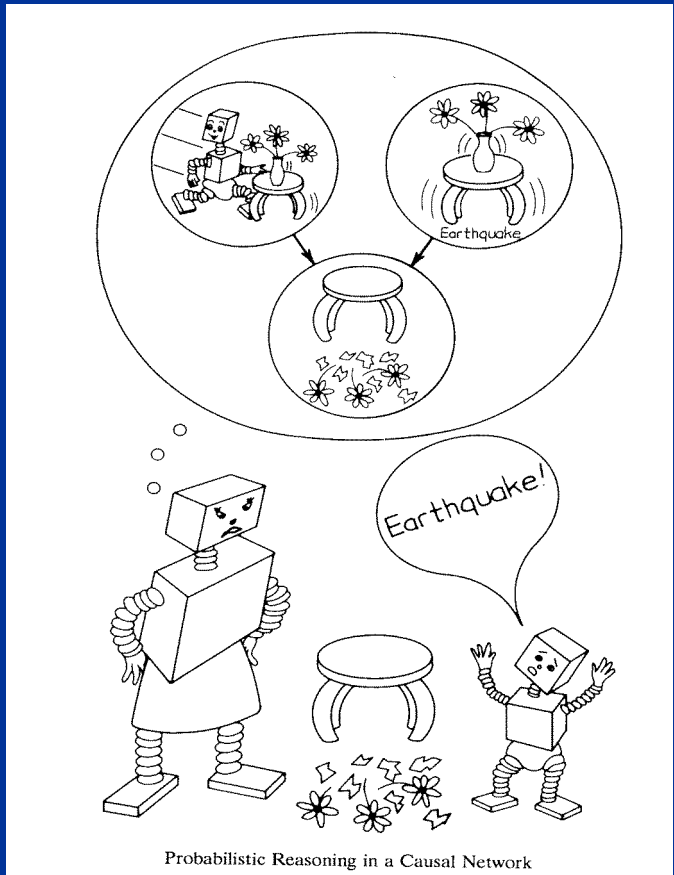
Modelos Gráficos Probabilistas

L. Enrique Sucar
INAOE

# Sesión 7:
# Redes Bayesianas – Inferencia:

# 2da Parte




[Neapolitan 90]

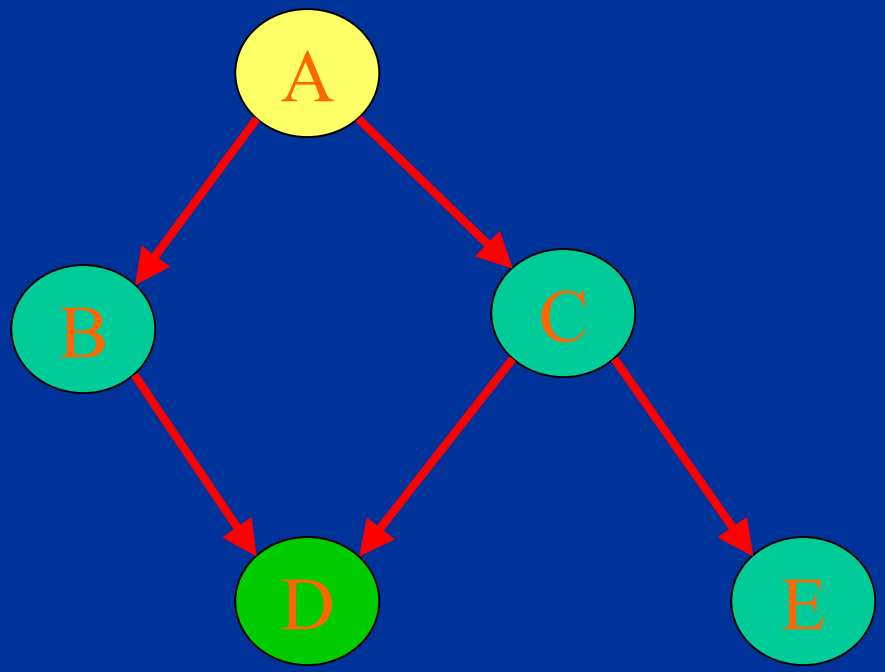
# Otros métodos de inferencia – redes multiconectadas

- Algoritmo para una variable:
  - Eliminación
- Algoritmos para todas las variables:
  - 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:

  $P(A \mid D)$



# 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:

$$P(XF \mid XE) = P(XE, XF) / P(XE)$$

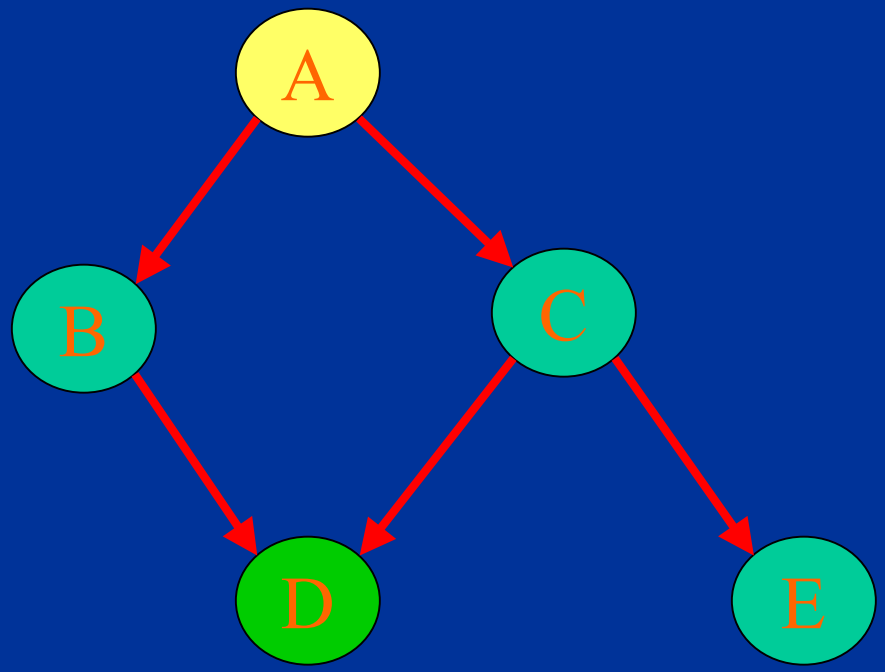
$$P(XE, XF) = \Sigma_{XR} P(XE, XF, XR)$$

