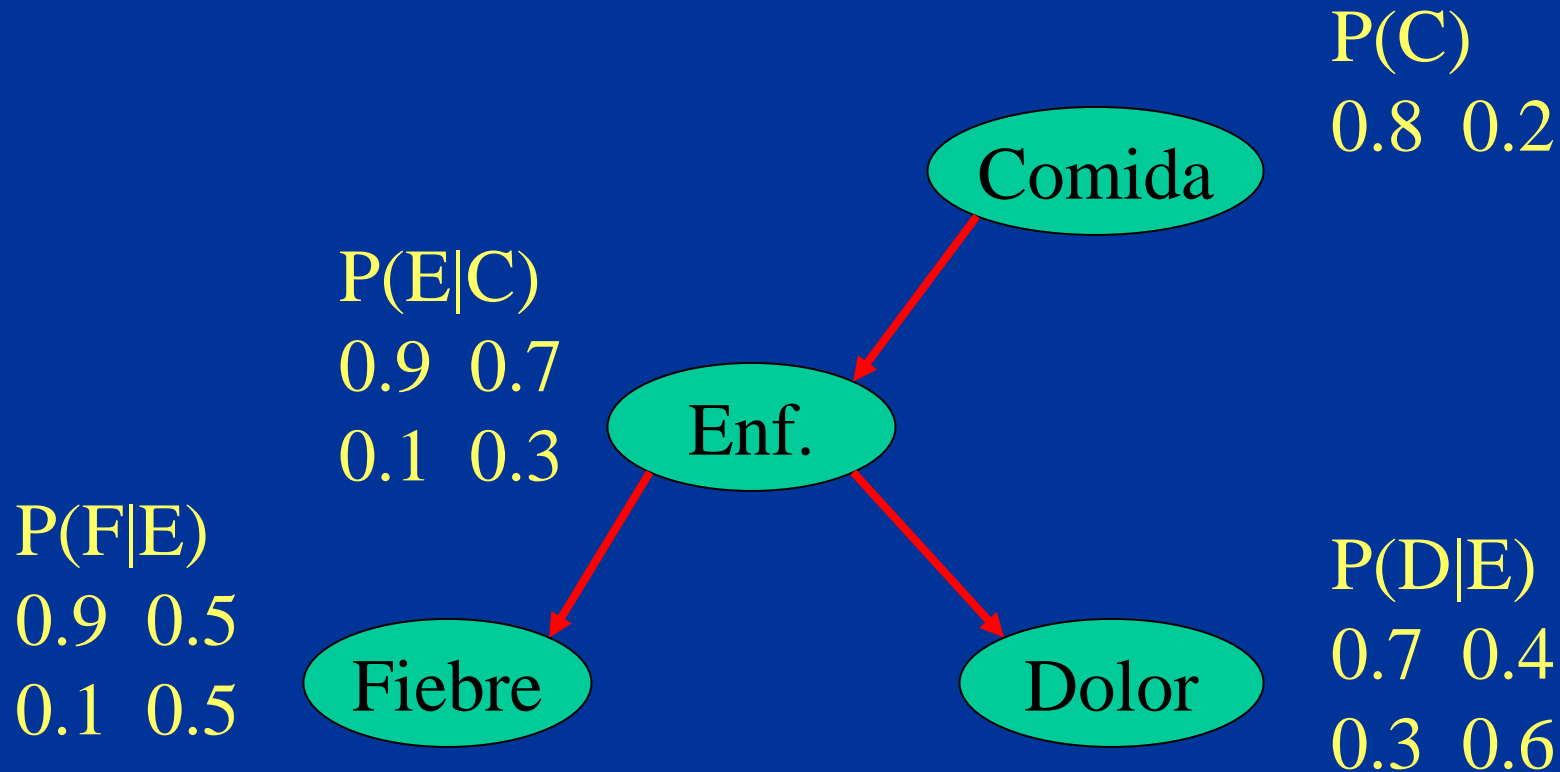
Ejemplo



Ejemplo



Ejemplo de cálculo



Ejemplo

- Probabilidad conjunta:

$$P(C,E,F,D) = P(C) P(E|C) P(F|E) P(D|E)$$

- Para calcular la P de enfermedad (E) dado fiebre ($F=f_1$)

$$P(E|F) = P(E,F) / P(F)$$

- Donde:

$$P(E,F) = \sum_c \sum_d P(C) P(E|C) P(F|E) P(D|E)$$

Ejemplo

- Reordenando:

$$P(E,F) = \sum_d P(F|E) P(D|E) \sum_c P(C) P(E|C)$$

- Hay que calcular esto para cada valor de E, dado f_1 . Para el caso e_1, f_1 :

$$\begin{aligned} P(e_1, f_1) &= \sum_d P(f_1|e_1) P(D|e_1) \sum_c P(C) P(e_1|C) \\ &= \sum_d P(f_1|e_1) P(D|e_1) [.9 \times .8 + .7 \times .2] \\ &= \sum_d P(f_1|e_1) P(D|e_1) [.86] \\ &= [.9 \times .7 + .9 \times .3] [.86] = [.9] [.86] = .774 \end{aligned}$$

Ejemplo

- En forma similar se calcula $P(e_2, f_1)$
- Luego, a partir de estos valores, se calcula la $P(f_1) = \sum_e P(E, F)$
- Finalmente se obtienen las probabilidades condicionales, $P(e_1|f_1)$ y $P(e_2|f_1)$

Conclusiones

- Como veremos más adelante:
 - Cada término que se suma en el algoritmo corresponde a un clique del grafo
 - El grafo que se obtiene con los arcos adicionales corresponde al grafo triangulado requerido para los algoritmos de agrupamiento
- La principal desventaja de este algoritmo es que se restringe a una variable, veremos ahora otros algoritmos que no tienen esta restricción

