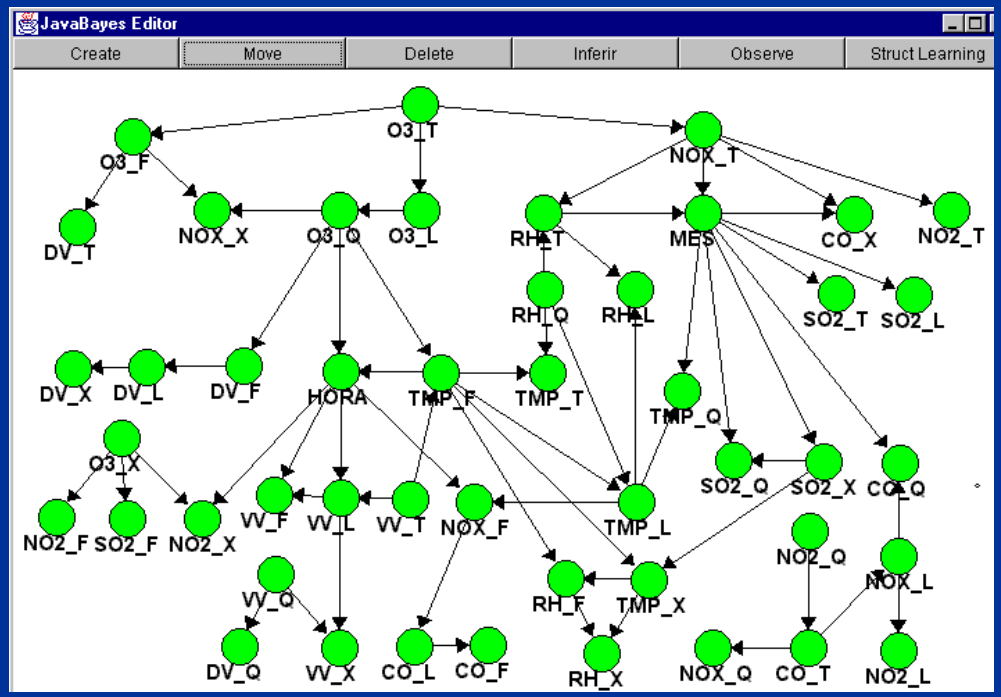
Parte I:
Modelos Gráficos Probabilistas



L. Enrique Sucar

INAOE

Información General

- Notas:

http://ccc.inaoep.mx/~esucar/Tut-Mixteca

- Correo:

esucar@inaoep.mx

Contenido

1. Introducción
2. Métodos básicos
3. Clasificadores bayesianos
4. Redes bayesianas
5. Aprendizaje de redes bayesianas
6. Campos de Markov

Referencias Generales

- J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan-Kaufman, 1988
- R. Neapolitan, Probabilistic Reasoning in Expert Systems, Wiley, 1990
- F. Jensen, Bayesian Networks and Decision Graphs, Springer, 2001
- A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge, 2009
- D. Koller & N. Friedman, Probabilistic Graphical Models: Principals and Techniques, MIT Press, 2009

I.1: Introducción

"Vivir es el arte de derivar conclusiones suficientes de premisas insuficientes"

[S. Butler]

Modelos Gráficos Probabilistas

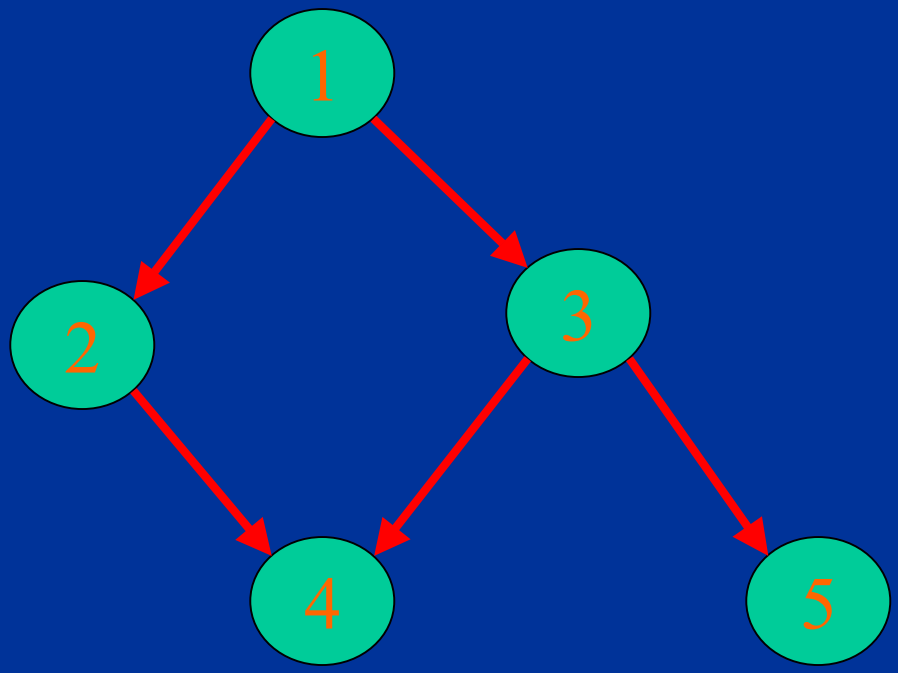
- Representaciones basadas en modelar la incertidumbre con variables probabilistas relacionadas mediante un modelo gráfico (dependencias)
- Muchas técnicas se pueden englobar dentro de este tipo de modelos

Modelos Gráficos Probabilistas

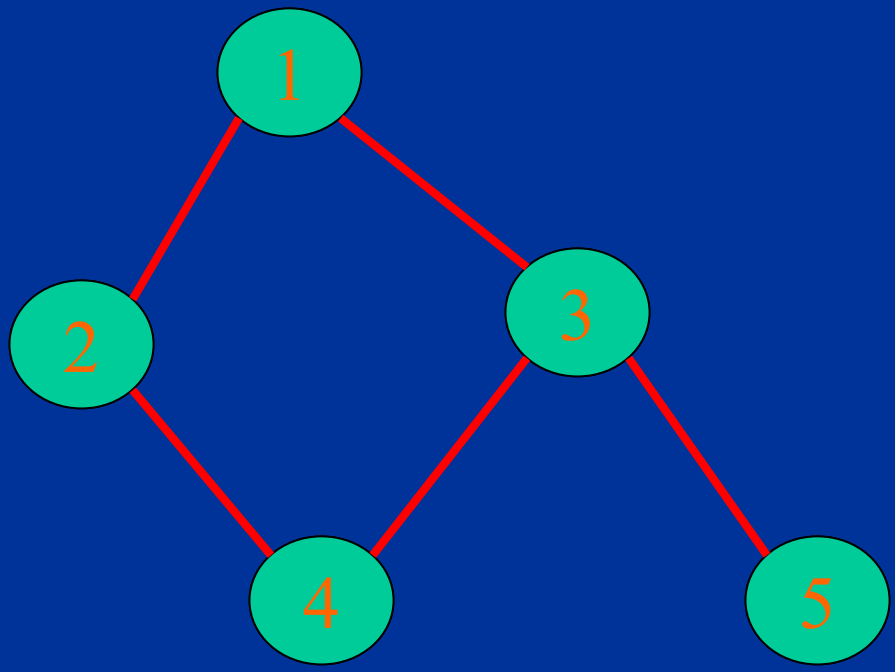
- Podemos clasificar los modelos gráficos probabilistas en 3 dimensiones principales:
  - Dirigidos vs. No-dirigidos
  - Estáticos vs. Dinámicos
  - Probabilista vs. Decisiones

Modelos Gráficos Probabilistas

- Dirigido

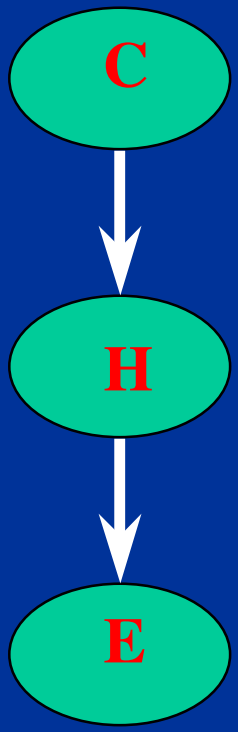


- No-dirigido

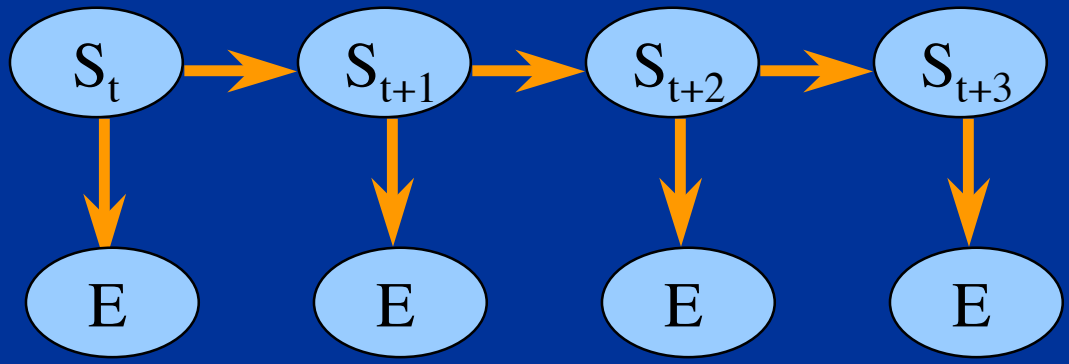


Modelos Gráficos Probabilistas

- Estático

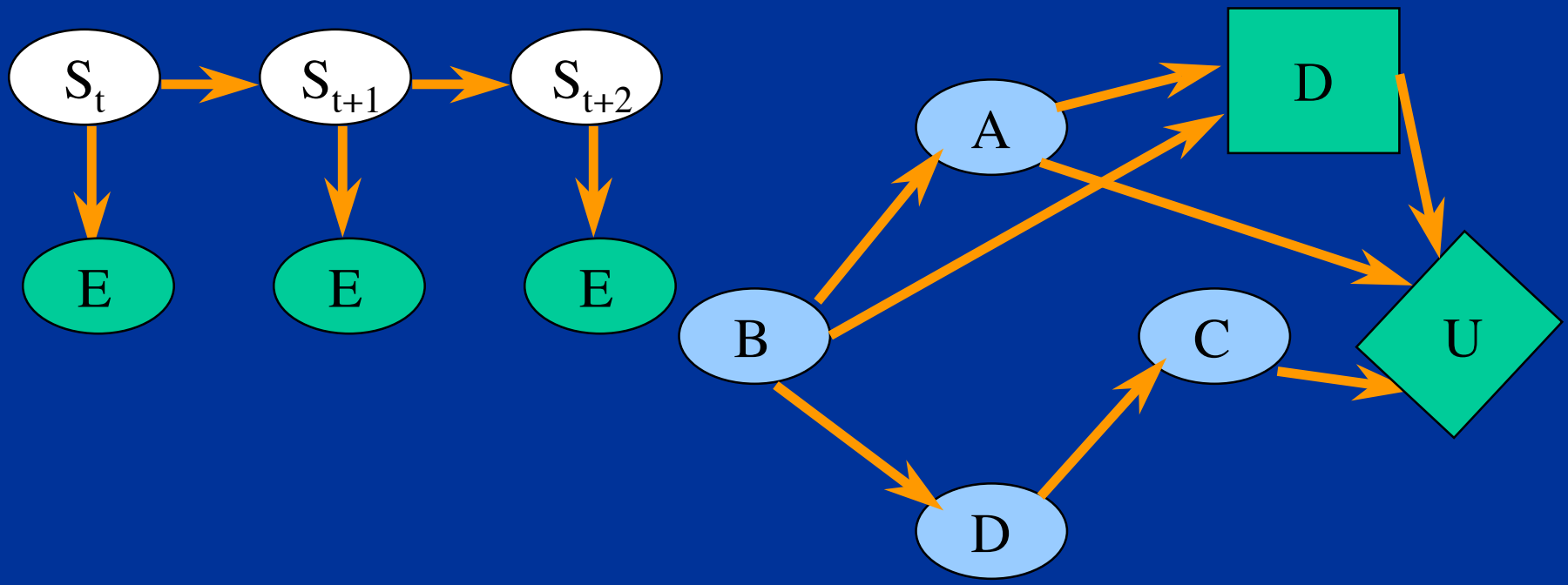


- Dinámico



Modelos Gráficos Probabilistas

- Probabilista
- Decisiones



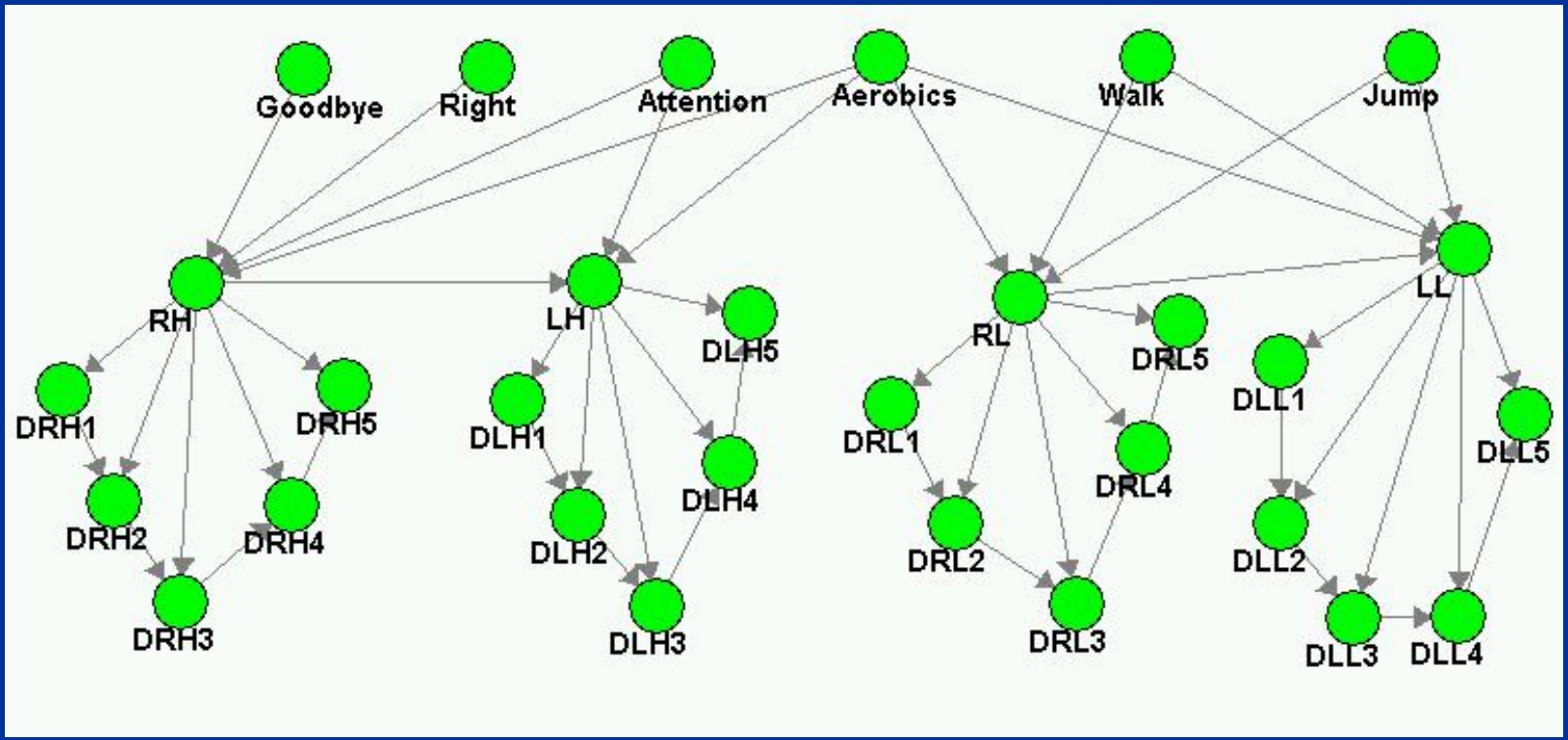
Modelos Gráficos Probabilistas

Modelo	D/N	S/D	P/D
Clasificador bayesiano	D/N	S	P
Cadenas ocultas de Markov	D	D	P
Campos / Redes de Markov	N	S	P
Redes bayesianas	D	S	P
Redes bayesianas dinámicas	D	D	P
Filtros de Kalman	D	D	P
Redes de Decisión	D	S	D
Procesos de decisión de Markov	D	D	D
POMDPs	D	D	D