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

# Eliminación

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

# Ejemplo




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

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

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):

$$P(A, D)= \Sigma_B \Sigma_C \Sigma_E P(A)P(B|A)P(C|A)P(\underline{D}|B,C)P(E|C)$$

$$P(A, D)= \Sigma_B \Sigma_C P(A)P(B|A)P(C|A)P(\underline{D}|B,C) \Sigma_E P(E|C)$$

$$P(A, D)= \Sigma_B P(A)P(B|A) \Sigma_C P(C|A)P(\underline{D}|B,C) \Sigma_E P(E|C)$$

$$P(A, D)= P(A) \Sigma_B P(B|A) \Sigma_C P(C|A)P(\underline{D}|B,C) \Sigma_E P(E|C)$$

# Ejemplo

- Si introducimos cierta notación:

$m_E(C) = \Sigma_E P(E|C)$, Entonces:

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

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

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

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

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

# Ejemplo

- De aquí podemos obtener P(D):

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

- Y entonces:

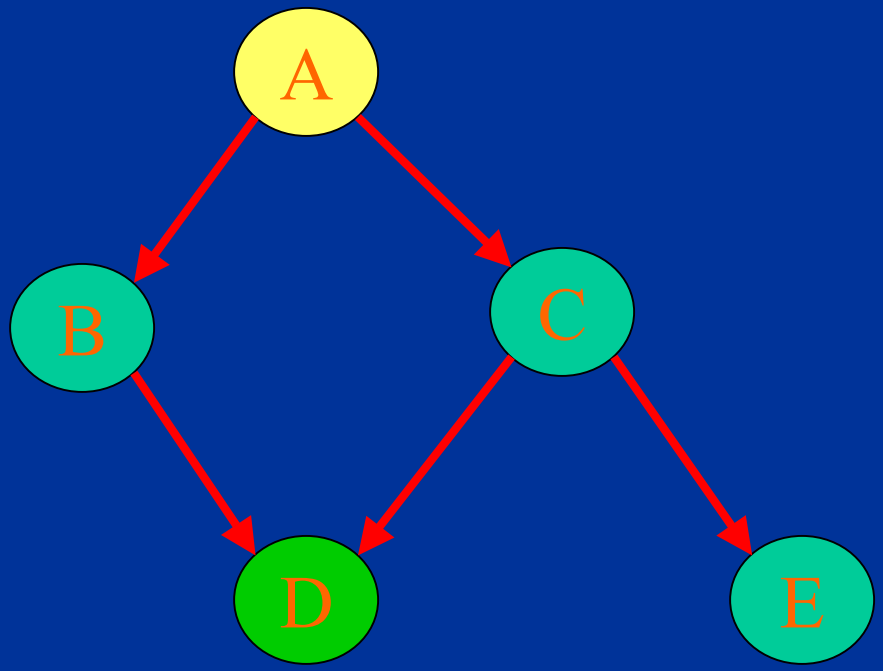
  $P(A|D) = P(A)\, m_B(A) / \Sigma_A P(A)\, m_B(A)$

- A partir de estas ideas se deriva el algoritmo general de ***Eliminación*** para inferencia en redes bayesianas