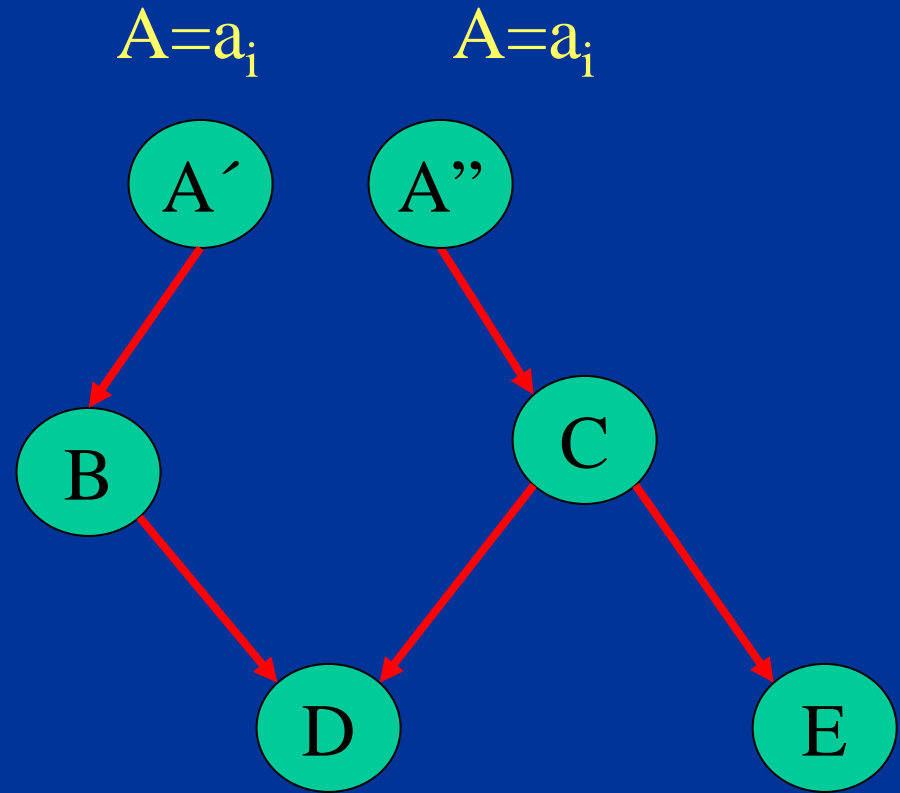
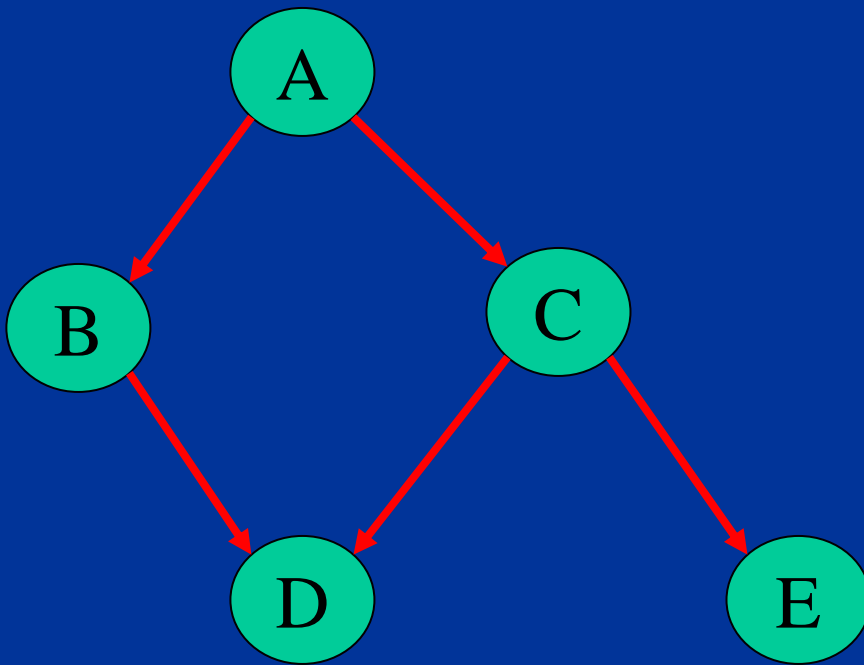
Cálculo de todas las variables

- Hay 3 tipos de métodos para calcular las probabilidades posteriores de todas las variables no conocidas en redes multiconectadas:
 - **Condicionamiento**
 - **Simulación estocástica**
 - **Agrupamiento**

Condicionamiento

- Si *instanciamos* (asignamos un valor) una variable, ésta bloquea las trayectorias de propagación.
- Entonces, asumiendo valores para un grupo seleccionado de variables podemos descomponer la gráfica en un conjunto de redes conectadas en forma sencilla.
- Propagamos para cada valor posible de dichas variables y luego promediamos las probabilidades ponderadas.

Condicionamiento



Procedimiento

- Al “cortar” en A, la probabilidad de cualquier variable (b) la podemos obtener mediante la regla de probabilidad total:

$$P(b|E) = \sum_i P(b|a_i, E) P(a_i|E)$$

- Donde:
 - $P(b|a_i, E)$: probabilidad posterior por propagación para cada valor de A
 - $P(a_i|E)$: “peso”

Procedimiento

- $P(a_i|E)$: “peso”, por la regla de bayes:

$$P(a_i|E) = \alpha P(a_i) P(E|a_i)$$

- Donde:

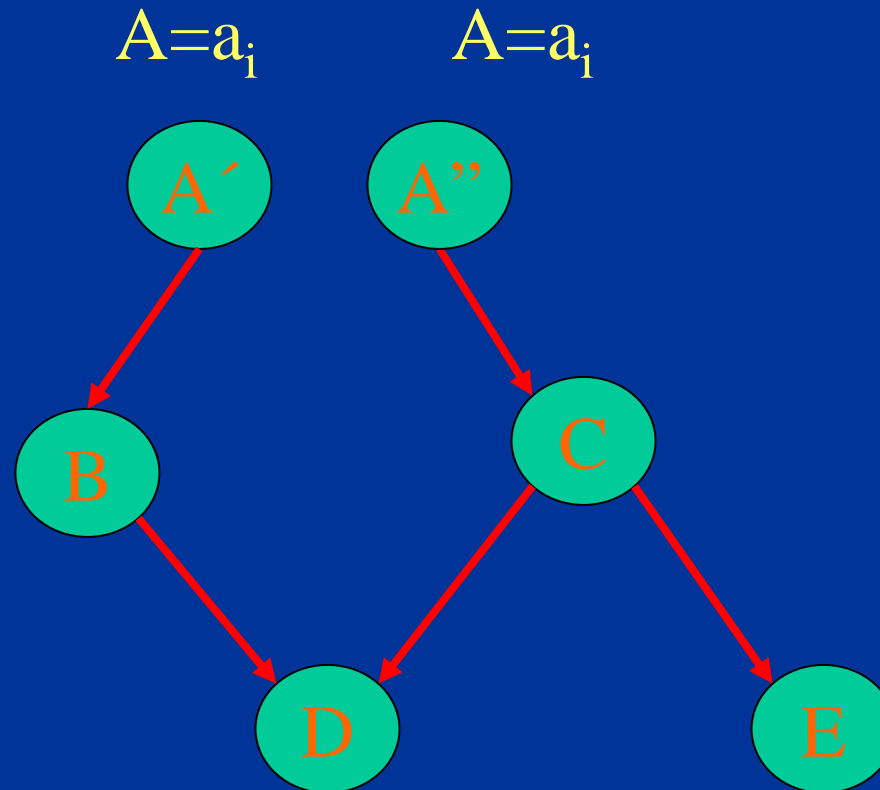
- el 1er término es la probabilidad *a priori* de A, se obtiene propagando sin evidencia
- El 2do término es la probabilidad del nodo evidencia dado A, se obtiene propagando sólo con A asignado

- Si hay varios nodos evidencia, el “peso” se obtiene en forma recursiva por la regla de bayes:

$$P(a_i|e1) = \alpha P(a_i) P(e1|a_i)$$

$$P(a_i|e1,e2) = \alpha P(a_i |e1) P(e2|a_i ,e1) , \dots$$

Ejemplo



- Considerando 2 valores para A y dado $D=0, E=1$

Ejemplo

1. Obtener P previa de A (en esta caso conocidas)
2. Obtener probabilidades de D y E dado cada valor de A : $P(D|A)$, $P(E|A)$, $A=0,1$
3. Propagar evidencia $E=1$, obtener pesos: $P(a|e)$, y probabilidades con $A=0,1$; por propagación
4. Propagar evidencia $D=0$, obtener pesos: $P(a|e,d)$, y probabilidades con $A=0,1$; por propagación
5. Obtener probabilidad posterior combinando los pesos y probabilidades con A instanciado

Simulación estocástica

- Se asignan valores aleatorios a las variables no instanciadas, se calcula la distribución de probabilidad y se obtienen valores de cada variable dando una muestra.
- Se repite el procedimiento para obtener un número apreciable de muestras y en base al número de ocurrencias de cada valor se determina la probabilidad de dicha variable.

Muestreo Lógico

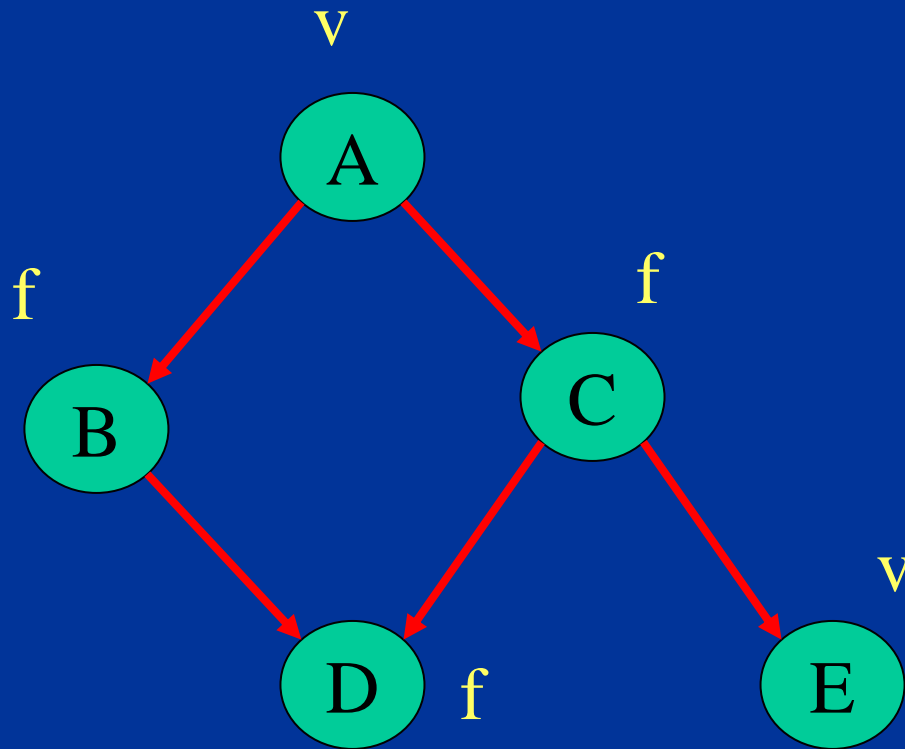
Para “N” muestras, repetir:

1. Dar valores aleatorios a los nodos raíz de acuerdo a sus probabilidades
2. En base a los valores anteriores, dar valores aleatorios a las siguientes variables (hijos de los nodos raíz) en función de la probabilidad condicional
3. Repetir (2) hasta llegar a los nodos hoja