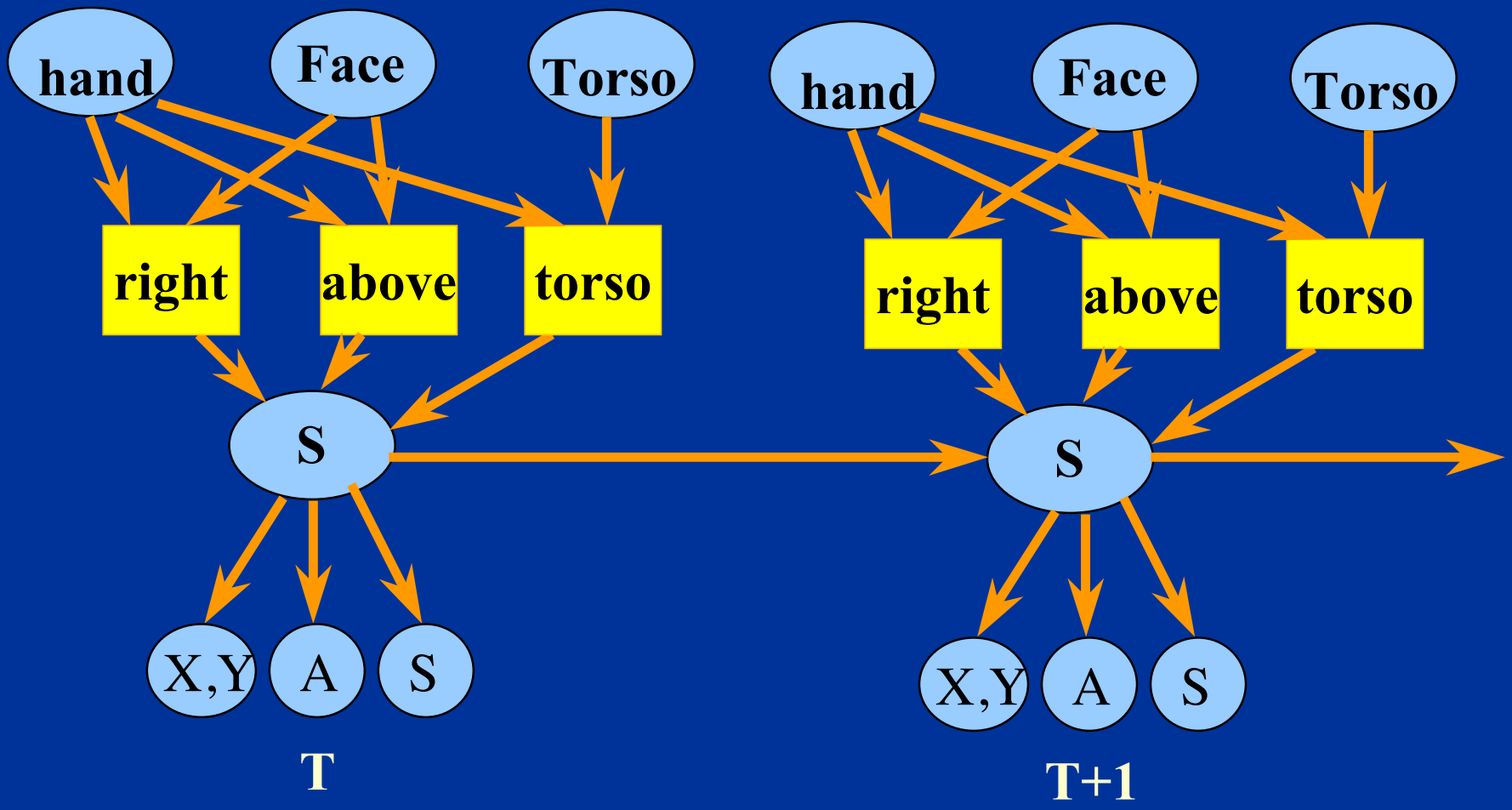
Aplicaciones

- Existen actualmente muchas aplicaciones de PGMs, como:
  - Reconocimiento de ademanes y actividades
  - Diagnóstico industrial
  - Modelado del estudiante
  - Navegación robótica
  - Validación de sensores
  - …