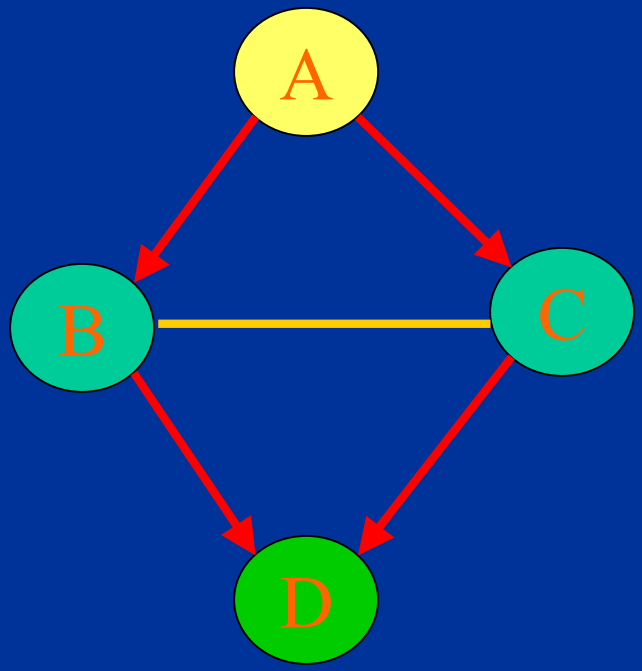
# Interpretación gráfica

- 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

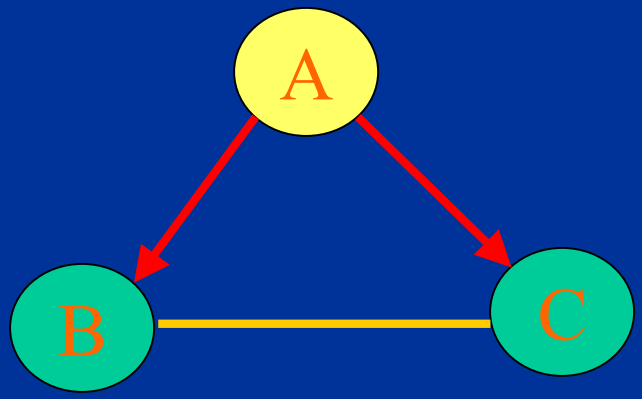
# Ejemplo



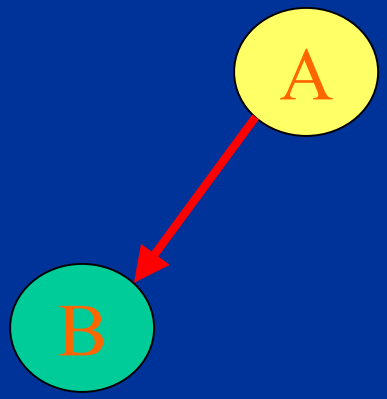
# Ejemplo



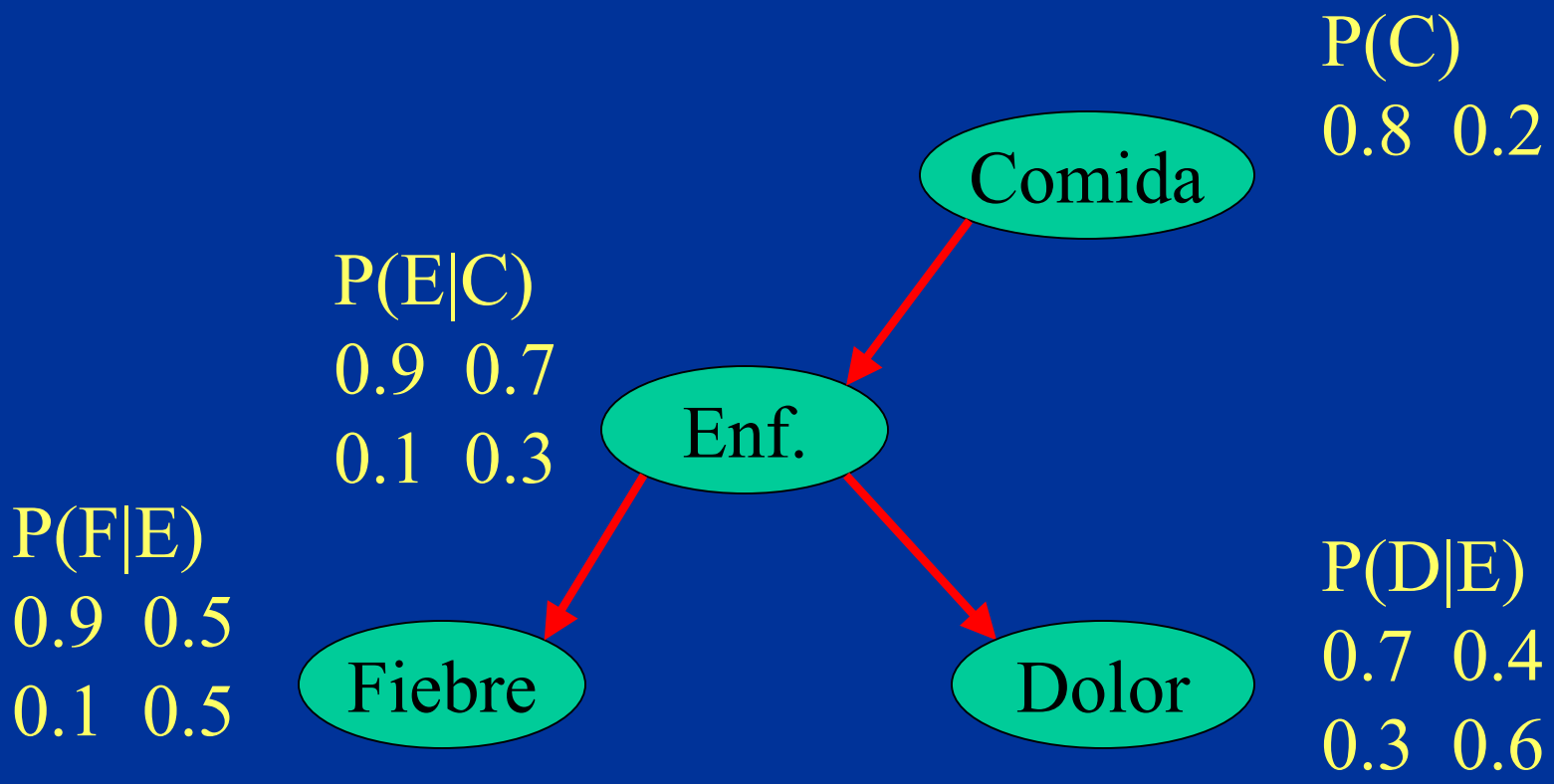
# Ejemplo



# 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:

$$P(E,F) = \Sigma_c \Sigma_d\, P(C)\, P(E|C)\, P(F|E)\, P(D|E)$$

# Ejemplo

- Reordenando:

$$P(E,F) = \Sigma_d\, P(F|E)\, P(D|E)\, \Sigma_c\, P(C)\, P(E|C)$$

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

$$P(e1,f1) = \Sigma_d\, P(f1|e1)\, P(D|e1)\, \Sigma_c\, P(C)\, P(e1|C)$$
$$= \Sigma_d\, P(f1|e1)\, P(D|e1)\, [.9 \text{ x. } 8 + .7 \text{ x } .2]$$
$$= \Sigma_d\, P(f1|e1)\, P(D|e1)\, [.86]$$
$$= [.9 \text{ x } .7 + .9 \text{ x } .3]\, [.86] = [.9]\, [.86] = .774$$

# Ejemplo

- En forma similar se calcula P(e2,f1)
- Luego, a partir de estos valores, se calcula la P(f1) = $\Sigma$e P(E,F)
- Finalmente se obtienen las probabilidades condicionales, P(e1|f1) y P(e2|f1)