Obtener probabilidades posteriores como frecuencias

- Si hay nodos evidencia, sólo considerar las muestras que correspondan a dichos valores

Muestreo Lógico: ejemplo



vfffv

fvvff

vffvf

ffvfv

vfvvf

ffffv

fvvvf

fffff

fffvf

vvvvf

Ejemplo

- Sin evidencia:
 - $P(A=V) = 4/10 = 0.4$
 - $P(B=V) = 3/10 = 0.3$
 - $P(C=V) = 5/10 = 0.5$
 - $P(D=V) = 5/10 = 0.5$
 - $P(E=V) = 3/10 = 0.3$
- Con evidencia: $D=V$ (aplican 5 muestras):
 - $P(A=V) = 3/5 = 0.6$
 - $P(B=V) = 2/5 = 0.4$
 - $P(C=V) = 3/5 = 0.6$
 - $P(E=V) = 1/5 = 0.2$

Muestreo pesado (*likelihood weighting*)

- Cuando se tiene evidencia, se “pierden” muchas muestras con muestreo lógico
- Una mejora es mantener todas y darles un peso de acuerdo a la probabilidad posterior de la evidencia en cada muestra

- Peso:

$$W(E|m) = P(e_1) P(e_2) \dots P(e_n)$$

donde la $P(\)$ es la probabilidad de acuerdo a sus padres

- La probabilidad se estima como la suma de los pesos de cada valor entre la suma de pesos total

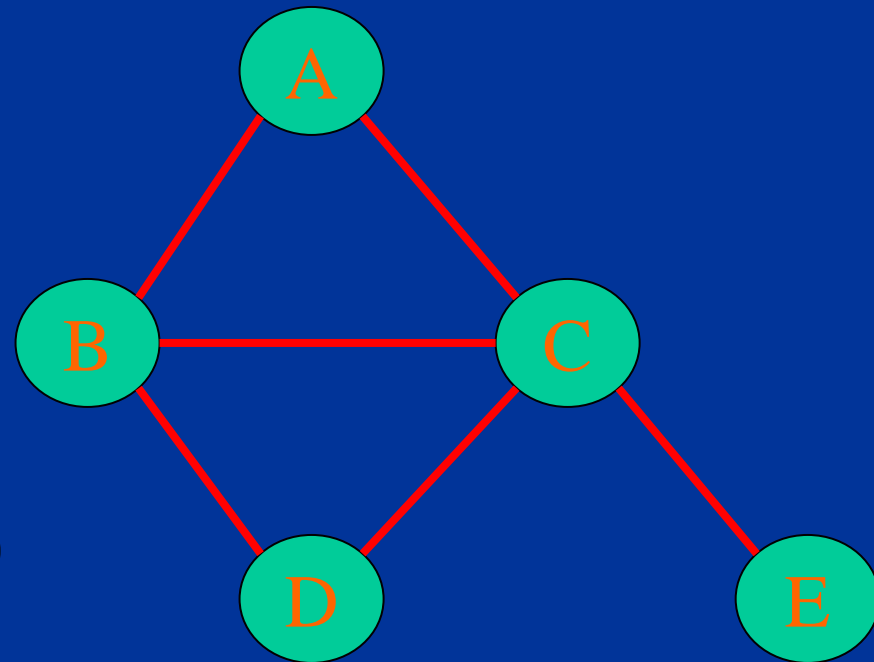
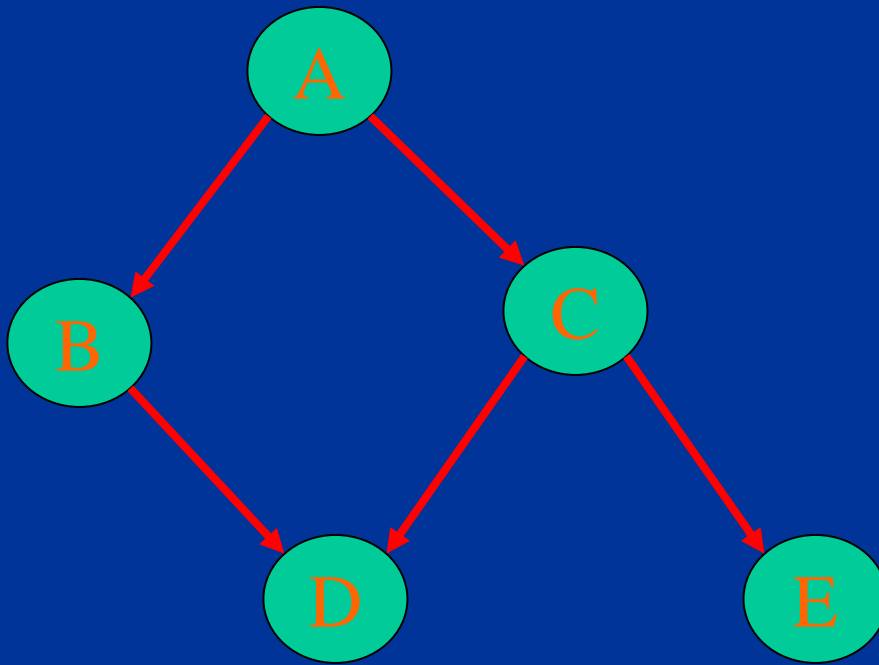
Agrupamiento

- El método de agrupamiento consiste en transformar la estructura de la red para obtener un árbol, mediante agrupación de nodos usando la teoría de grafos.
- La propagación se realiza sobre el árbol de macro-nodos obtenido, donde cada macro-nodo corresponde a un clique o *unión* de la RB original (*junction tree*)