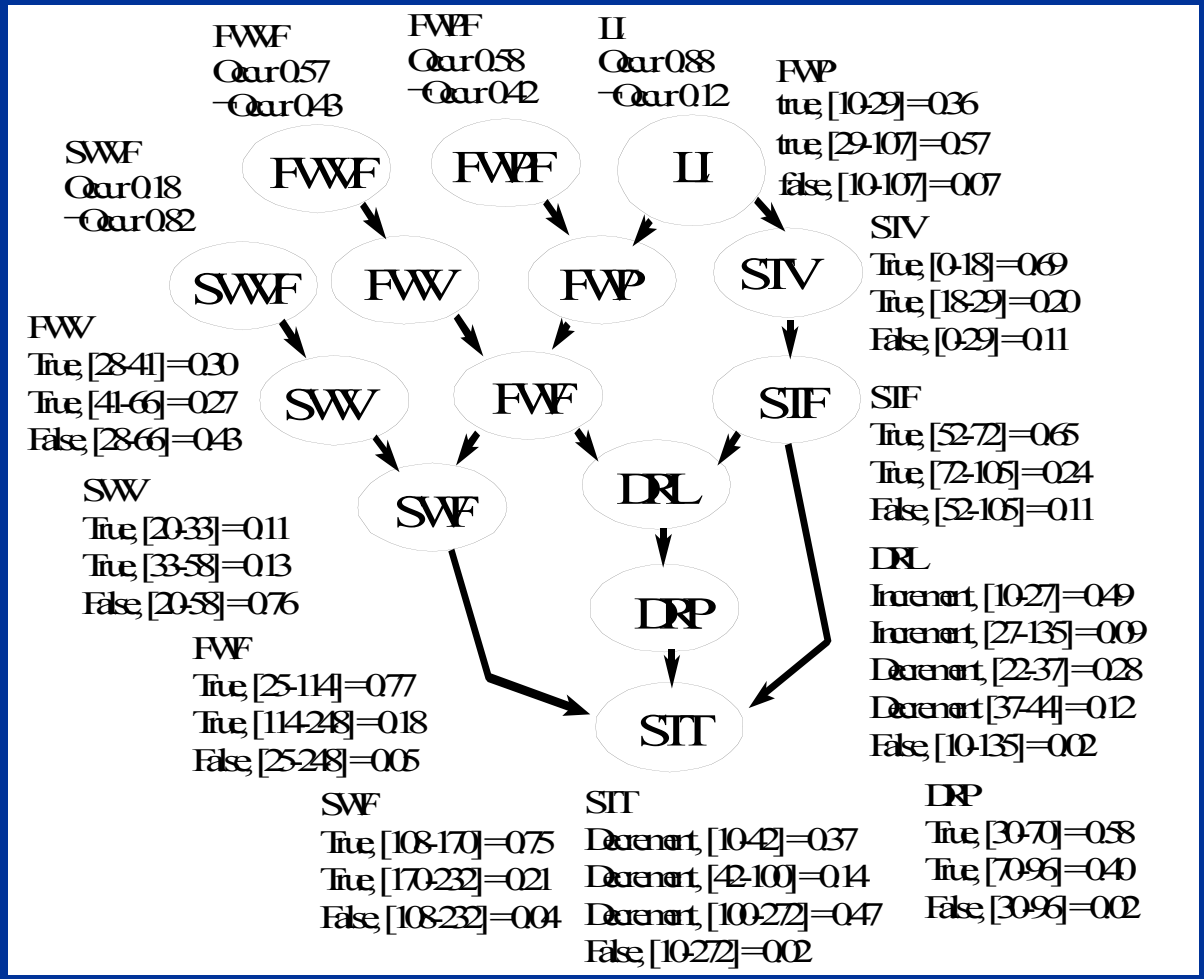
Reconocimiento de actividades humanas



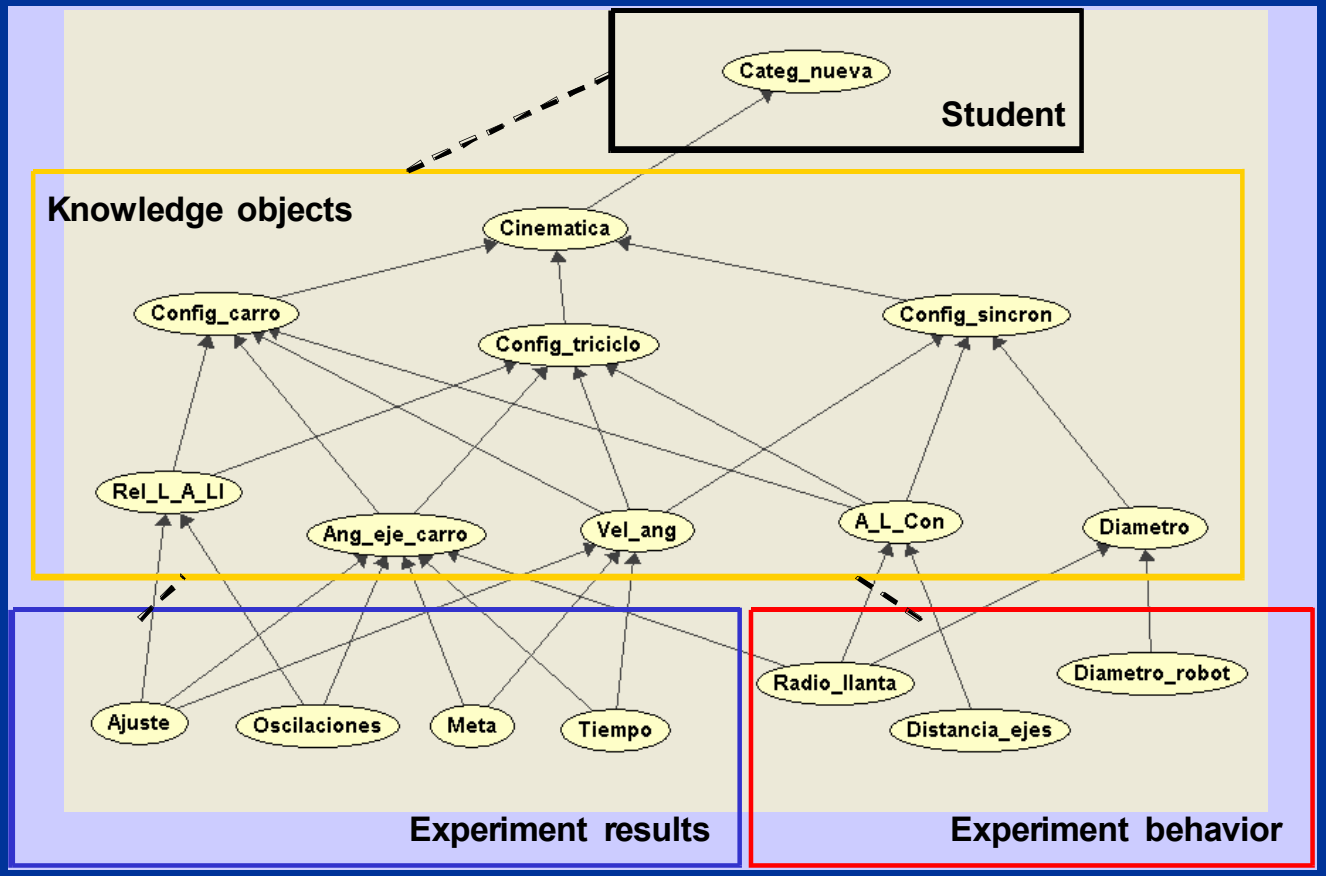
Reconocimiento de ademanes



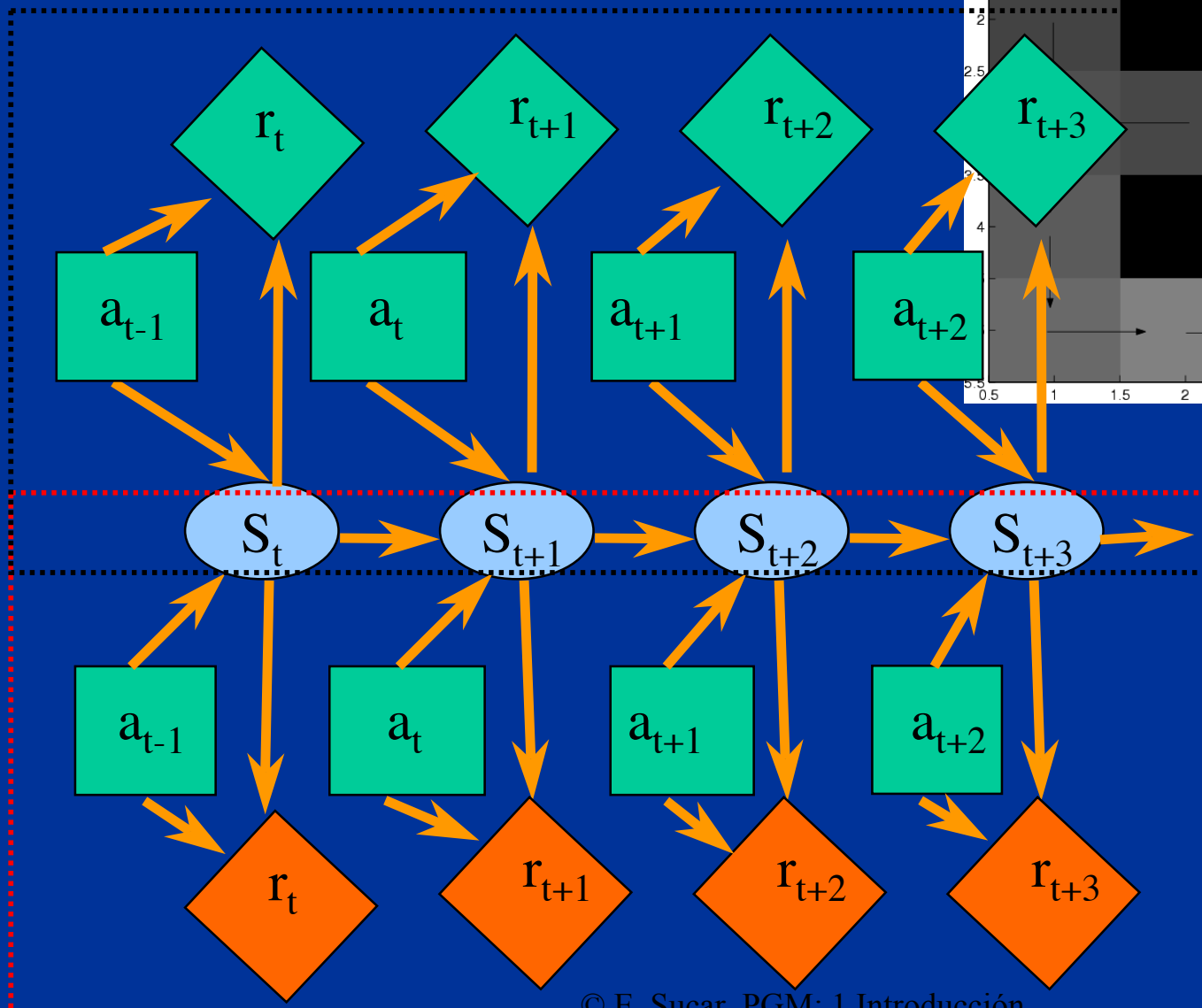
Diagnóstico industrial

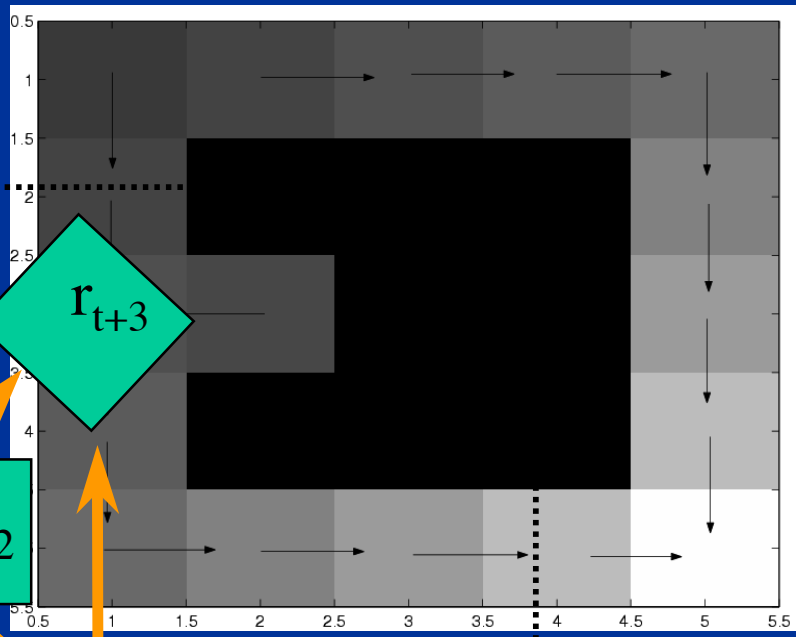


Modelado del estudiante

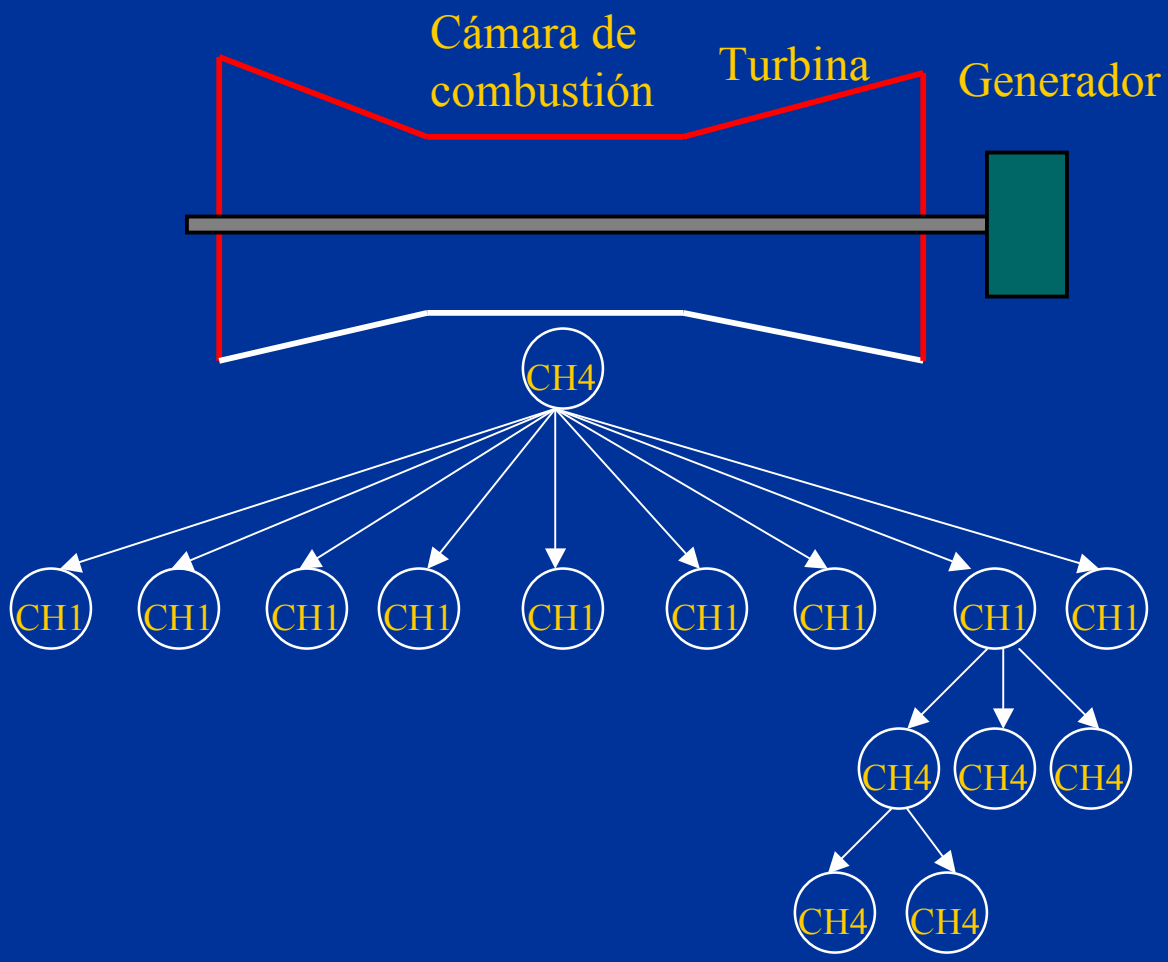


Navegación robótica





Validación de sensores



Referencias

- Ng & Abramson, “Uncertainty Management in Expert Systems”, IEEE Expert, Abril 1990.
- [Russell y Norvig] Cap. 14
- [Pearl] Cap.1

I.2: Métodos Básicos

"…las reglas mátemáticas de la probabilidad no son simplemente reglas para calcular frecuencias de variables aleatorias; son también las únicas reglas consistentes para realizar inferencia de cualquier tipo …"

[E. T. Jaynes 2003]

Formulación

- Muchos problemas se pueden formular como un conjunto de variables sobre las que tenemos cierta información y queremos obtener otra, por ejemplo:
  - Diagnóstico médico o industrial
  - Percepción (visión, voz, sensores)
  - Clasificación (bancos, empleadores, ...)
  - Modelado de estudiantes, usuarios, etc.

Ejemplo

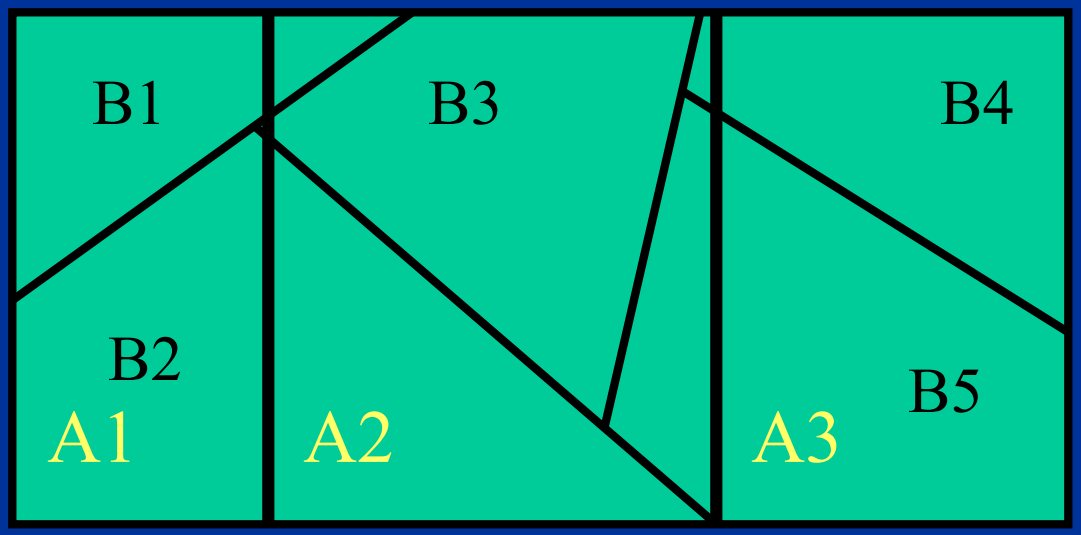
- Determinar si una persona es sujeta de crédito:
  - X1: otorgar crédito (si/no)
  - X2: ingreso anual (entero positivo)
  - X3: créditos anteriores (si/no)
  - X4: edad (entero positivo)
  - X5: ocupación (empleado, empresario, …)

Formulación

- Desde el punto de vista de probabilidad se puede ver como:
  - Un conjunto de variables aleatorias: X1, X2, X3, ...
  - Cada variable es generalmente una partición del espacio
  - Cada variable tiene una distribución de probabilidad (conocida o desconocida)

Variables y Particiones

- A = {A1, A2, A3}
- B = {B1, B2, B3, B4, B5}



Preguntas

- Dada cierta información (como valores de variables y probabilidades), se requiere contestar ciertas preguntas, como:
  - Probabilidad de que una variable tome cierto valor [marginal *a priori*]
  - Probabilidad de que una variable tome cierto valor dada información de otra(s) variable(s) [condicional o *a posteriori*]

Preguntas

- Valor de mayor probabilidad de una o más variables [abducción]
- Valor de mayor probabilidad de una o más variables dada información de otra(s) variable(s) [abducción parcial o explicación]
- Parámetros del modelo dados datos históricos de las variables [estimación o aprendizaje]

Enfoque básico (*fuerza bruta*)

- Dada la probabilidad conjunta de las variables, para todos los posibles valores de cada una (asumimos por ahora que son discretas):

$$P(X1, X2, X3, ..., Xn)$$

- podemos estimar todas las probabilidades requeridas

Inferencia

- Probabilidad marginal (cuál es la probabilidad de las diferentes ocupaciones):

$$p(X) = \Sigma_{Y, Z}\, p(X,Y, Z)$$

- Probabilidad condicional (cuál es la probabilidad de otorgar el crédito dado cierto nivel de ingreso) :

$$p(X \mid Y) = p(X,Y) / p(Y)$$

- Donde:

$$p(X,Y) = \Sigma_{Z}\, p(X,Y, Z)$$

Abducción

- Valor más probable (qué tipo de ocupación es el más común):

$$\text{Arg}_X \, [\max p(X) = \max \Sigma_{Y,Z} \, p(X,Y,Z)]$$

- Valor condicional más probable (debo o no otorgar el crédito):

$$\text{Arg}_X \, [\max p(X \mid y1) = \max p(X,y1) / p(y1)]$$

- Valor conjunto más probable (que combinación de ocupación y edad es la más probable):

$$\text{Arg}_{X,Y} \, [\max p(X,Y) = \max \Sigma_Z \, p(X,Y,Z)]$$

Ejemplo

- Problema de decidir cuando jugar golf?
- Variables
  - Ambiente
  - Temperatura
  - Viento
  - Humedad
  - Jugar

Ejemplo

- Consideremos inicialmente dos variables: ambiente (S,N,Ll) y temperatura (A,M,B)
- Dada la tabla de P conjunta, encontrar:
  - Probabilidad de ambiente, temperatura
  - Probabilidad de ambiente conocida la temperatura (y viceversa)
  - Combinación de A y T más probable
  - Ambiente más probable dada la temperatura (y viceversa)

Ejemplo

Limitaciones

- El tamaño de la tabla y el número de operaciones crece exponencialmente con el número de variables
- La “tabla” conjunta nos dice poco sobre el fenómeno que estamos analizando
- Puede ser difícil estimar las probabilidades requeridas (por expertos o a partir datos)

Estimación de Parámetros

- Dados un conjunto de valores de las variables (registros), se busca estimar las probabilidades conjuntas requeridas
- Considerando datos completos:
  - Las probabilidades se pueden *estimar* contando el número de casos de cada valor

$$P(X_i, Y_j) \sim N_{i,j} / N$$

  - Esto corresponde al *estimador de máxima verosimilitud* cuando no hay valores faltantes

Ejemplo

- Dados datos sobre lo que “jugadores” han hecho en situaciones pasadas, podemos estimar la probabilidad conjunta
- Consideremos el caso de 2 variables (ambiente y temperatura) y 14 registros de datos

Ejemplos

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	no	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvia	media	alta	si	N

Ejemplo

Limitaciones

- Se requiere una gran cantidad de datos para estimaciones confiables
- Se complica si hay datos faltantes
- Puede ser mejor estimar probabilidades marginales o condicionales (menos datos, más fácil para el experto)
- También puede ser complejo el tener *demasiados* datos (minería de datos)

Alternativas

- El problema de complejidad computacional utilizando el enfoque básico (tanto en espacio para representar el modelo, como en tiempo para el cálculo de probabilidades), nos lleva a buscar alternativas
- Los modelos gráficos probabilistas proveen esta alternativa, mediante representaciones muchos más compactas (y entendibles) y técnicas eficientes para el cálculo de las probabilidades

I.3: Clasificadores bayesianos

Clasificación

- El concepto de clasificación tiene dos significados:
  - No supervisada: dado un conjunto de datos, establecer clases o agrupaciones (*clusters*)
  - Supervisada: dadas ciertas clases, encontrar una regla para clasificar una nueva observación dentro de las clases existentes

Clasificación

- El problema de clasificación (supervisada) consiste en obtener el valor más probable de una variable (hipótesis) dados los valores de otras variables (evidencia, atributos)

$$\text{Arg}_H \left[\text{Max } P(H \mid E_1, E_2, \ldots E_N) \right]$$
$$\text{Arg}_H \left[\text{Max } P(H \mid \mathbf{E}) \right]$$

$$\mathbf{E} = \{E_1, E_2, \ldots E_N\}$$

Tipos de Clasificadores

- Métodos estadísticos clásicos
  - Clasificador bayesiano simple (*naive Bayes*)
  - Descriminadores lineales
- Modelos de dependencias
  - Redes bayesianas
- Aprendizaje simbólico
  - Árboles de decisión, reglas, …
- Redes neuronales, SVM, …

Clasificación

- Consideraciones para un clasificador:
  - Exactitud – proporción de clasificaciones correctas
  - Rapidez – tiempo que toma hacer la clasificación
  - Claridad – que tan comprensible es para los humanos
  - Tiempo de aprendizaje – tiempo para obtener o ajustar el clasificador a partir de datos

Regla de Bayes

- La probabilidad posterior se puede obtener en base a la regla de Bayes:

$$P(H \mid \mathbf{E}) = P(H)\, P(\mathbf{E} \mid H) / P(\mathbf{E})$$

$$P(H \mid \mathbf{E}) = P(H)\, P(\mathbf{E} \mid H) / \Sigma_i P(\mathbf{E} \mid H_i)\, P(H_i)$$

- Normalmente no se requiere saber el valor de probabilidad, solamente el valor más probable de H

Regla de Bayes

- Para el caso de 2 clases H:{0, 1}, la regla de decisión de Bayes es:

$$H^*(E) = \begin{cases} 1 \text{ si } P(H=1 \mid \mathbf{E}) > 1/2 \\ 0, \text{ de otra forma} \end{cases}$$

- Se puede demostrar que la regla de Bayes es óptima

Valores Equivalentes

- Se puede utilizar cualquier función monotónica para la clasificación:

$$\text{Arg}_H \left[\text{Max } P(H \mid \mathbf{E}) \right]$$

$$\text{Arg}_H \left[\text{Max } P(H)\, P(\mathbf{E} \mid H) / P(\mathbf{E}) \right]$$

$$\text{Arg}_H \left[\text{Max } P(H)\, P(\mathbf{E} \mid H) \right]$$

$$\text{Arg}_H \left[\text{Max } \log \{P(H)\, P(\mathbf{E} \mid H)\} \right]$$

$$\text{Arg}_H \left[\text{Max } (\log P(H) + \log P(\mathbf{E} \mid H)) \right]$$

Clasificador bayesiano simple

- Estimar la probabilidad: $P(\mathbf{E} \mid H)$ es complejo, pero se simplifica si se considera que los atributos son independientes dada la hipotesis:

$$P(E_1, E_2, \ldots E_N \mid H) = P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)$$

- Por lo que la probabilidad de la hipótesis dada la evidencia puede estimarse como:

$$P(H \mid E_1, E_2, \ldots E_N) = \frac{P(H)\, P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)}{P(\mathbf{E})}$$

- Esto se conoce como el clasificador bayesiano simple

Clasificador bayesiano simple

- Como veíamos, no es necesario calcular el denominador:

$$P(H \mid E_1, E_2, \ldots E_N) \sim$$

$$P(H)\, P(E_1 \mid H)\, P(E_2 \mid H) \ldots P(E_N \mid H)$$

- P(H) se conoce como la *probabilidad a priori*, $P(E_i \mid H)$ es la probabilidad de los atributos dada la hipótesis (*verosimilitud*), y $P(H \mid E_1, E_2, \ldots E_N)$ es la *probabilidad posterior*

Ejemplo

- Para el caso del golf, cuál es la acción más probable (jugar / no-jugar) dado el ambiente y la temperatura?