# 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

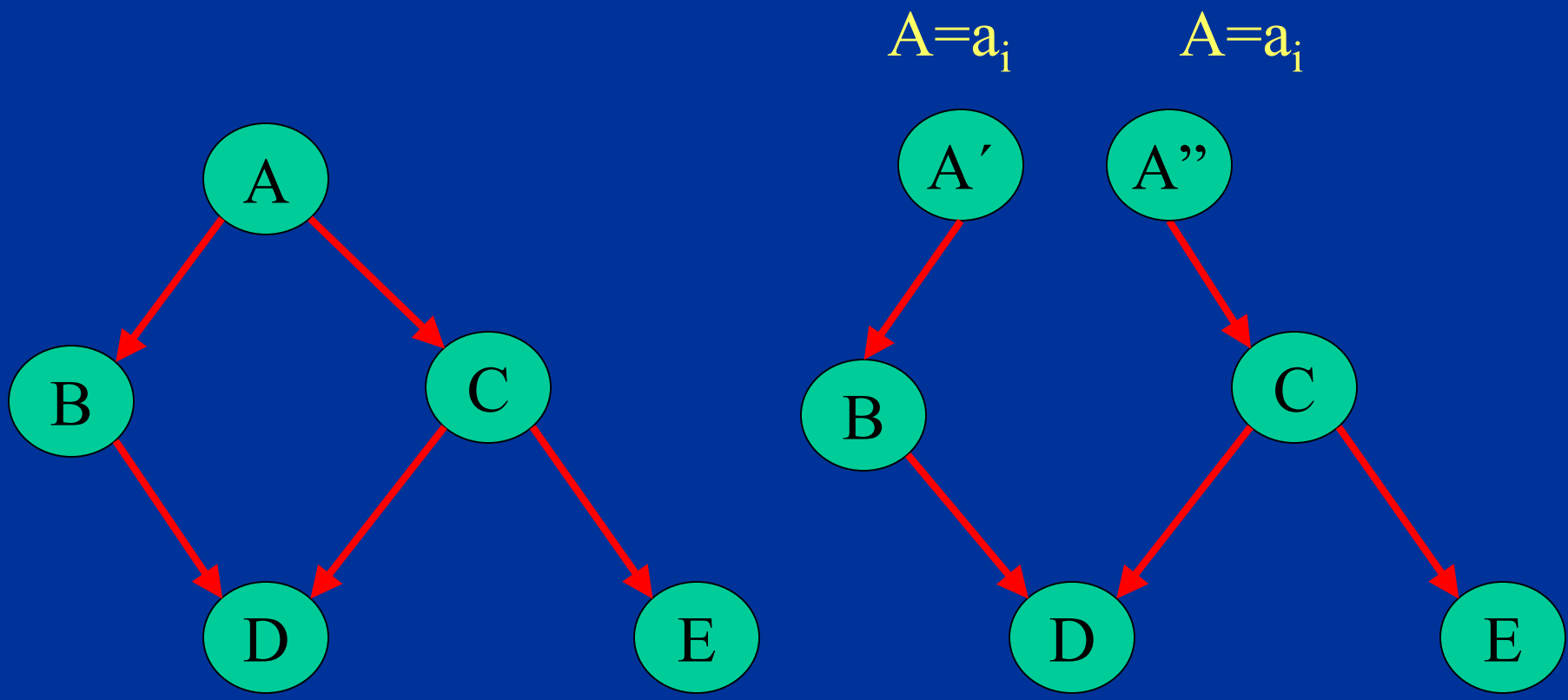
# Cálculo de todas las variables

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

# Condicionamiento

- Si *instanciamos* (asignamos un valor) a 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:

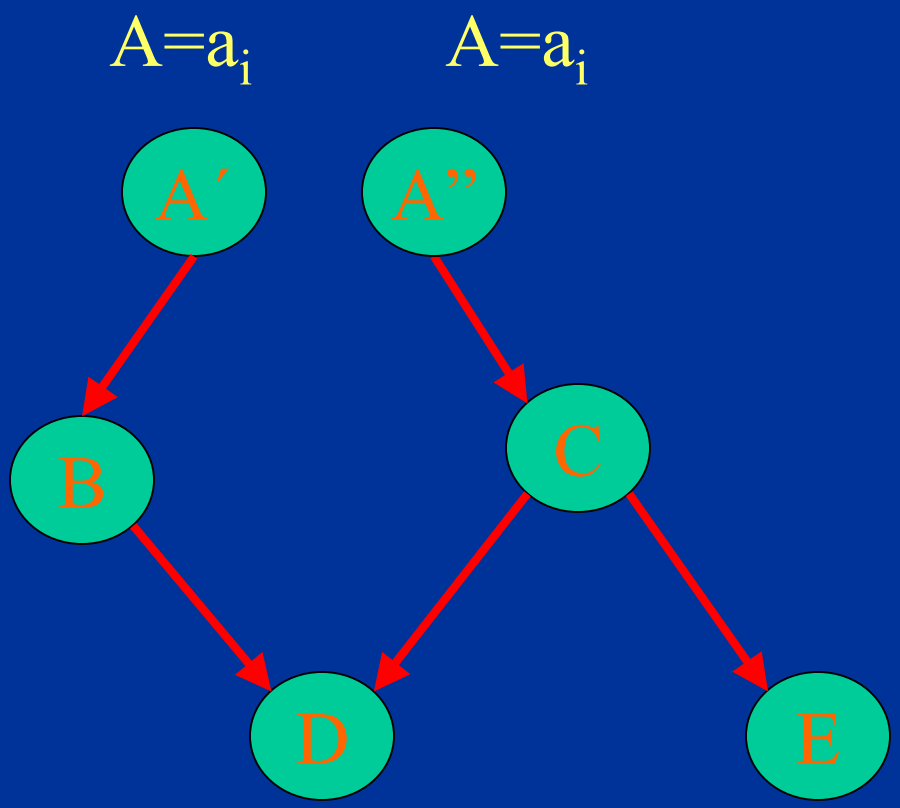
$$P(a_i|E) = \alpha\, P(a_i)\, P(E|a_i)$$

- 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:

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

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

# 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 asignadas, 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

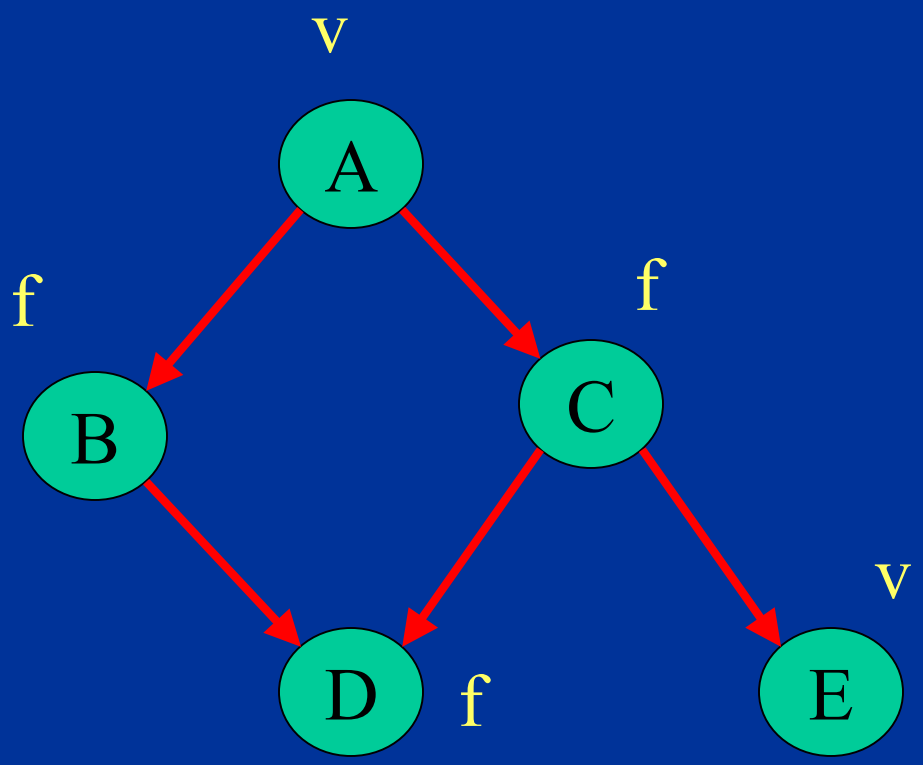
Para "N" muestras, repetir:

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:

$$W(E|m) = P(e1)\ P(e2)\ \ldots\ P(en)$$

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

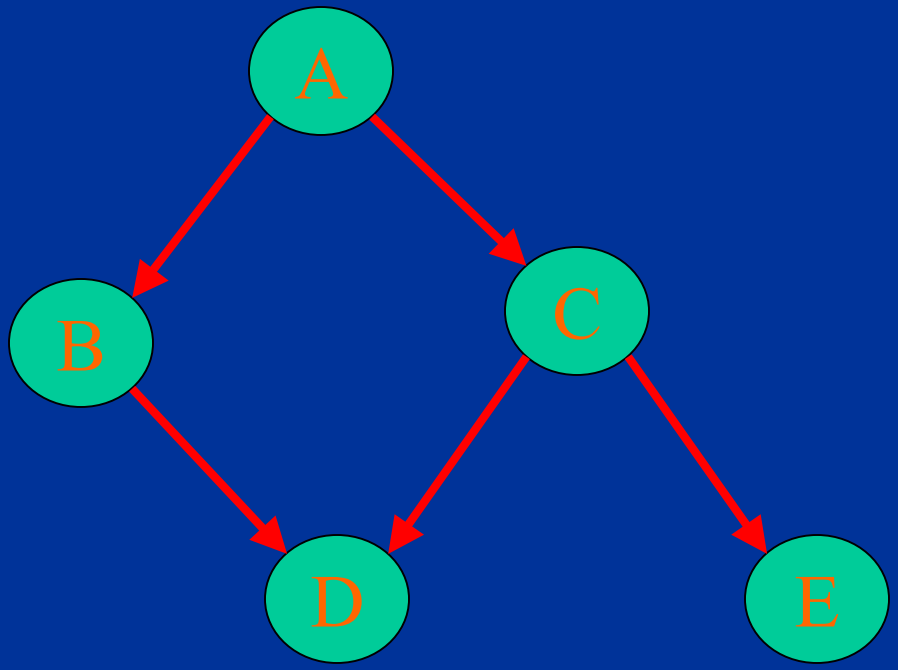
# Agrupamiento

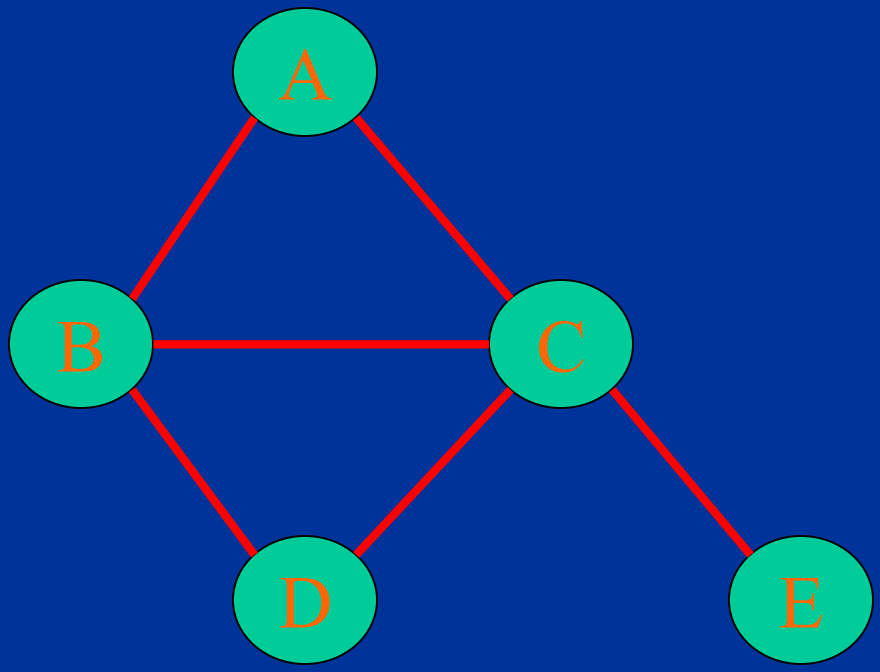
- El método de agrupamiento consiste en transformar la estructura de la red para obtener un árbol, mediante agrupación de nodos usando la teoría de grafos.
- La propagación se realiza sobre el árbol de macro-nodos obtenido, donde cada macro-nodo corresponde a un clique o *unión* de la RB original (*junction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

# Agrupamiento

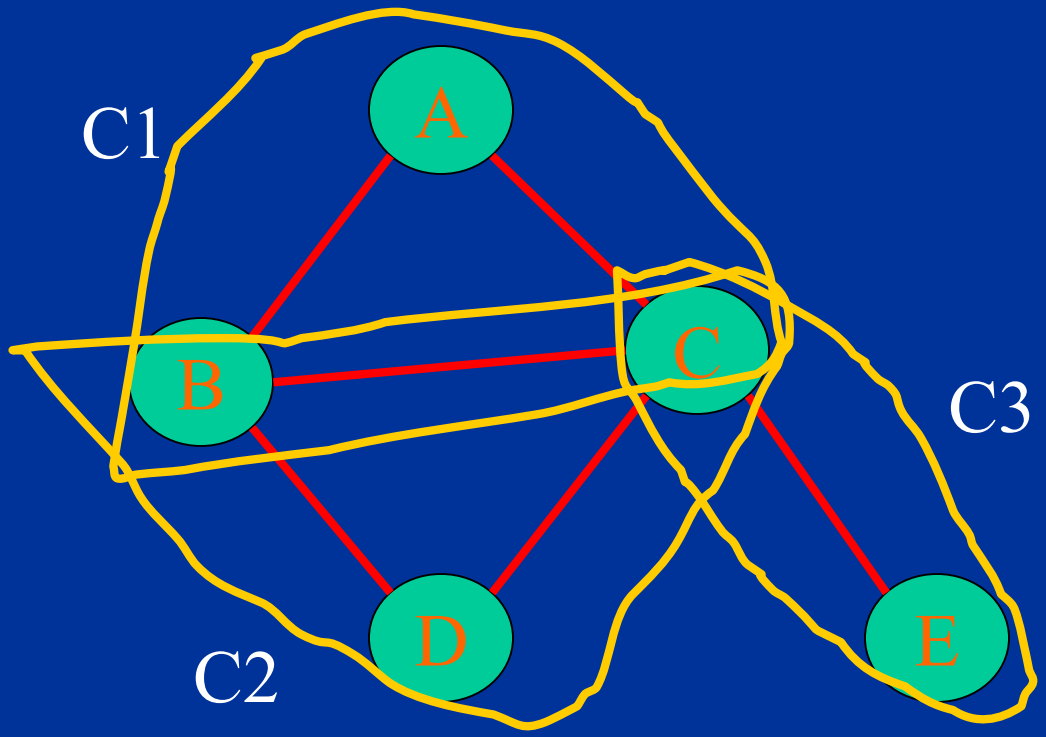
- Transformación:
  - Eliminar direccionalidad de los arcos
  - Ordenamiento de los nodos por máxima cardinalidad
  - Moralizar el grafo (arco entre nodos con hijos comunes)
  - Triangular el grafo
  - Obtener los *cliques* y ordenar
  - Construir árbol de *cliques*