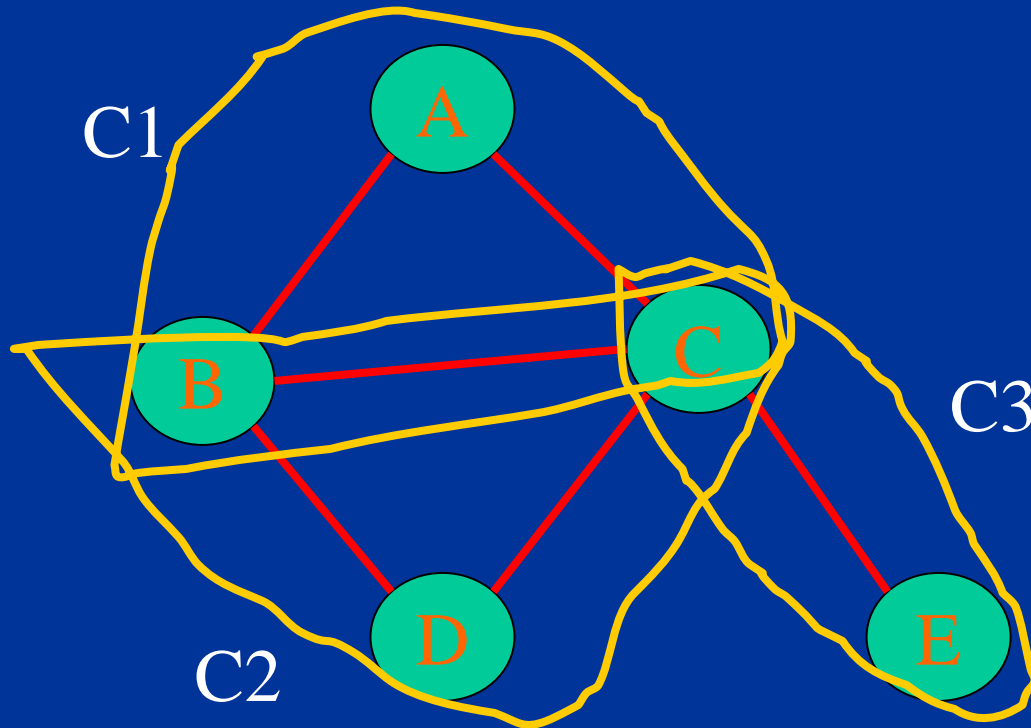
Agrupamiento

- Transformación:
 - Eliminar direccionalidad de los arcos
 - Ordenamiento de los nodos por máxima cardinalidad
 - Moralizar el grafo (arco entre nodos con hijos comunes)
 - Triangular el grafo
 - Obtener los *cliques* y ordenar
 - Construir árbol de *cliques*

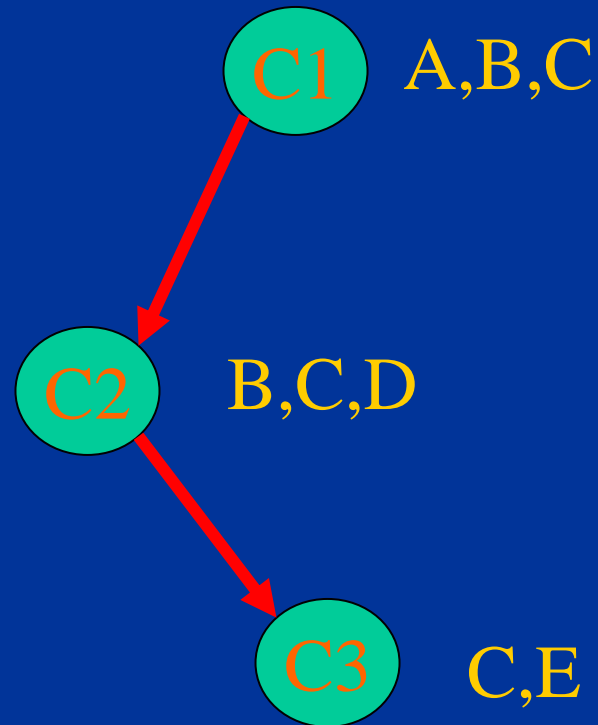
Ejemplo



Ordenamiento de Cliques



Árbol de Cliques



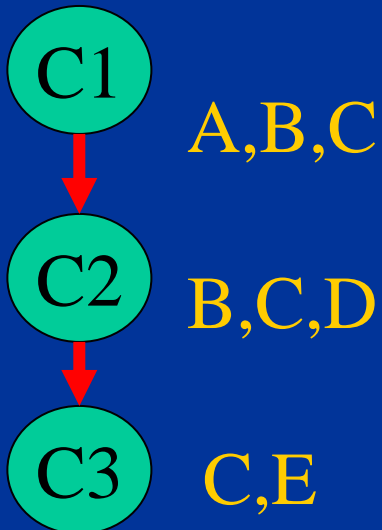
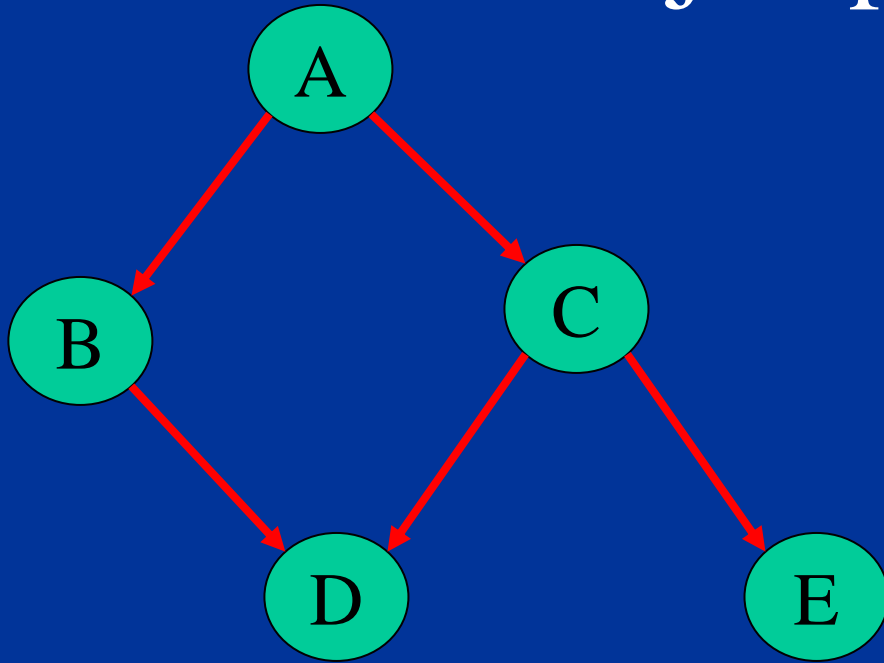
Propagación