Hoja de cálculo de Microsoft Excel

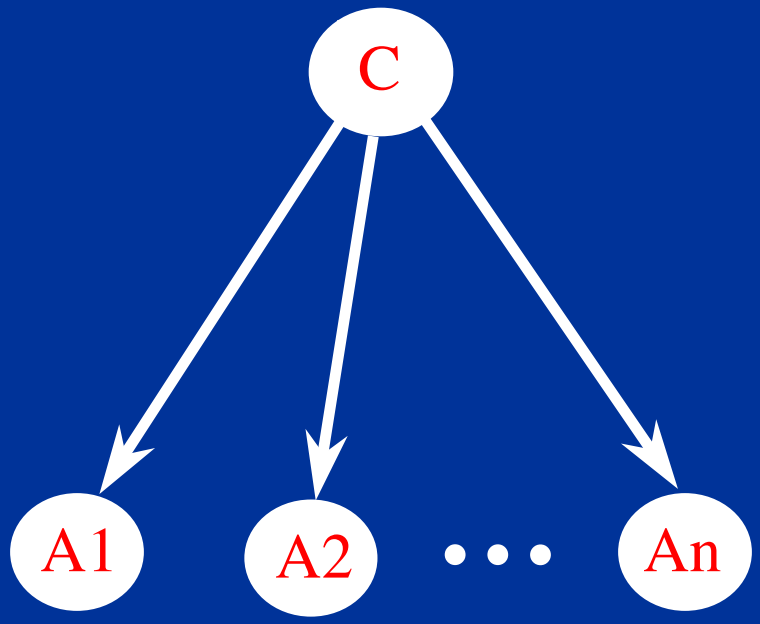
Ventajas

- Bajo tiempo de clasificación
- Bajo tiempo de aprendizaje
- Bajos requerimientos de memoria
- “Sencillez”
- Buenos resultados en muchos dominios

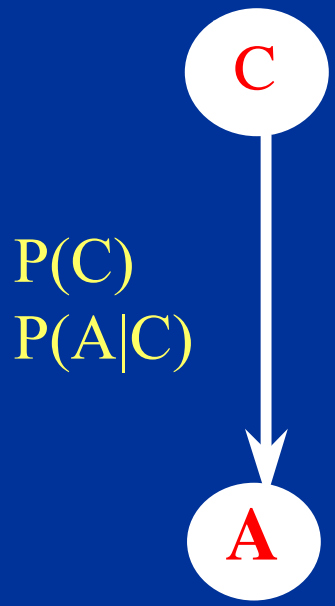
Limitaciones

- En muchas ocasiones la suposición de independencia condicional no es válida
- Para variables continuas, existe el problema de discretización
- Alternativas – dependencias:
  - Estructuras que consideran dependencias
  - Mejora estructural del clasificador
- Alternativas – variables continuas:
  - Descriminador lineal (variables gaussianas)
  - Técnicas de discretización

CBS – modelo gráfico



Enfoques para clasificación

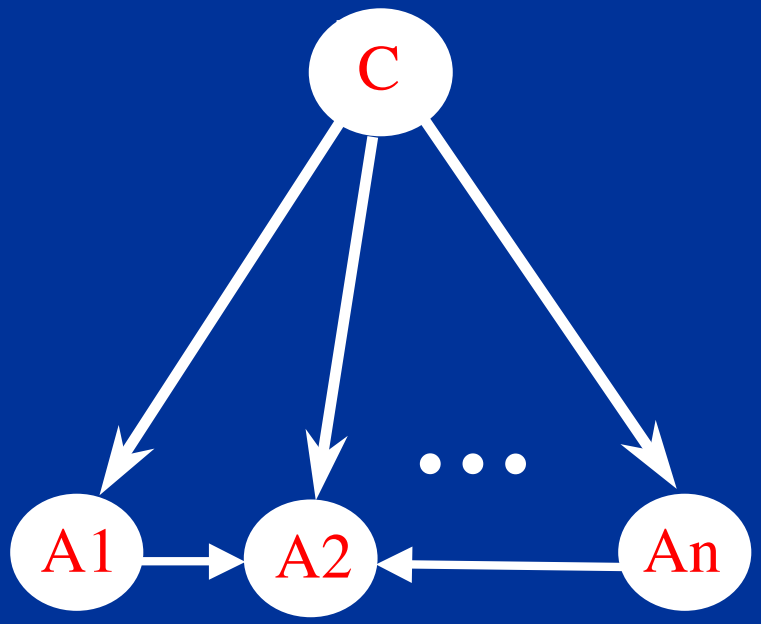


Generativo

Descriminativo

Extensiones

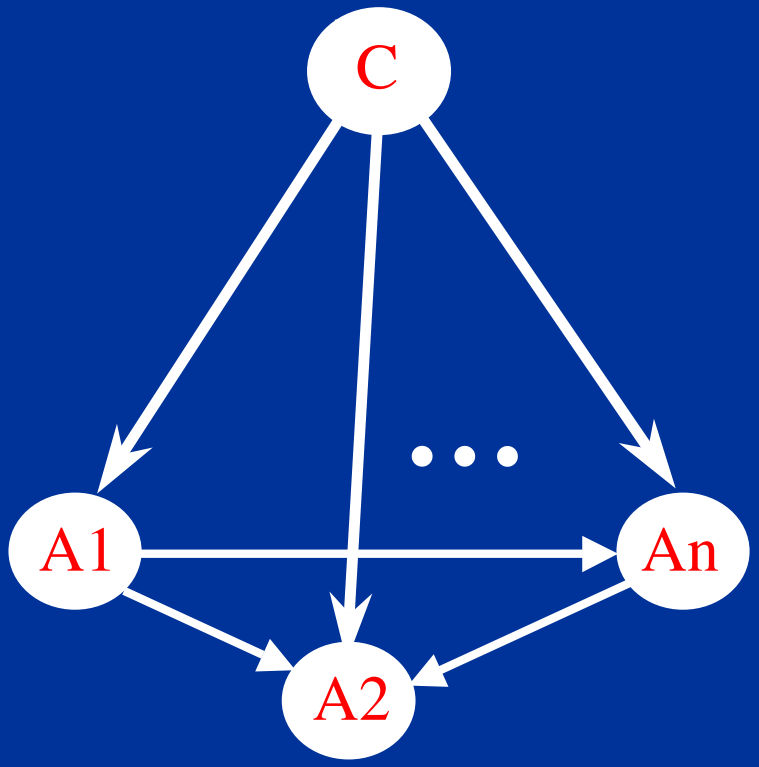
- TAN



Se incorpora algunas dependencias entre atributos mediante la construcción de un "árbol" entre ellos (más adelante veremos como se aprende el árbol)

Extensiones

- BAN



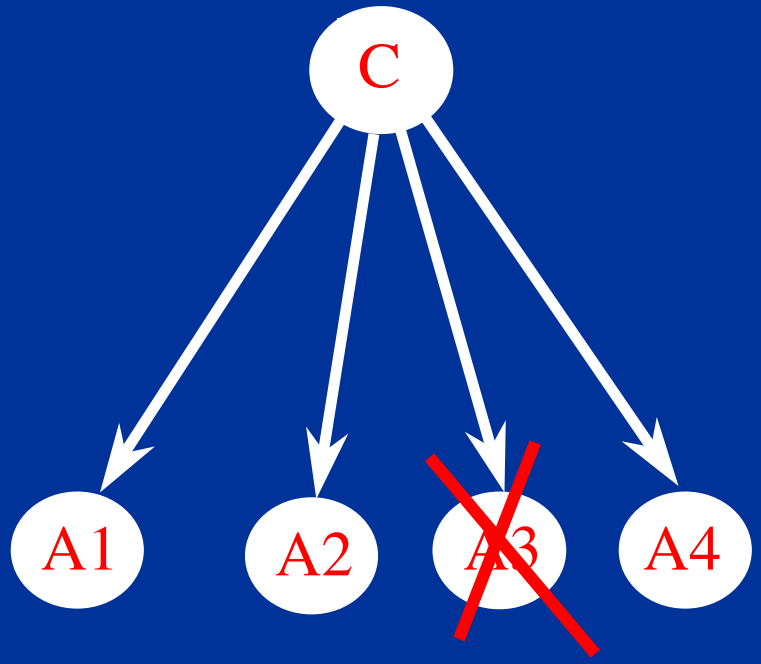
Se incorpora una “red” para modelar las dependencias entre atributos (también más adelante veremos como aprender una red).

Mejora estructural

- Otra alternativa para mejorar el CBS es partir de una estructura “simple” y modificarla mediante:
  - Eliminación de atributos irrelevantes (selección de atributos)
  - Verificación de las relaciones de independencia entre atributos y alterando la estructura:
    - Eliminar nodos
    - Combinar nodos
    - Insertar nodos

Eliminación de atributos

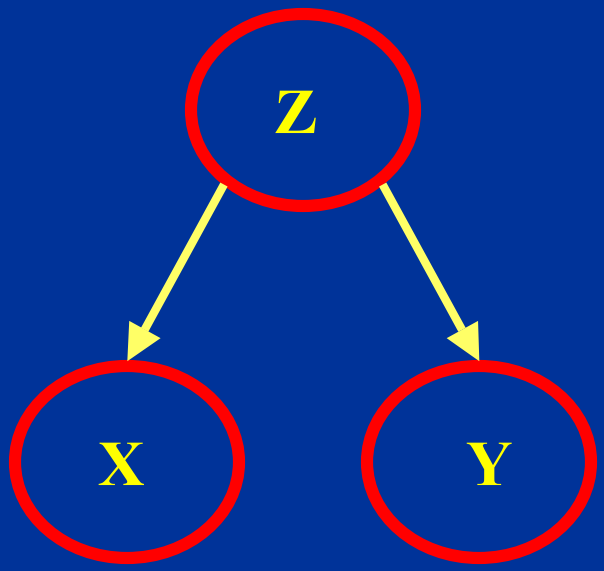
- Medir la "dependencia" entre la clase y atributos (por ejemplo con la información mutua), y eliminar aquellos con "poca" aportación

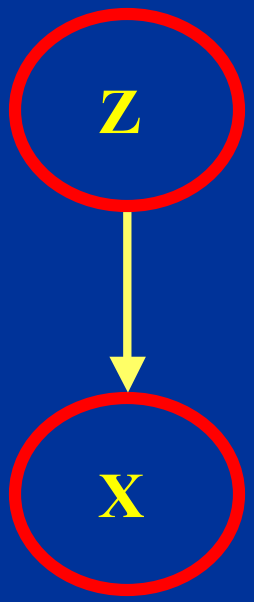


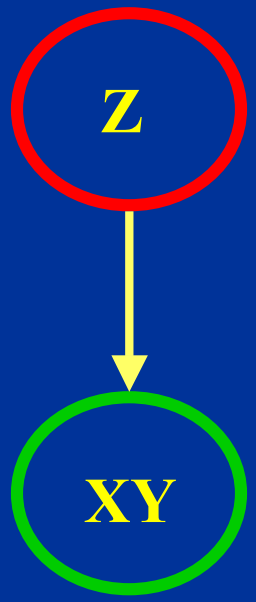
Mejora estructural

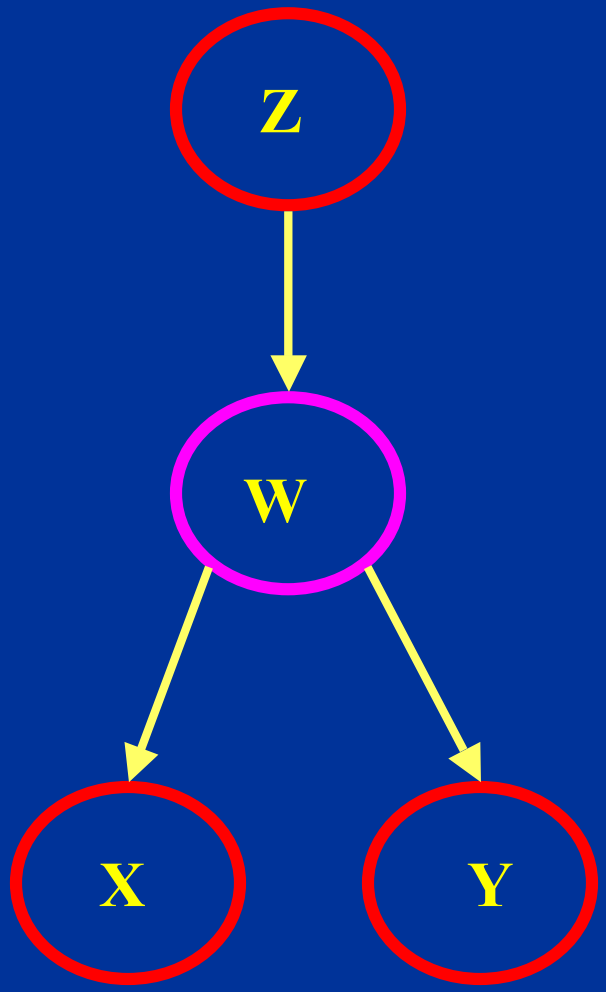
- Medir la dependencia entre pares de atributos dada la clase (por ejemplo mediante la información mutua condicional), alterar la estructura si hay 2 dependientes:
  1. Eliminación: quitar uno de los dos (redundantes)
  2. Unión: juntar los 2 atributos en uno, combinando sus valores
  3. Inserción: insertar un atributo "virtual" entre la clase y los dos atributos que los haga independientes.

Mejora Estructural









Atributos redundantes

- Prueba de dependencia entre cada atributo y la clase
- Información mutua:

$$MI(C, Ai) = \sum P(C, Ai) \log [P(C, Ai) / P(C) P(Ai)]$$

- Eliminar atributos que no provean información a la clase

Atributos dependientes

- Prueba de independencia de cada atributo dada la clase
- Información mutua condicional

$$CMI(A_i, A_j \mid C) = \sum P(A_i,A_j|C) \log [P(A_i,A_j|C) / P(A_i|C)\, P(A_j|C)]$$

- Eliminar, unir o (insertar) atributos

Discretización

- Si los atributos no siguen una distribución gaussiana, la alternativa es convertirlos a discretos agrupando los valores en un conjunto de rangos o intervalos
- Dos tipos de técnicas de discretización:
  - No supervisada: no considera la clase
  - Supervisada: en base a la clase

Discretización no supervisada

- Intervalos iguales
- Intervalos con los mismos datos
- En base al histograma

Discretización supervisada

- Considerando los posibles "cortes" entre clases:
  - Probar clasificador (con datos diferentes)
  - Utilizar medidas de información (p. ej., reducir la entropía)
- Problema de complejidad computacional

Costo de mala clasificación

- En realidad, no sólo debemos considerar la clase más probable si no también el costo de una mala clasificación
  - Si el costo es igual para todas las clases, entonces es equivalente a seleccionar la de mayor probabilidad
  - Si el costo es diferente, entonces se debe minimizar el costo esperado

Costo de mala clasificación

- El costo esperado (para dos clases, + y -) está dado por la siguiente ecuación:

$$CE = FN\ p(-)\ C(-|+) + FP\ p(+)\ C(+|-)$$

FN: razón de falsos negativos

FP: razón de falsos positivos

p: probabilidad de negativo o positivo

C(-|+): costo de clasificar un positivo como negativo

C(+|-): costo de clasificar un negativo como positivo

- Considerando esto y también la proporción de cada clase, existen técnicas más adecuadas para comparar clasificadores como la *curva ROC* y las *curvas de costo*

Referencias

- Clasificadores:
  - D. Michie, D.J. Spiegelhalter , C.C. Taylor, “Machine Learning, Neural and Statistical Classification”, Ellis Horwood, 1994
  - L. E. Sucar, D. F. Gillies, D. A. Gillies, "Objective Probabilities in Expert Systems", Artificial Intelligence Journal, Vol. 61 (1993) 187-208.
  - J. Cheng, R. Greiner, “Comparing Bayesian network classifiers”, UAI´99, 101-108.
  - M. Pazzani, “Searching for attribute dependencies in Bayesian classifiers”, Preliminary Papers of Intelligence and Statistics, 424-429.
  - M. Martínez, L.E. Sucar, “Learning an optimal naive Bayesian classifier”, ICPR, 2006

Referencias

- Evaluación:
  - C. Drummond, R. C. Holte, “Explicitly representing expected cost: an alternative to the ROC representation”.