# Ejemplo

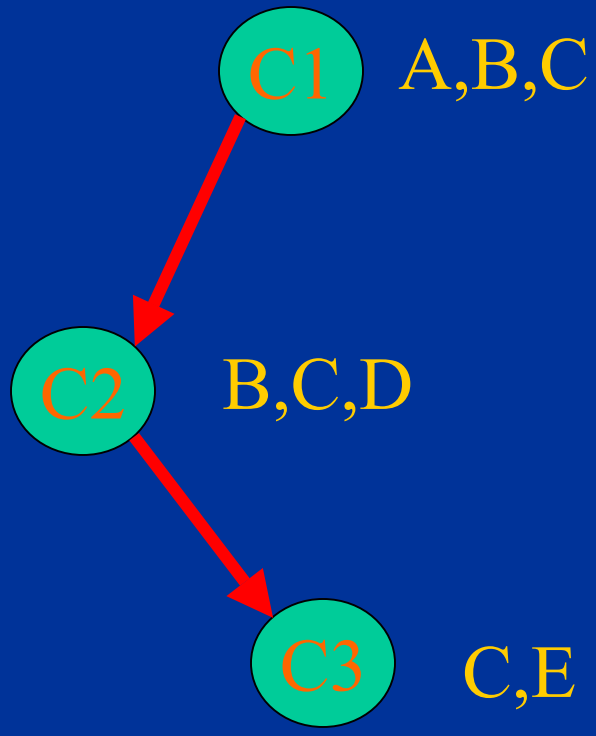


# Ordenamiento de Cliques



# Árbol de Cliques



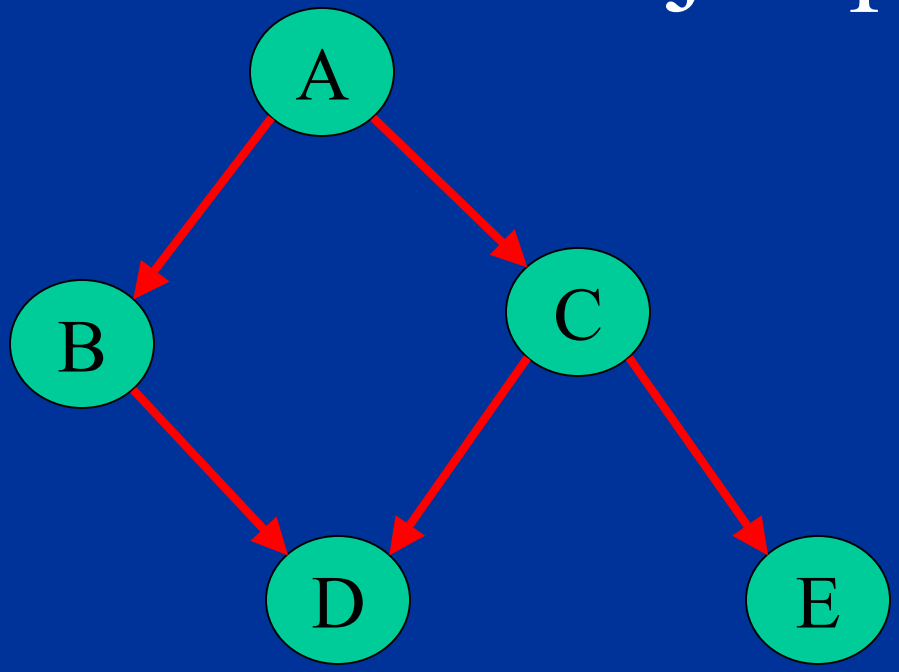
# Propagación

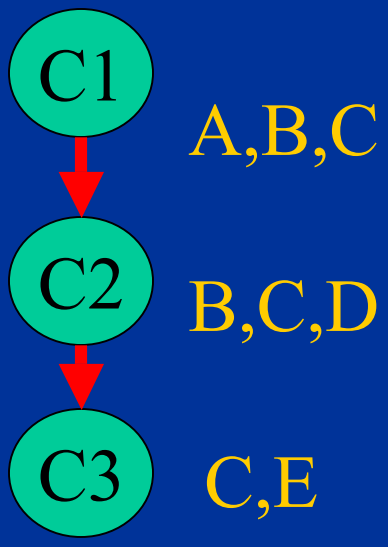
- La propagación es mediante el envío de mensajes en el árbol de *cliques* (en forma similar a árboles)
- Inicialmente se calcula la probabilidad conjunta (potencial) de cada *clique*, y la condicional dado el padre
- Dada cierta evidencia se recalculan las probabilidades de cada *clique*
- La probabilidad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

# Procedimiento – preprocesamiento:

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

# Ejemplo




- C:
  - A,B,C
  - B,C,D
  - C,E
- S:
  - 0
  - 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 $\lambda$ :

$$\lambda(Ci) = \Sigma_R \Psi(Ci)$$

- Actualización:

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

- Enviar $\lambda$ a padre

# Propagación sin evidencia:

- Cálculo de $\pi$ . Para todos los hijos "i" del clique "j":

$$\pi(Ci) = \Sigma_{Cj - Si} P'(Ci)$$

- Enviar $\pi$ a cada hijo

# Propagación sin evidencia:

- Propagación de $\lambda$ :
  - Calcular $\lambda$ para los clique hoja
  - Actualizar potencial del clique "j" al recibir cada $\lambda$ de los hijos "i":

$$\Psi(Cj)' = \lambda(Ci)\ \Psi(Cj)$$

  - Al recibir todas la $\lambda$ propagar al padre:
  - Al llegar al nodo raíz obtener P':