- La propagación es mediante el envío de mensajes en el árbol de *cliques* (en forma similar a árboles)
- Inicialmente se calcula la probabilidad conjunta (potencial) de cada *clique*, y la condicional dado el padre
- Dada cierta evidencia se recalculan las probabilidades de cada *clique*
- La probabilidad individual de cada variable se obtiene de la del *clique* por marginalización

Procedimiento – preprocesamiento:

1. Se obtienen los conjuntos de nodos de cada *clique* – C_i
2. Se obtienen los conjuntos de nodos comunes con *cliques* previos – S_i
3. Se obtienen los conjuntos de nodos que están en C_i pero no en S_i : $R_i = C_i - S_i$
4. Se calcula la probabilidad (potencial) de cada *clique* – $\psi(\text{clqi}) = \prod P(\text{nodos})$

Ejemplo



- C:
 - A, B, C
 - B, C, D
 - C, E

- S:
 - \emptyset
 - B, C
 - C

- R:
 - A, B, C
 - D
 - E

- Ψ :
 - $P(A) P(B|A) P(C|A)$
 - $P(D|B, C)$
 - $P(E|C)$

Propagación sin evidencia:

- Cálculo de λ :

$$\lambda(C_i) = \sum_R \Psi(C_i)$$

- Actualización:

$$\Psi(C_i)' = \Psi(C_i) / \lambda(C_i)$$

- Enviar λ a padre

Propagación sin evidencia:

- Cálculo de π . Para todos los hijos “i” del clique “j”:

$$\pi(C_i) = \sum_{C_j - S_i} P'(C_i)$$

- Enviar π a cada hijo

Propagación sin evidencia:

- Propagación de λ :
 - Calcular λ para los clique hoja
 - Actualizar potencial del clique “j” al recibir cada λ de los hijos “i”:

$$\Psi(C_j)' = \lambda(C_i) \Psi(C_j)$$

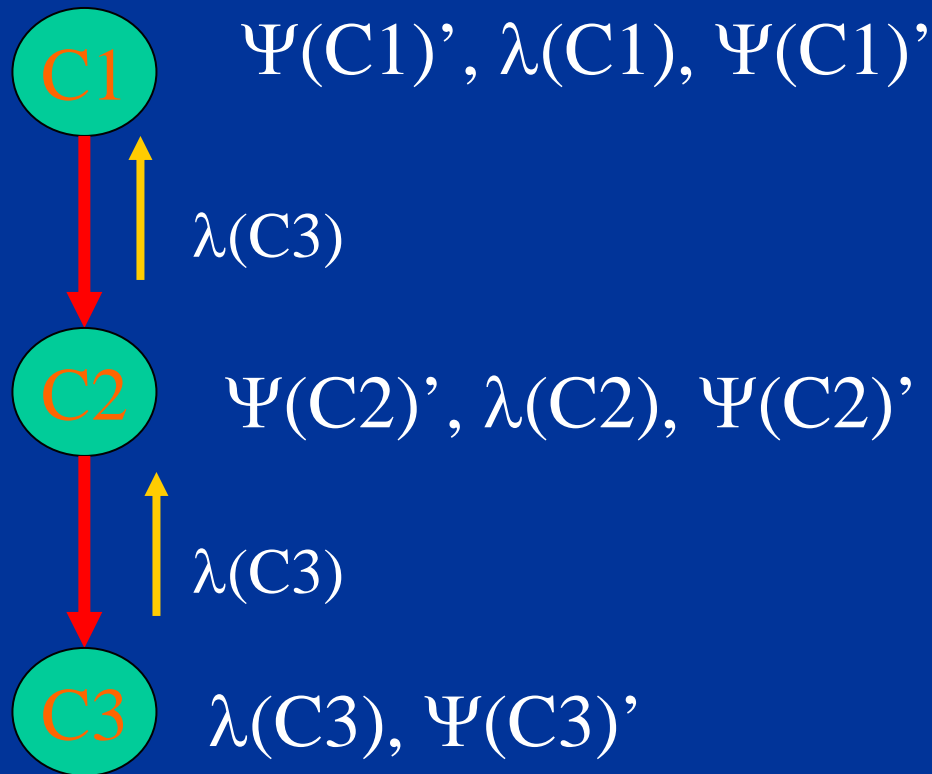
- Al recibir todas la λ propagar al padre:
- Al llegar al nodo raíz obtener P' :