I.4: Redes Bayesianas

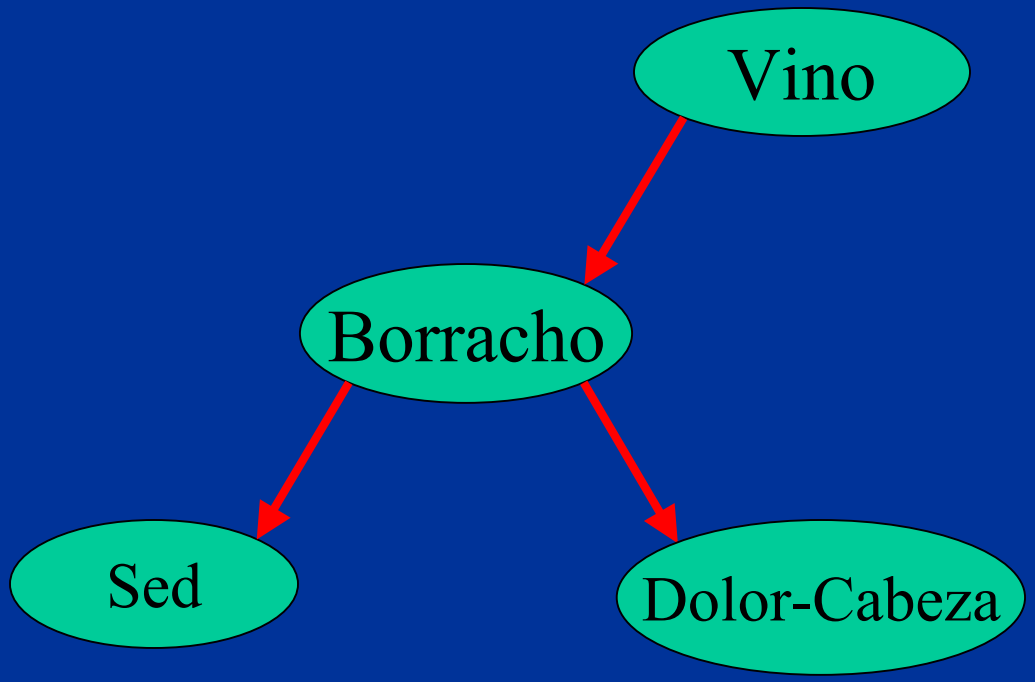
"La probabilidad no es realmente sobre números, es sobre la estructura del razonamiento"

[G. Shafer]

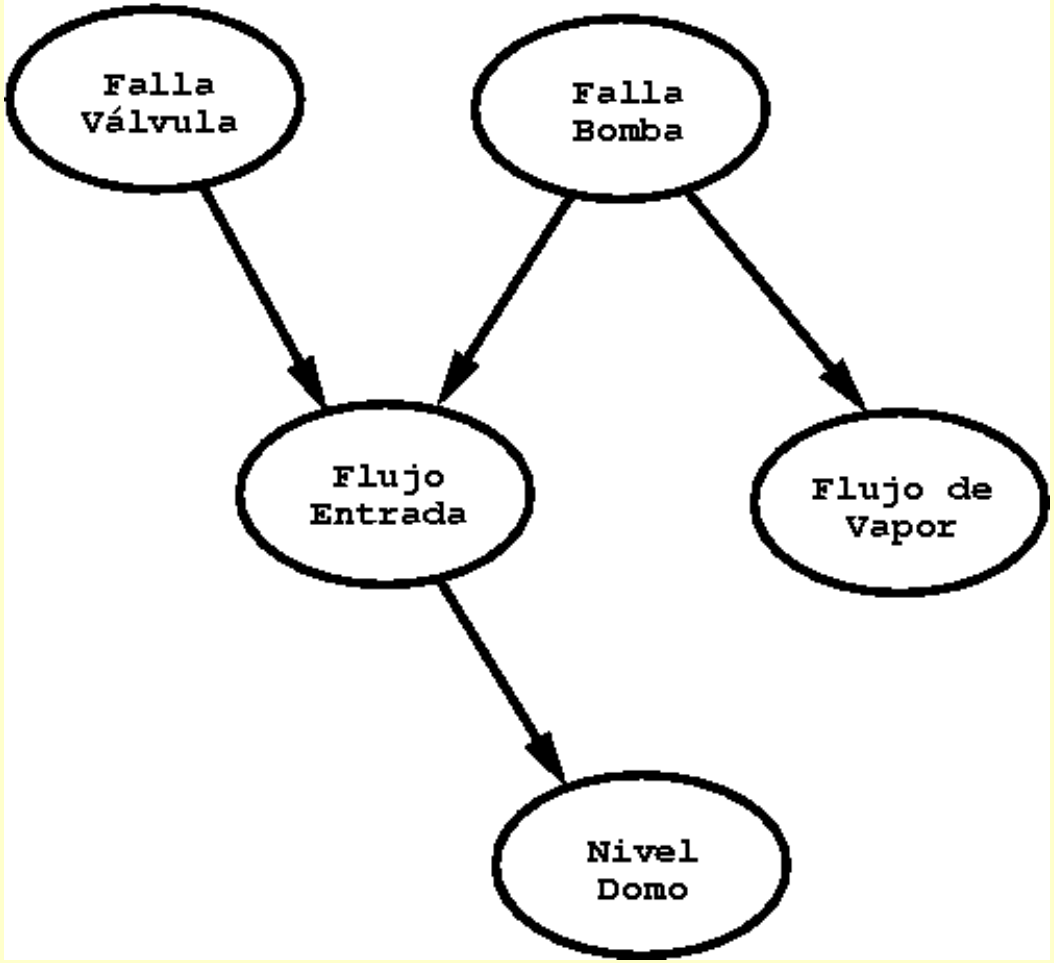
Representación

- Las redes bayesianas son una representación gráfica de dependencias para razonamiento probabilístico, en la cual los nodos y arcos representan:
  - Nodos: Variables proposicionales.
  - Arcos: Dependencia probabilística
- La variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste.

Ejemplo de una red bayesiana



Otro ejemplo



Estructura

- La topología o estructura de la red nos da información sobre las dependencias probabilísticas entre las variables.
- La red también representa las independencias condicionales de una variable (o conjunto de variables) dada otra variable(s).

Ejemplo

- Para el caso del domo:

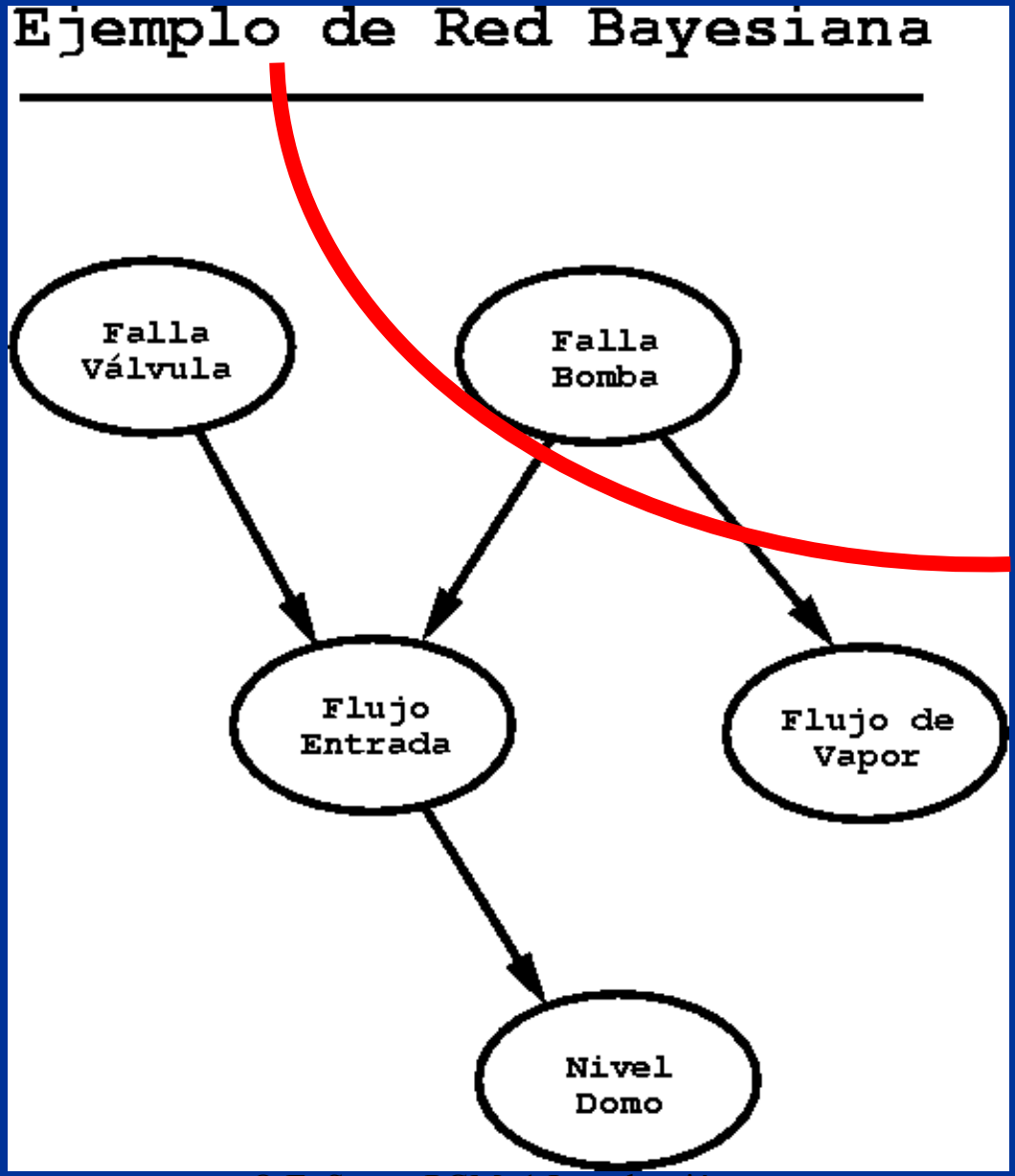
{Fva} es cond. indep. de {Fv, Fe, Nd} dado {Fb}

- Esto es:

$$P(Fva \mid Fv, Fe, Nd, Fb) = P(Fva \mid Fb)$$

- Esto se representa gráficamente por el nodo Fb separando al nodo Fva del resto de las variables.

Ejemplo de Red Bayesiana



Independencias condicionales

- En una RB todas la relaciones de independencia condicional representadas en el grafo corresponden a relaciones de independencia en la distribución de probabilidad.

- Dichas independencias simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades).

Representación Gráfica

- Una red bayesiana representa en forma gráfica las dependencias e independencias entre variables aleatorias, en particular las independencias condicionales
- Independencia en la distribución
  - $P(X \mid Y,Z) = P(X \mid Z)$
- Independencia en el grafo
  - X "separada" de Y por Z

Representación Gráfica

Notación:

- Independencia en la distribución
  - I(X,Z,Y)
- Independencia en el grafo
  - < X | Z | Y >

Separación "D"

- El conjunto de variables A es independiente del conjunto B dado el conjunto C, si no existe trayectoria entre A y B en que
  1. Todos los nodos convergentes están o tienen descendientes en C
  2. Todos los demás nodos están fuera de C

Separación “D”

- Tres casos básicos
  - Arcos divergentes
  - Arcos en secuencia
  - Arcos convergentes

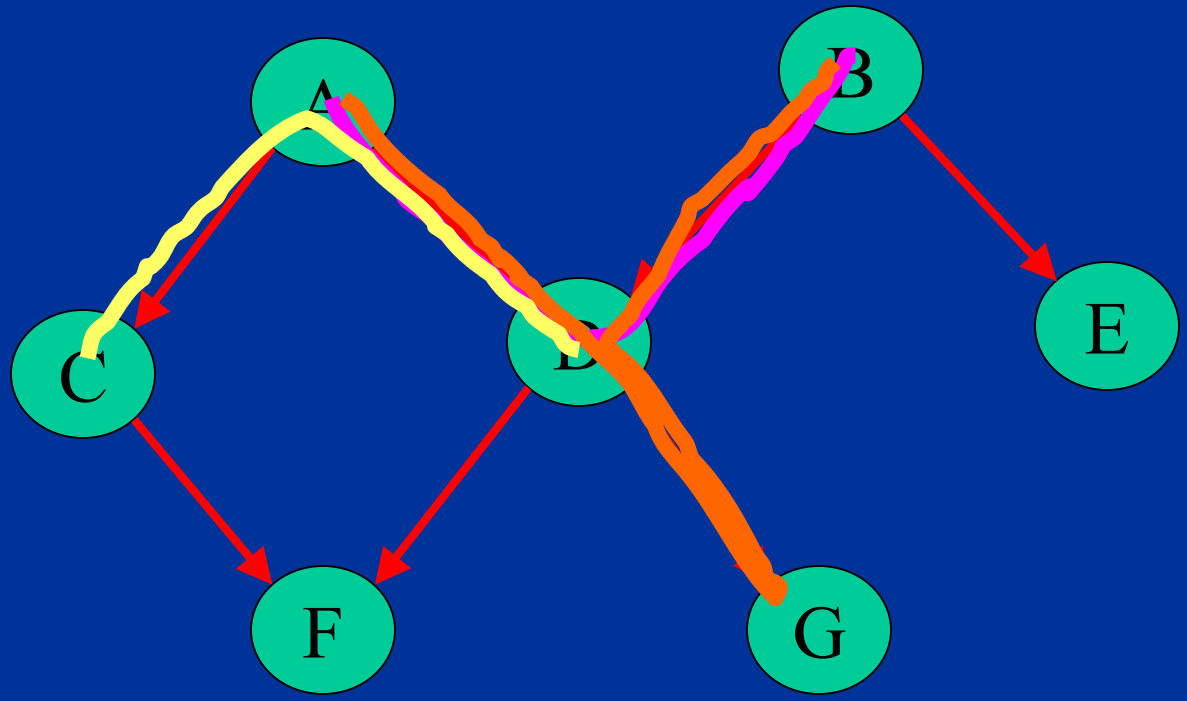
Separación “D” – casos básicos

- caso 1: Secuencia:

- caso 2: Divergentes:

- caso 3: Convergentes:

Ejemplos Separación-D



¿I(A,CD,F)?
¿I(A,CD,B)?
¿I(BD,A,C)?
¿I(A,G,B)?
¿I(A,D,G)?
¿I(C,BEG,D)?

Correspondencia Grafo-Modelo

- Dada una distribución de probabilidad o modelo (M) y una representación gráfica de dependencias o grafo (G) debe existir una correspondencia entre las independencias representados en ambos
- Tres tipos básicos - *mapas*

Correspondencia Grafo-Modelo

- Mapa-D: las variables independientes están separadas en el grafo
- Mapa-I: las variables separadas en el grafo son independientes
- Mapa perfecto: mapa-I & mapa-D

- No es siempre posible tener un mapa perfecto (hay distribuciones con relaciones de independencia que no se pueden representar como un GAD)

Correspondencia Grafo-Modelo

- Mapa-I mínimo: las variables separadas en el grafo son independientes y al quitar cualquier arco se destruye esta condición

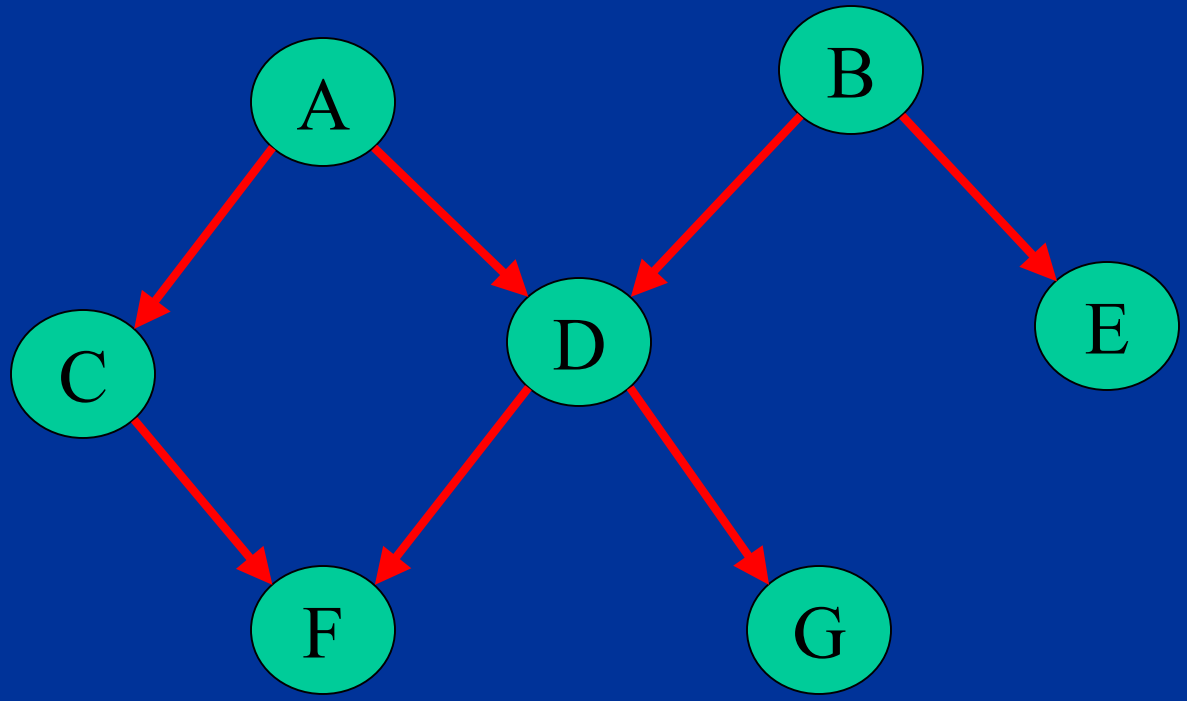
- Una red bayesiana es un grafo acíclico dirigido (GAD) que corresponde a un mapa-I mínimo de una distribución de probabilidad P