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

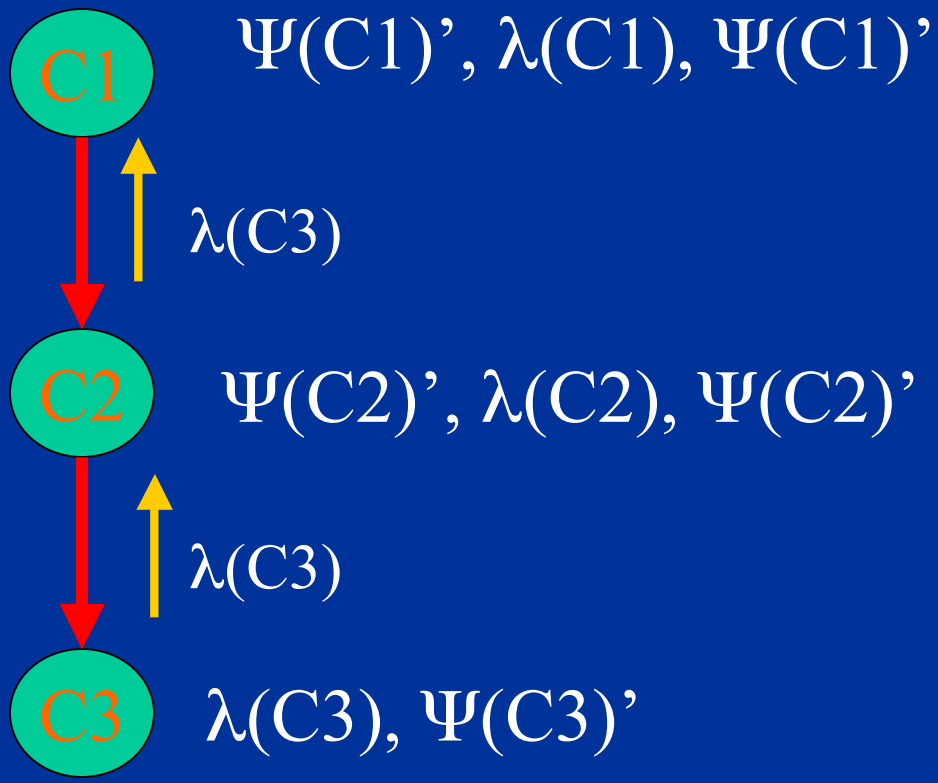
# Propagación sin evidencia:

- Propagación de $\pi$ :
  - Obtener $\pi$ del clique raíz para cada hijo
  - Enviar $\pi$ a cada hijo
  - Actualizar P'(Ci):

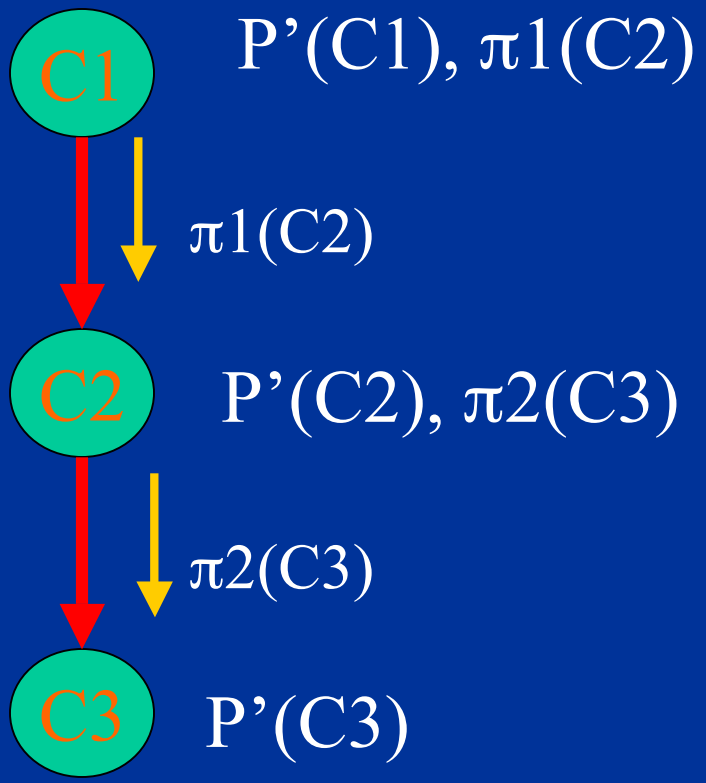
$$P'(Ci) = \pi(Ci)\ \Psi'(Ci)$$

  - Enviar $\pi$ 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:
  - CLQi = CLQi – {E}
  - Si = Si - {E}
  - Ri = Ri - {E}
- Se obtienen los potenciales para cliques con nodos evidencia de acuerdo a los valores de dichos nodos:
  - $\Psi(Ci)' = \Psi(Ci)' \leftarrow$ E=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: {0}, {B,C}, {C}
  - R: {A,B,C} {0}, {0}
- 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:

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

- En el ejemplo:

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

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

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

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

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

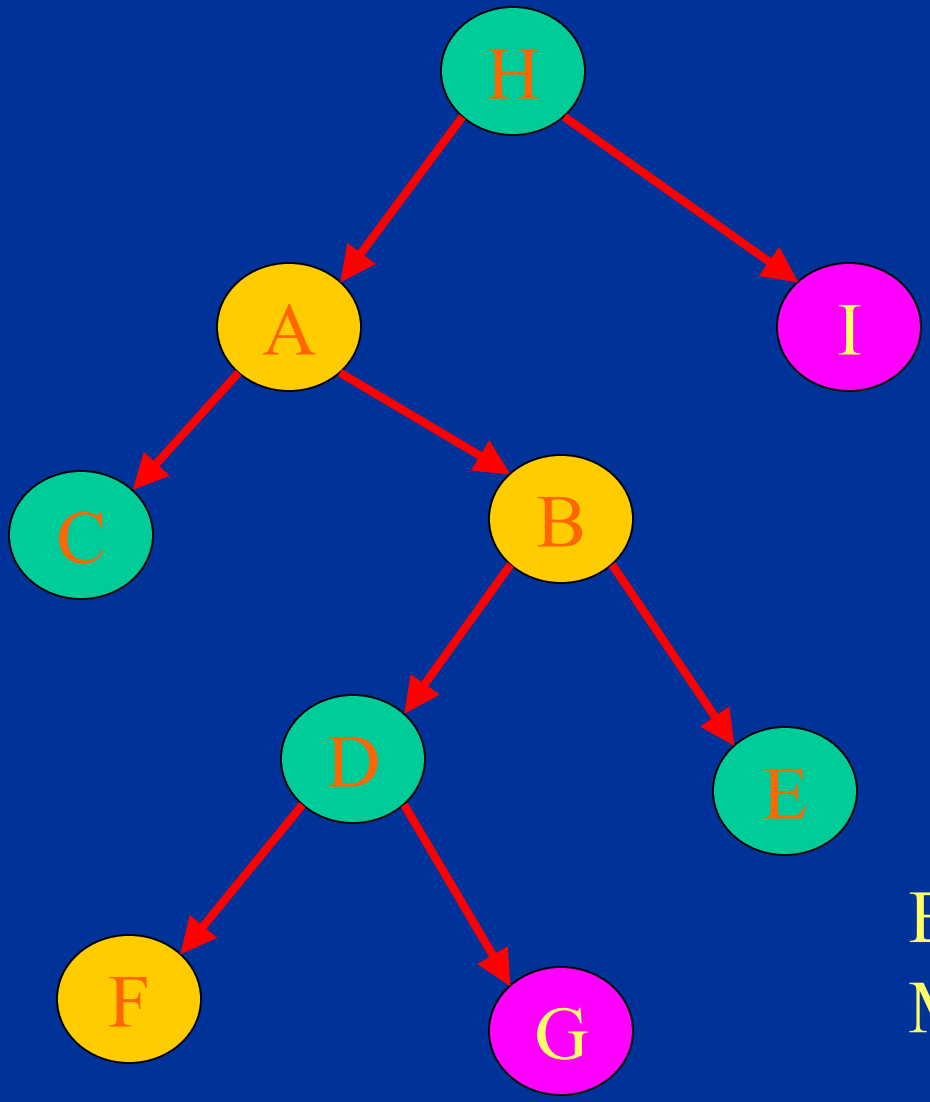
# Complejidad

- 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



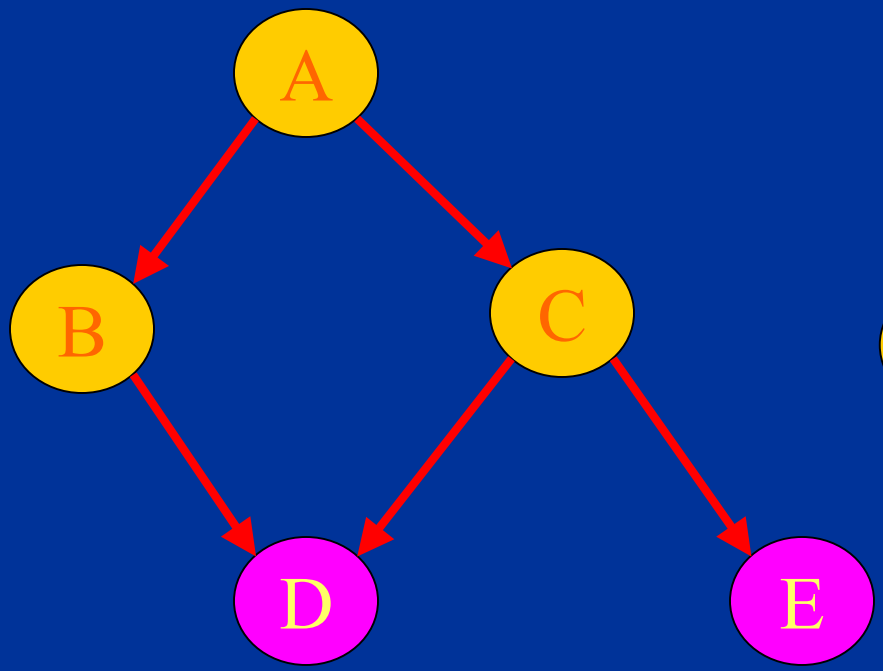

Ejemplo:
Max P(A,B,F|G,I)

# Procedimiento

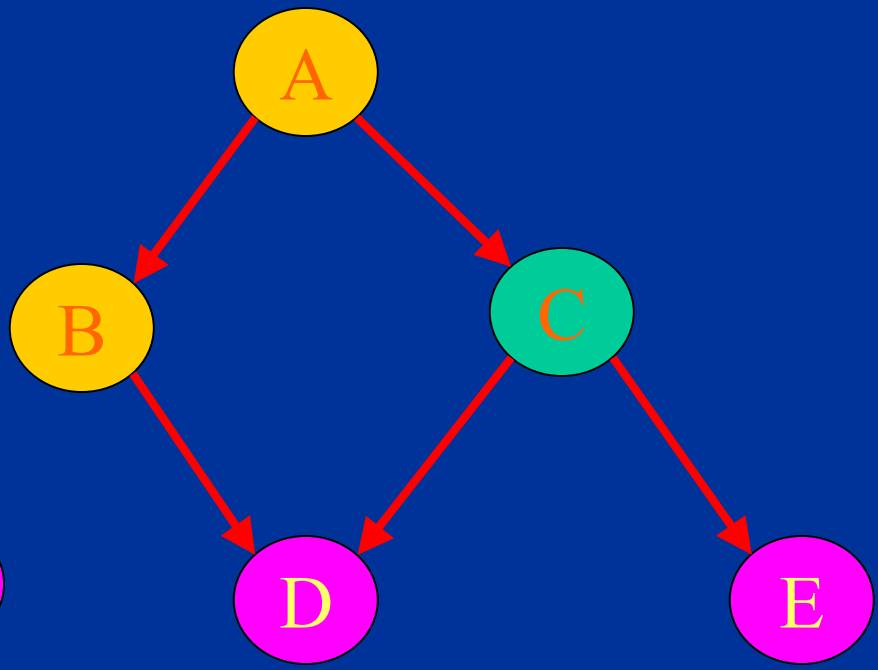
- Caso 1-Abducción total: conjunto explicación = todas las variables no instanciadas
  - Mismo algoritmo que para propagación substituyendo sumatoria por MAX
- Caso 2-Abducción parcial: conjunto explicación = cualquier subconjunto de variables no asignadas
  - Se utiliza el mismo algoritmo, usando MAX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

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

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

# Referencias

- Koller & Friedman - Cap. 9,10
- Pearl 88 – Cap. 4,5
- Neapolitan 90 – Cap. 6,7,8
- Jensen 01 – Cap. 5