$$P'(C_j) = \Psi(C_j)'$$

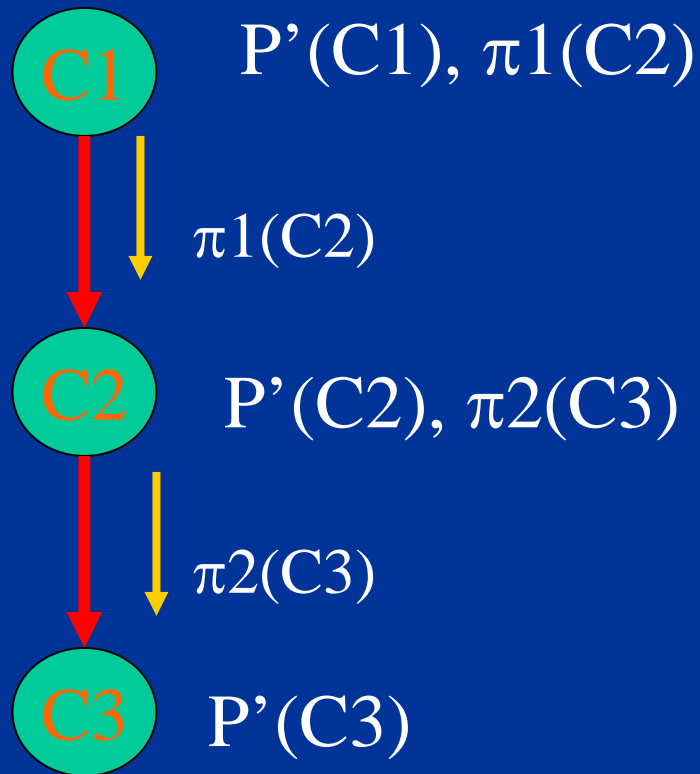
Propagación sin evidencia:

- Propagación de π :
 - Obtener π del clique raíz para cada hijo
 - Enviar π a cada hijo
 - Actualizar $P'(C_i)$:
$$P'(C_i) = \pi(C_i) \Psi'(C_i)$$
 - Enviar π a cada hijo hasta llegar a los nodos hoja

Ejemplo – propagación λ



Ejemplo – propagación π



Propagación con evidencia:

- Cuando hay nodos conocidos (Evidencia – E), se actualizan los potenciales, R y S de cada clique en función de la evidencia:
 - $CLQ_i = CLQ_i - \{E\}$
 - $S_i = S_i - \{E\}$
 - $R_i = R_i - \{E\}$
- Se obtienen los potenciales para cliques con nodos evidencia de acuerdo a los valores de dichos nodos:
 - $\Psi(C_i)' = \Psi(C_i)' \leftarrow E = \text{evidencia}$
- Después se sigue el mismo proceso de propagación que sin evidencia

Ejemplo:

- Supongamos que se conocen D y E:
 - C: {A,B,C}, {B,C}, {C}
 - S: { \emptyset }, {B,C}, {C}
 - R: {A,B,C} { \emptyset }, { \emptyset }
- Potenciales:
 - $\Psi(\{A,B,C\})$
 - $\Psi(\{B,C\}) \leftarrow D=di$
 - $\Psi(\{C\}) \leftarrow E=ei$

Probabilidades de las variables:

- Se obtienen a partir de las probabilidades de los cliques por marginalización:

$$P(X) = \sum_{Y,Z,\dots} P'(\text{clq})$$

- En el ejemplo:

$$P(A) = \sum_{B,C} P'(\text{clq-1})$$

$$P(B) = \sum_{A,C} P'(\text{clq-1})$$

$$P(C) = \sum_{A,B} P'(\text{clq-1})$$

$$P(D) = \sum_{B,C} P'(\text{clq-2})$$

$$P(E) = \sum_C P'(\text{clq-3})$$

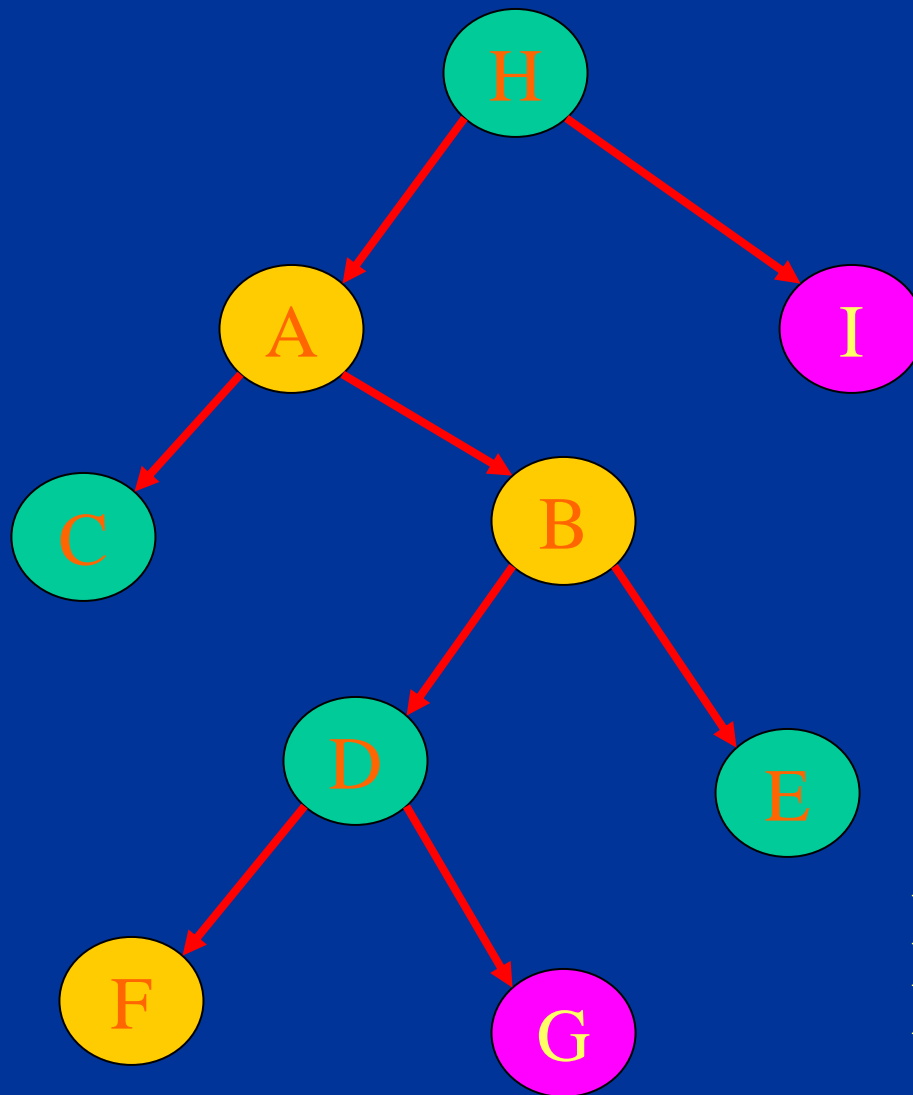
Complejidad

- En el peor caso, la propagación en redes bayesianas es un problema NP-duro
- En la práctica, en muchas aplicaciones se tienen redes no muy densamente conectadas y la propagación es eficiente aún para redes muy grandes (función del *clique* mayor)
- Para redes muy complejas (muchas conexiones), la mejor alternativa son técnicas de simulación estocástica o técnicas aproximadas