Especificación Estructural

- En una RB, cualquier nodo X es independiente de todos los nodos que no son sus descendientes dados sus nodos padres Pa(X) – "contorno de X"

- La estructura de una RB se especifica indicando el contorno (padres) de cada variable

Especificación Estructural

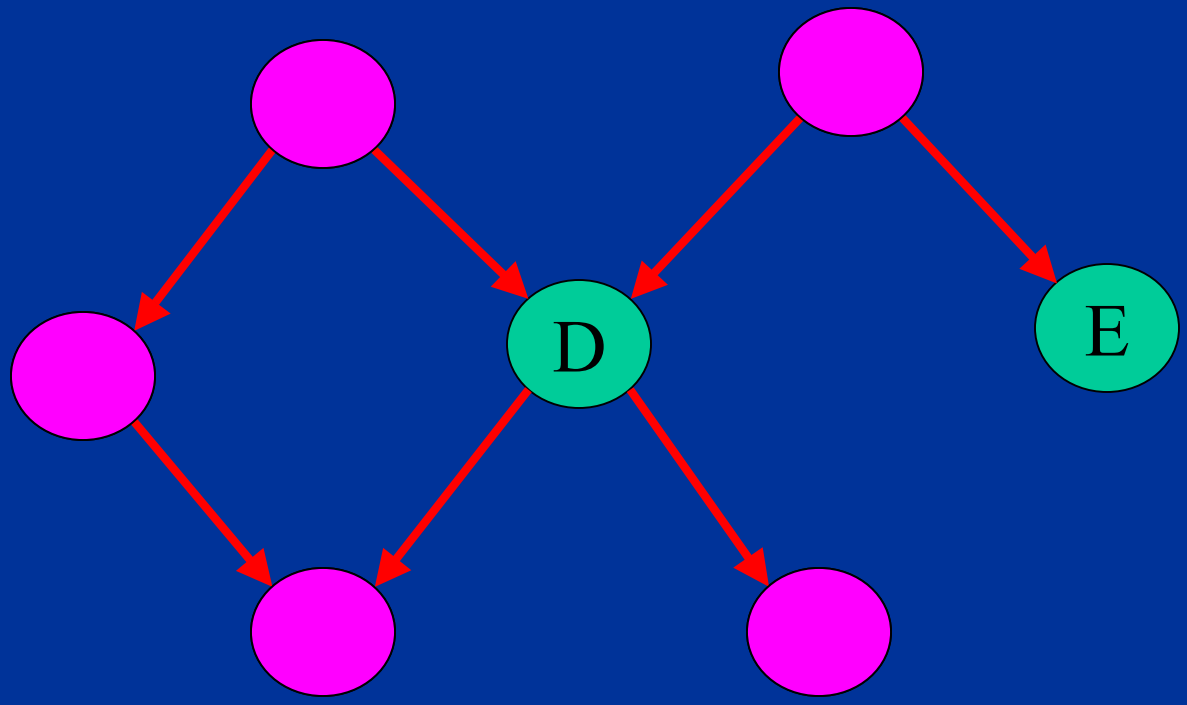


Pa(A) = 0
Pa(B) = 0
Pa(C) = A
Pa(D) = A, B
Pa(E) = B
Pa(F) = C, D
Pa(G) = D

Cobija de Markov

- La "cobija de Markov" de un nodo es el conjunto de nodos que lo hacen independiente del resto de la red
- Para una RB la cobija de Markov está formada por:
  - Nodos padre
  - Nodos hijo
  - Otros padres de los hijos

Cobija de Markov

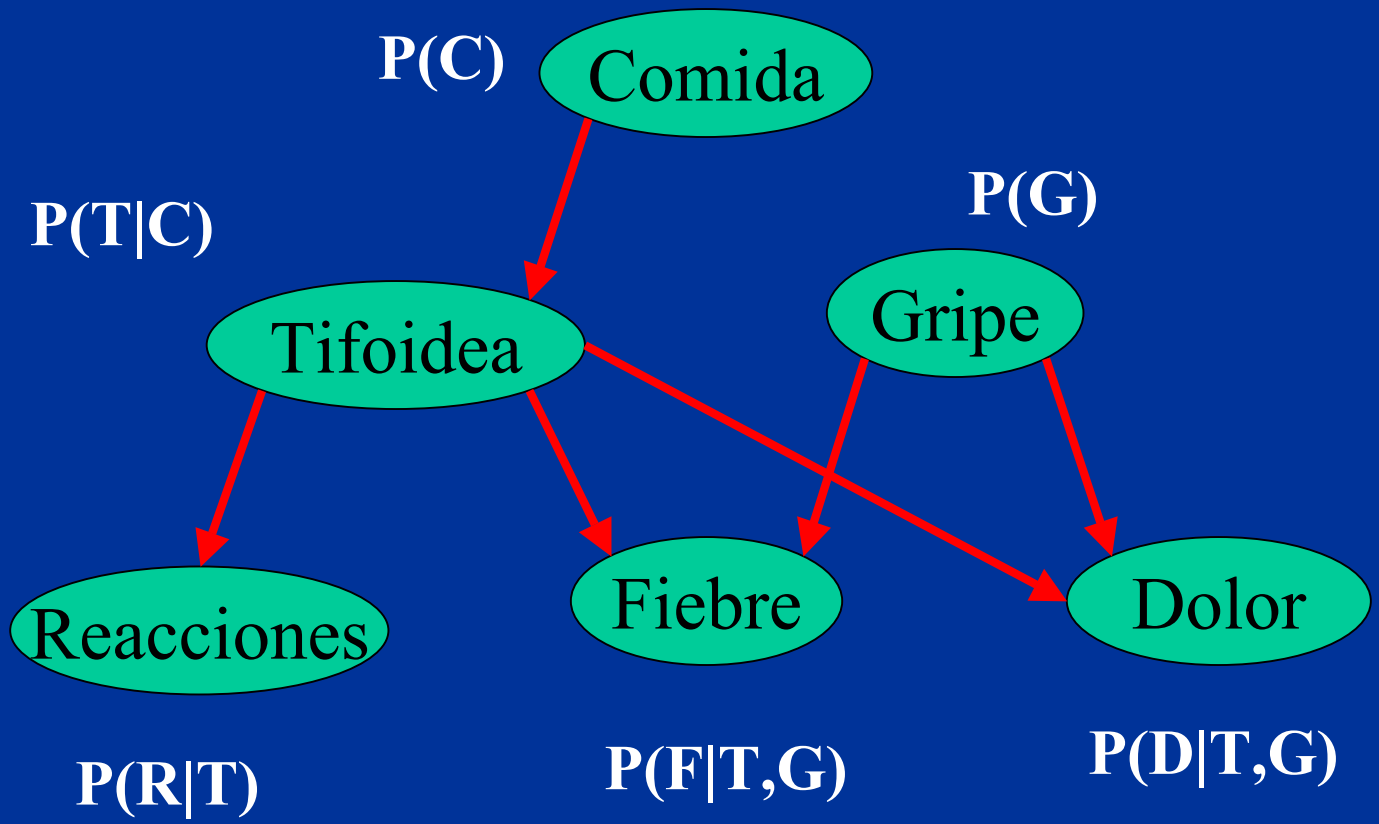


CM (D) ?

Parámetros

- Complementan la definición de una red bayesiana las probabilidades condicionales de cada variable dados sus padres.
  - **Nodos raíz**: vector de probabilidades marginales
  - **Otros nodos**: matriz de probabilidades condicionales dados sus padres

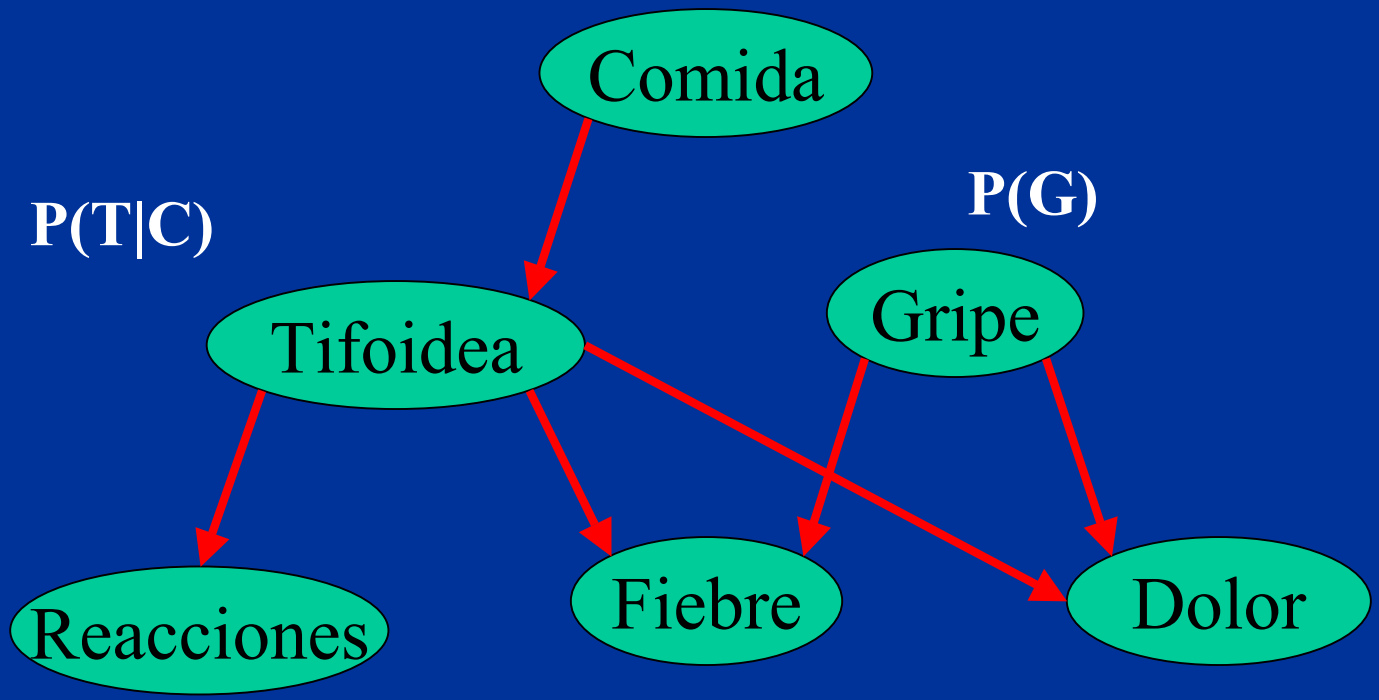
Ejemplo

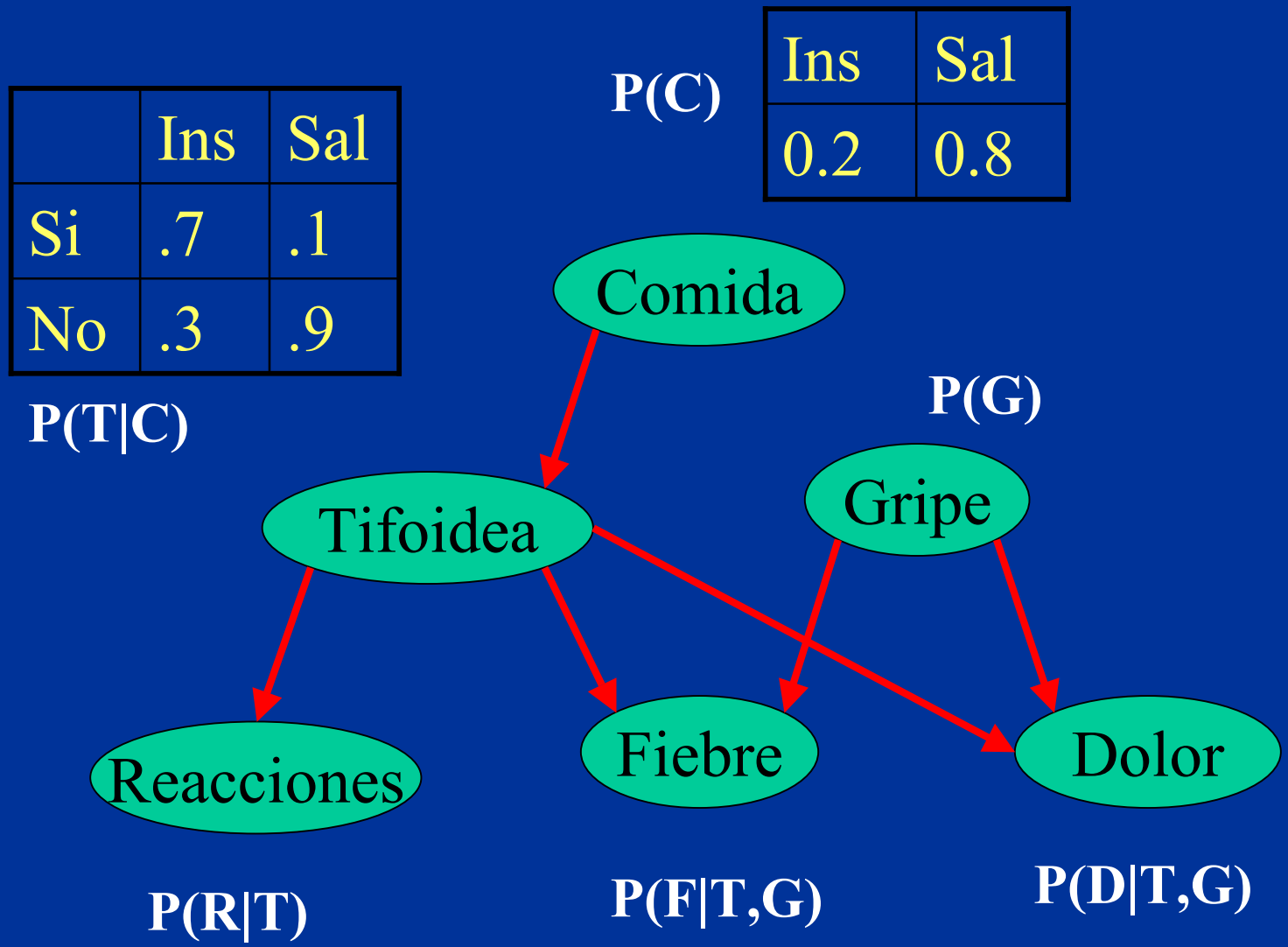

P(C)
Comida
P(T|C)
P(G)
Tifoidea
Gripe
Reacciones
Fiebre
Dolor
P(R|T)
P(F|T,G)
P(D|T,G)

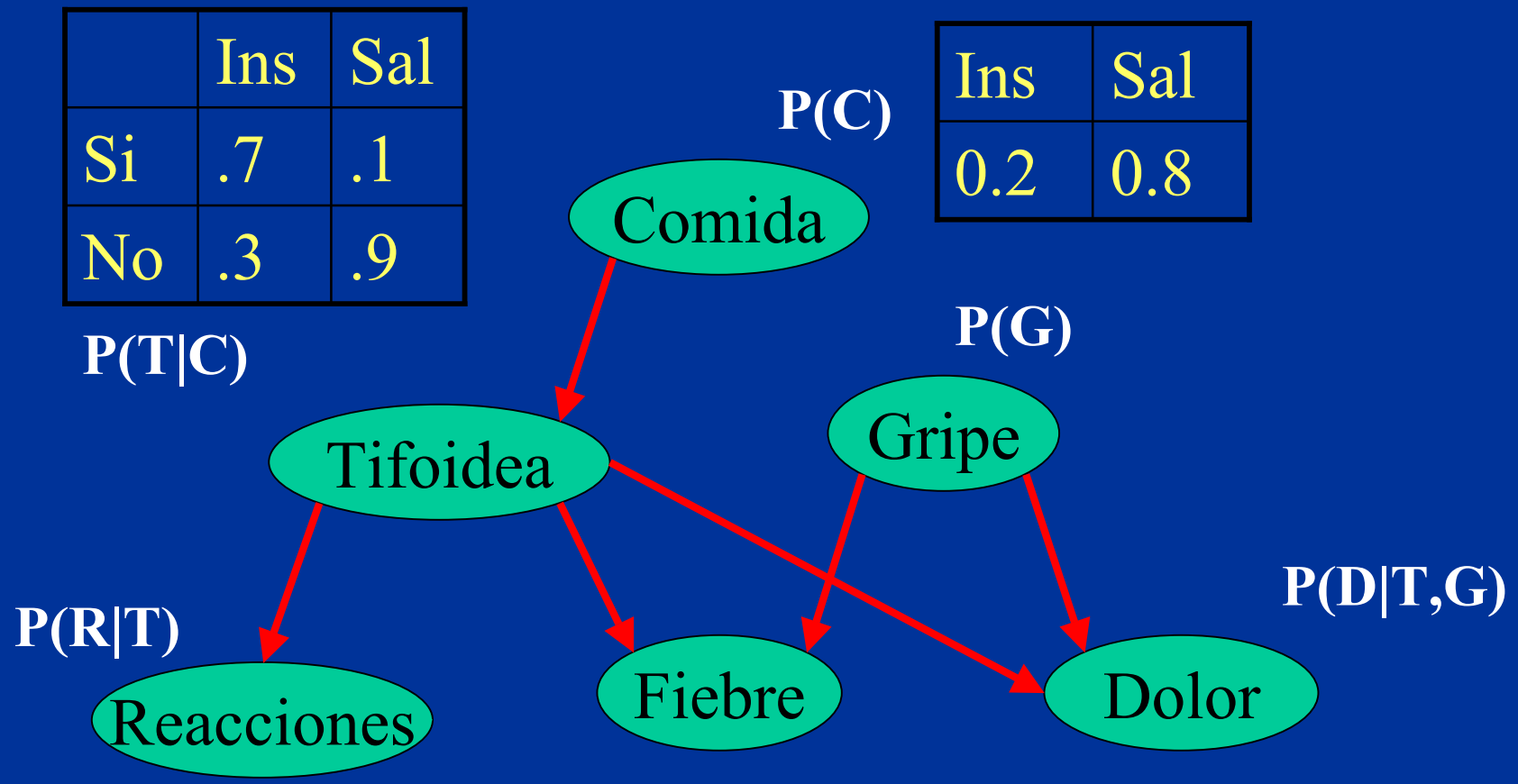
Ejemplo

P(C)

Ins	Sal
0.2	0.8




P(C)
Ins
Sal
0.2
0.8
Comida
P(T|C)
Si
.7
.1
No
.3
.9
P(G)
Tifoidea
Gripe
Reacciones
Fiebre
Dolor
P(R|T)
P(F|T,G)
P(D|T,G)



	Ins	Sal
Si	.7	.1
No	.3	.9

P(T|C)

P(C)

Ins	Sal
0.2	0.8

P(G)

P(R|T)

P(D|T,G)

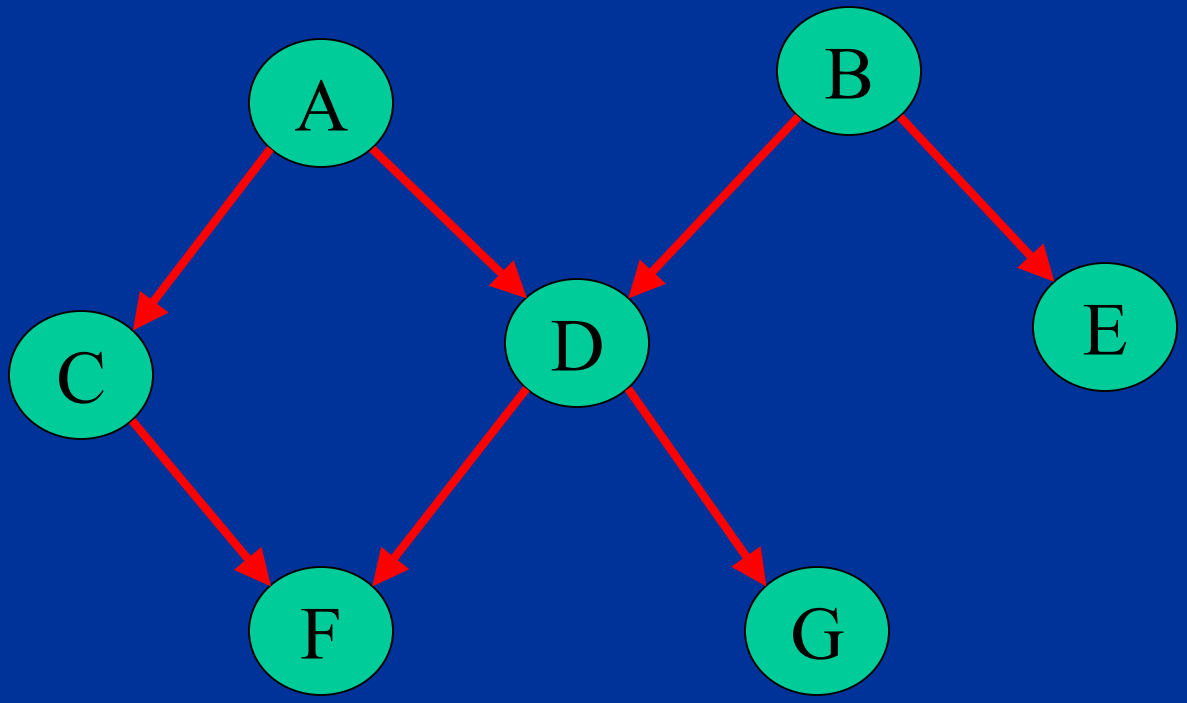
P(F|T,G)

	Si, Si	Si,No	No,Si	No,No
F	0.8	0.6	0.5	0.1
~F	0.2	0.4	0.5	0.9

Especificación Paramétrica

- Dado que los contornos (padres) de cada nodo especifican la estructura, mediante las probabilidades condicionales de dichos nodos podemos especificar también las probabilidades requeridas
- Aplicando la regla de la cadena y las independencias condicionales, se puede verificar que con dichas probabilidades se puede calcular la probabilidad conjunta

Especificación Paramétrica



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

Especificación Paramétrica

- En general, la probabilidad conjunta se especifica por el producto de las probabilidades de cada variable dados sus padres:

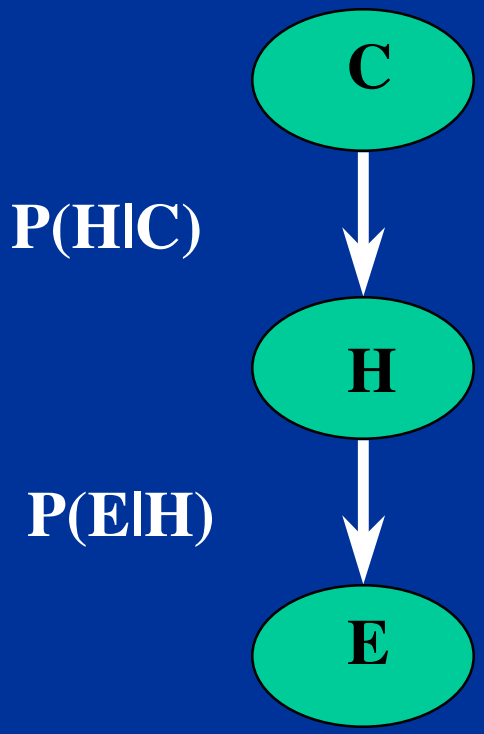
$$P(X1,X2, ..., Xn) = \prod P(Xi \mid Pa(Xi))$$

Inferencia probabilística

- En RB, la inferencia probabilística consiste en:

  "dadas ciertas variables conocidas (evidencia), calcular la probabilidad posterior de las demás variables (desconocidas)"
- Es decir, calcular: P(Xi | **E**), donde:
  - **E** es un subconjunto de variables de la RB (posiblemente vació)
  - Xi es cualquier variable en la RB, no en **E**

Inferencia bayesiana



***Causal*:**

C → H

***Evidencial*:**

E → H

***Mixta*:**

C, E → H

Tipos de Técnicas

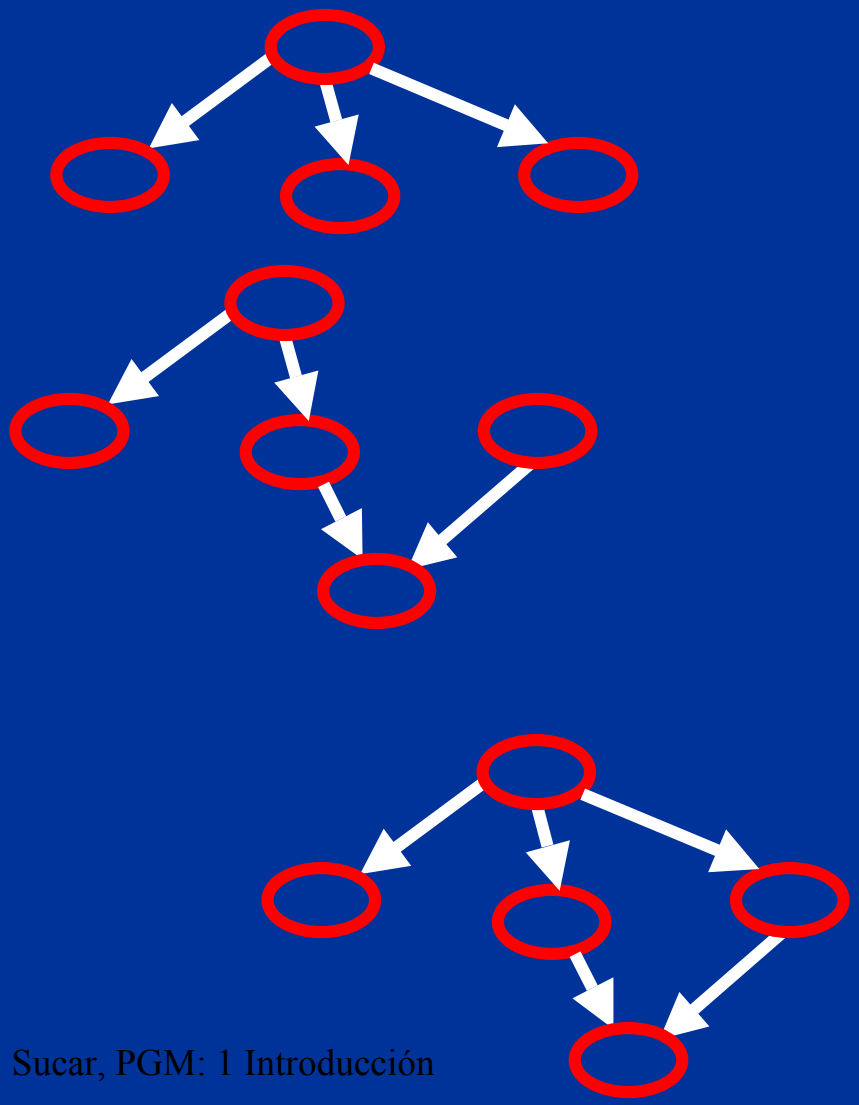
- Calcular probabilidades posteriores:
  - Una variable, cualquier estructura: algoritmo de eliminación (*variable elimination*)
  - Todas las variable, estructuras sencillamente conectadas (árboles, poliárboles): propagación
  - Todas las variables, cualquier estructura:
    - Agrupamiento (*junction tree*)
    - Simulación estocástica
    - Condicionamiento

Tipos de Técnicas

- Obtener variable(s) de mayor probabilidad dada cierta evidencia – abducción:
  - Abducción total
  - Abducción parcial

Tipos de estructuras

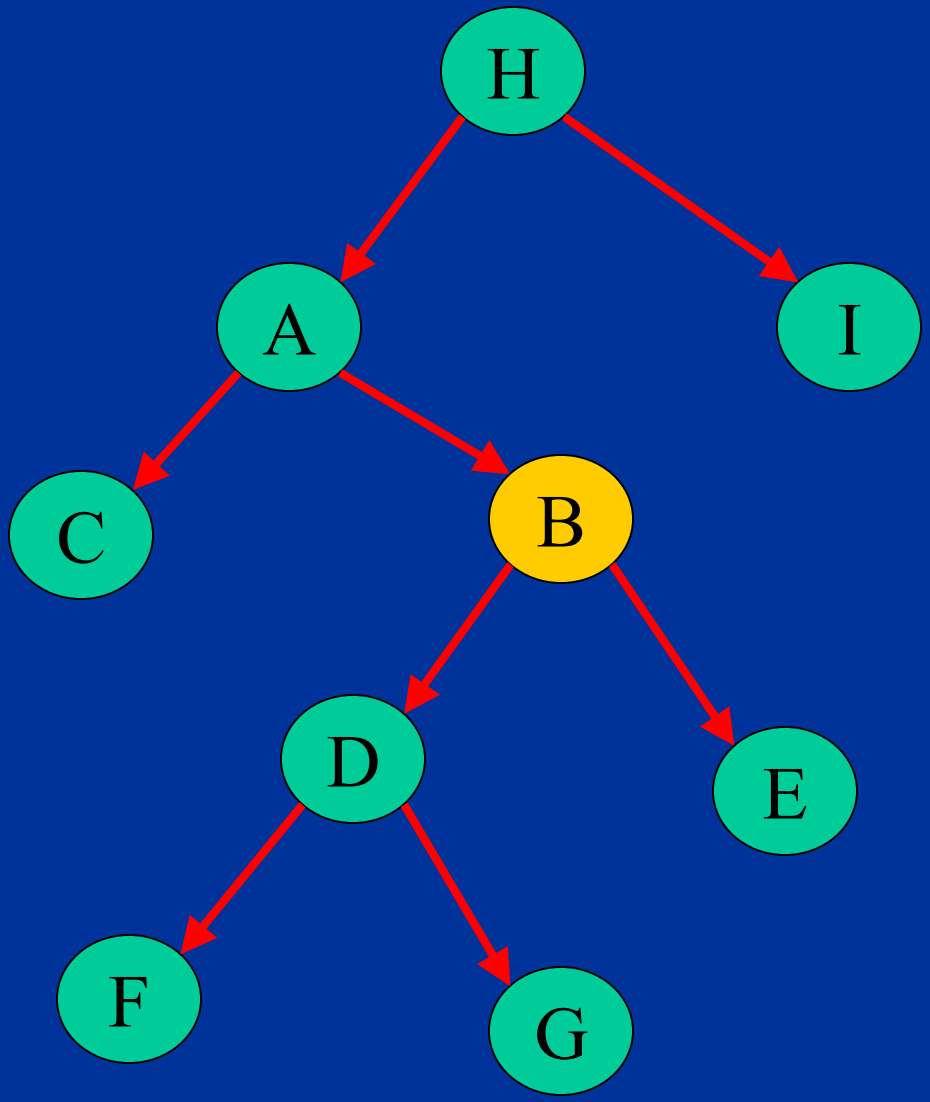
- Sencillamente conectadas
  - Árboles
  - Poliárboles
- Multiconectadas



Propagación en Árboles

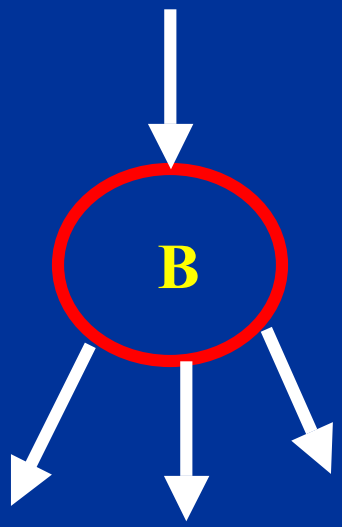
Cada nodo corresponde a una variable discreta, B ($B_1, B_2, \ldots, B_m$) con su respectiva matriz de probabilidad condicional, $P(B|A)=P(B_j|A_i)$