Abducción

- La “abducción” se define como encontrar la mejor “explicación” (valores de un cierto conjunto de variables) dada cierta evidencia
- Normalmente se buscan los valores del conjunto “explicación” que tiene mayor probabilidad
- En general, el conjunto de mayor probabilidad NO es igual a los valores individuales de mayor probabilidad

Abducción



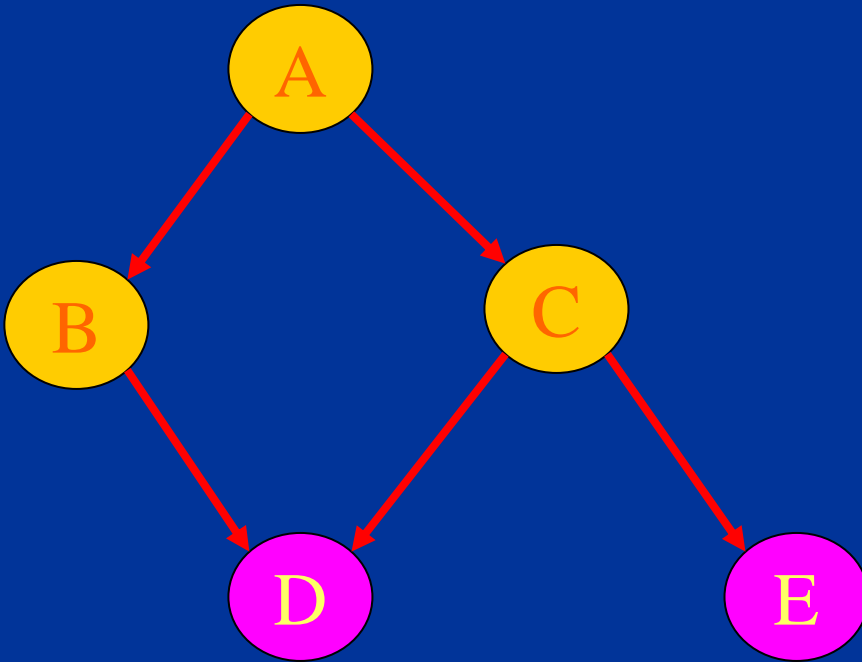
Ejemplo:
 $\text{Max } P(A, B, F | G, I)$

Procedimiento

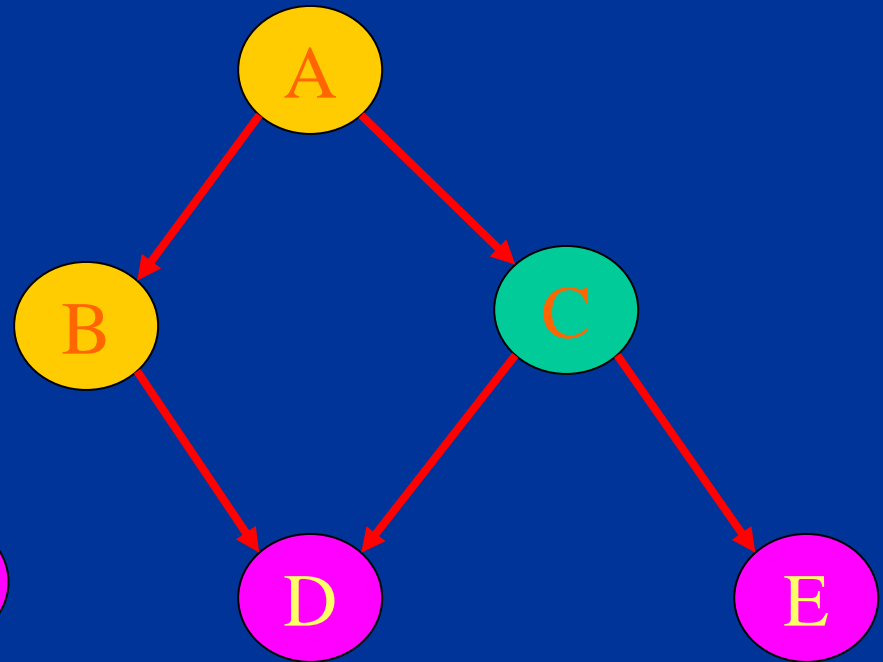
- Caso 1: conjunto explicación = todas las variables no instanciadas
 - Mismo algoritmo que para propagación substituyendo sumatoria por MAX
- Caso 2 : conjunto explicación = cualquier subconjunto de variables no asignadas
 - Se utiliza el mismo algoritmo, usando MAX para las variables explicación y sumatoria para las demás

Ejemplo

Caso 1



Caso 2



Ejemplo

- Caso 1: D,E – evidencia, A,B,C – explicación

$$\max P(A,B,C|D,E)$$

- Caso 2: D,E – evidencia, A,B– explicación

$$\max P(A,B|D,E) =$$

$$\max [\sum_C P(A,B,C|D,E)]$$

Referencias

- Pearl 88 – Cap. 4,5
- Neapolitan 90 – Cap. 6,7,8
- Notas Jordan – Cap. 3

Actividades

- Leer referencias sobre redes bayesianas
- Examen parcial