Propagación en Árboles

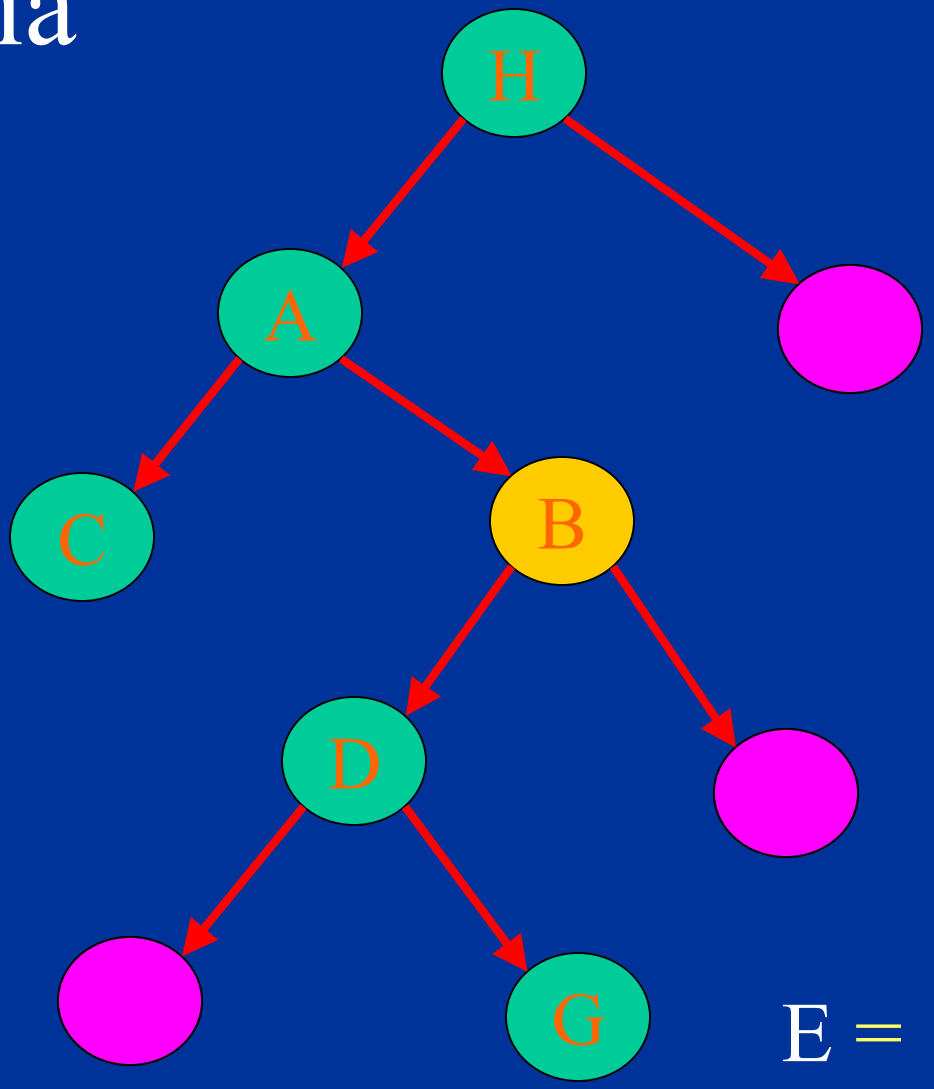


Dada cierta evidencia *E* -representada por la instanciación de ciertas variables- la probabilidad posterior de cualquier variable *B*, por el teorema de Bayes:

$$P(B_i \mid E) = P(B_i)\, P(E \mid B_i) / P(E)$$



Evidencia



$E = \{I,F,E\}$

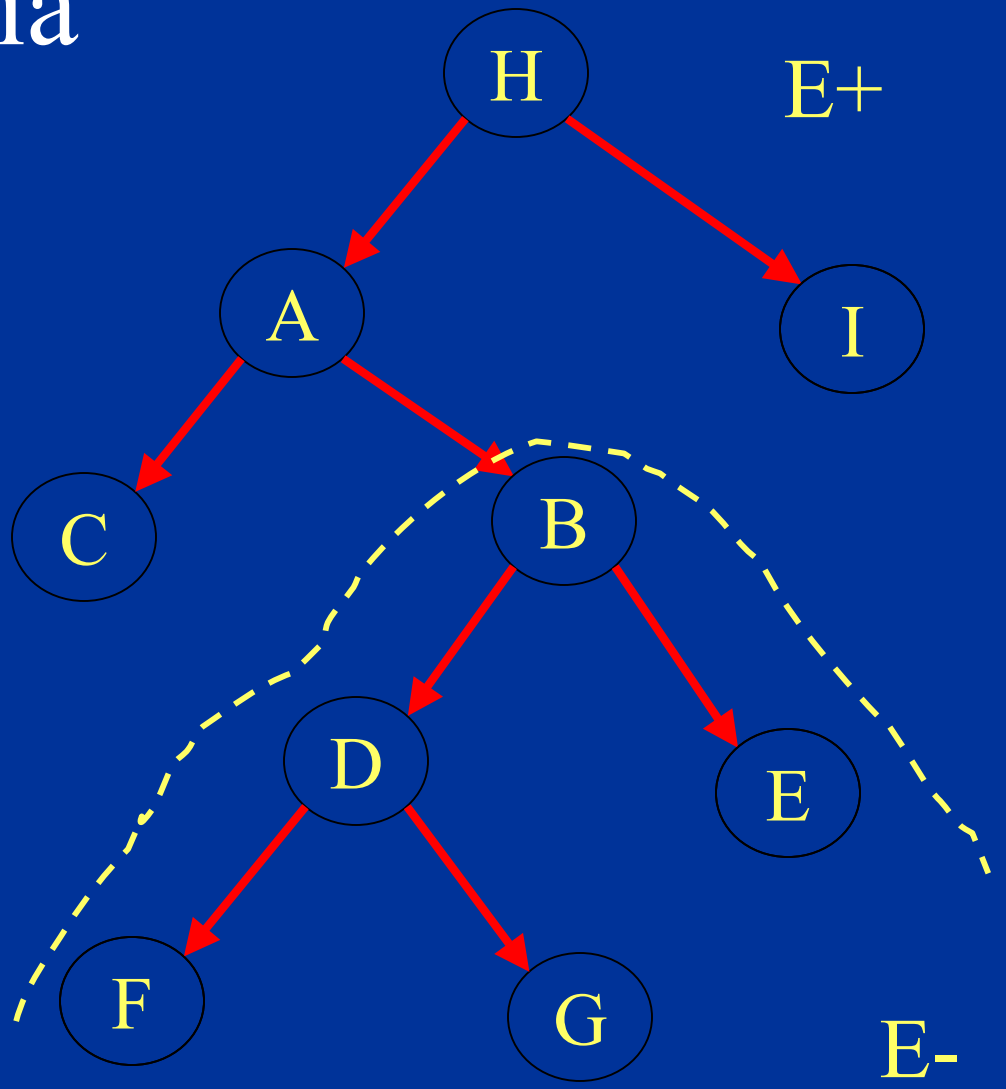
Evidencia

Ya que la estructura de la red es un árbol, el Nodo *B* la separa en dos subárboles, por lo que podemos dividir la evidencia en dos grupos:

E-: Datos en el árbol que cuya raíz es *B*

E+: Datos en el resto del árbol

Evidencia



Entonces:

$P(B_i \mid E) = P(B_i)\, P(E^-, E^+ \mid B_i) / P(E)$

Pero dado que ambos son independientes y aplicando nuevamente Bayes:

$P(B_i \mid E) = \alpha\, P(B_i \mid E^+)\, P(E^- \mid B_i)$

Donde α es una constante de normalización

Definiciones:

Si definimos los siguientes términos:

$$\lambda(B_i) = P(E^- \mid B_i)$$

$$\rho(B_i) = P(B_i \mid E^+)$$

Entonces:

$$P(B_i \mid E) = a\, \rho(B_i)\, \lambda(B_i)$$

Desarrollo

- En base a la ecuación anterior, se puede integrar un algoritmo distribuido para obtener la probabilidad de un nodo dada cierta evidencia
- Para ello se descompone el cálculo de cada parte:
  - Evidencia de los hijos (λ)
  - Evidencia de los demás nodos (π)

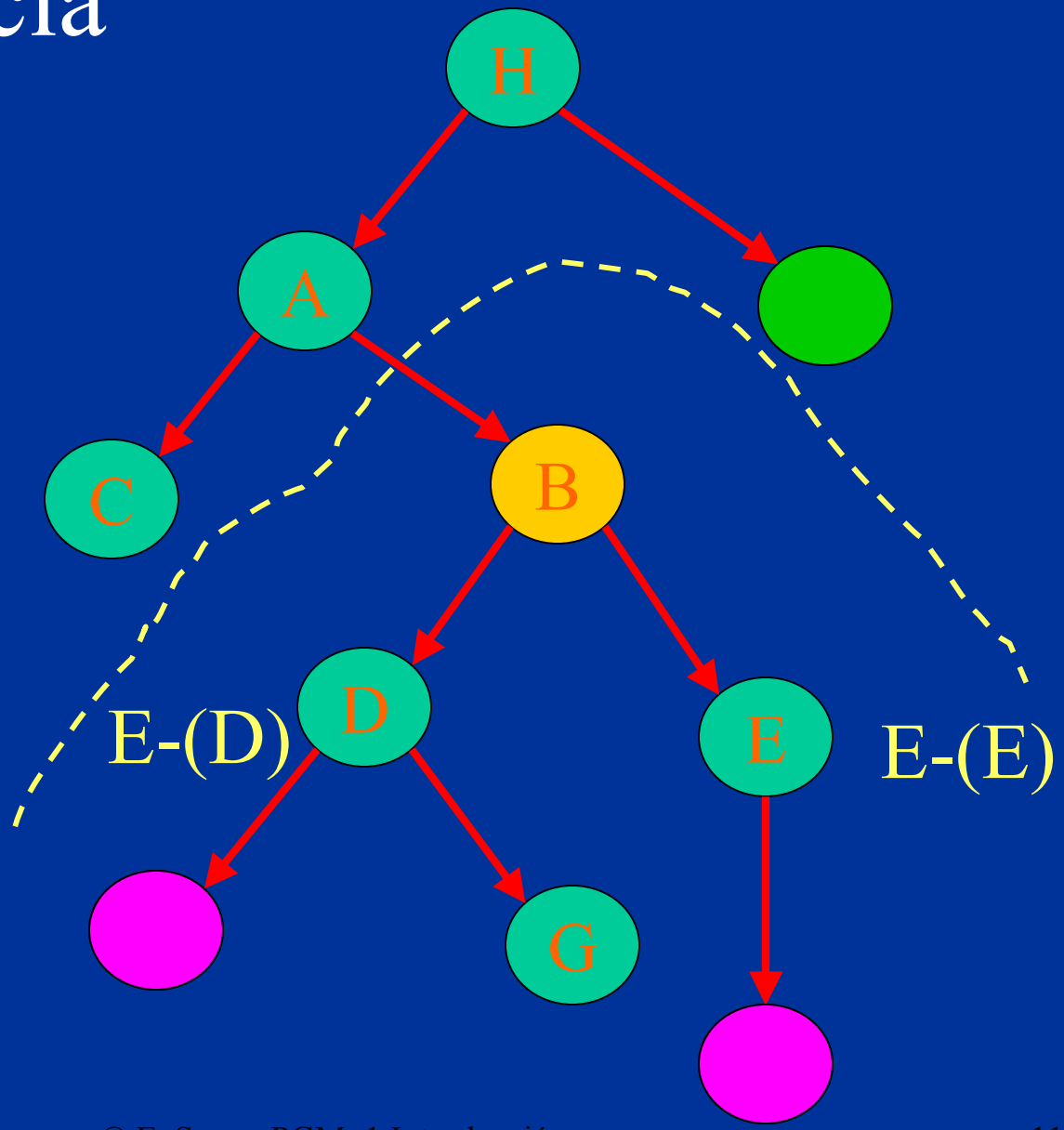
Evidencia de los hijos (l)

- Dado que los hijos son condicionalmente independientes dado el padre:

$$\lambda(B_i) = P(E^- \mid B_i) = \prod_k P(E_k^- \mid B_i)$$

- Donde E_k^- corresponde a la evidencia del subárbol del hijo k

Evidencia hijos



Evidencia de los hijos (λ)

- Condicionando respecto a los posibles valores de los hijos de B:

$$\lambda(B_i) = \prod_k \left[\sum_j P(E_k^- \mid B_i, S_j^k) P(S_j^k \mid B_i) \right]$$

- Donde S^k es el hijo k de B, y la sumatoria es sobre los valores de dicho nodo (teorema de probabilidad total)

Evidencia de los hijos (λ)

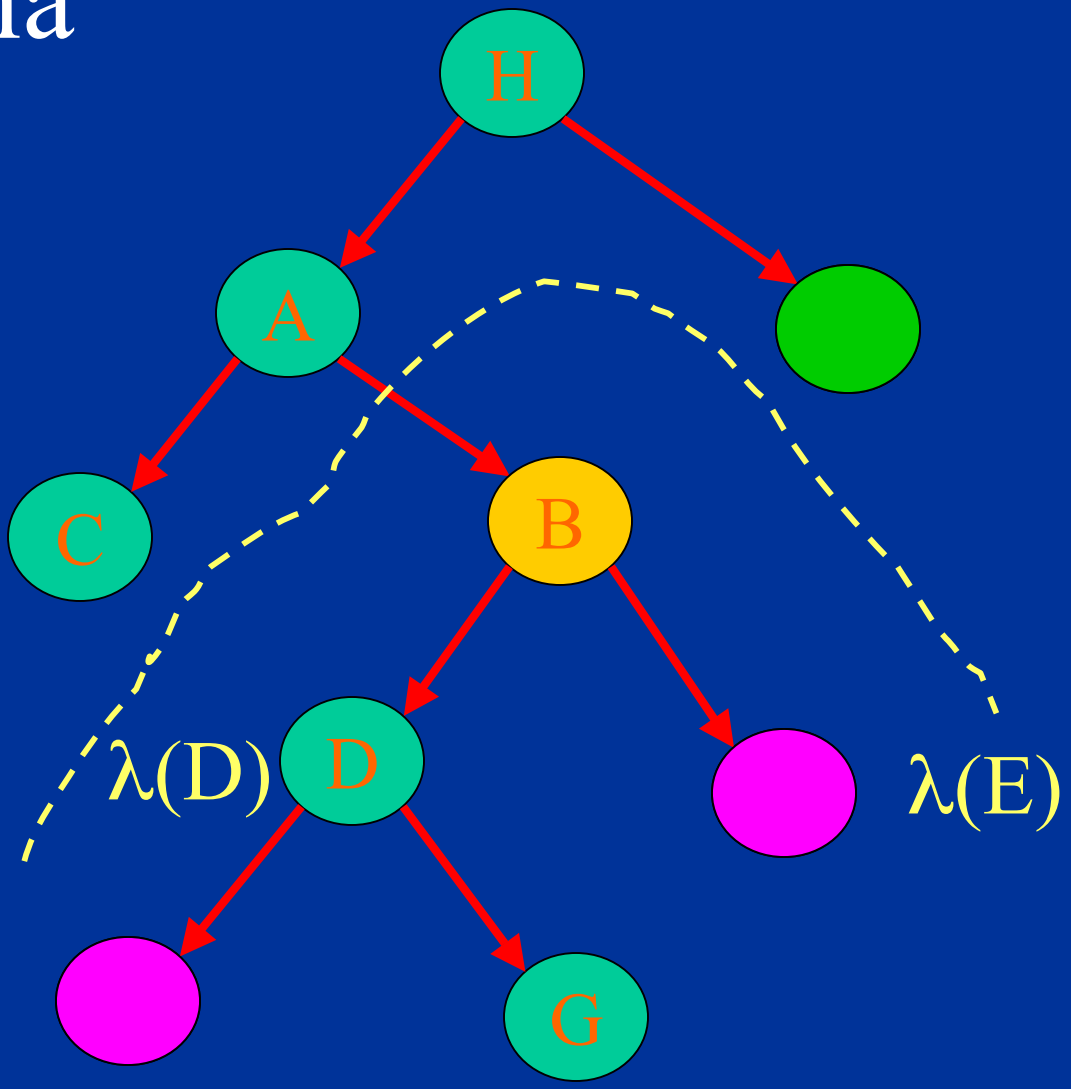
- Dado que B es condicionalmente independiente de la evidencia dados sus hijos:

$$\lambda(B_i) = \prod_k [\sum_j P(E_k^- \mid S_j^k) P(S_j^k \mid B_i)]$$

- Substituyendo la defínción de λ:

$$\lambda(B_i) = \prod_k [\sum_j P(S_j^k \mid B_i) \lambda(S_j^k)]$$

Evidencia hijos



Evidencia de los hijos (l)

- Recordando que λ es un vector (un valor por cada posible valor de B), lo podemos ver en forma matricial:

$$\lambda = \lambda \quad P(S \mid B)$$

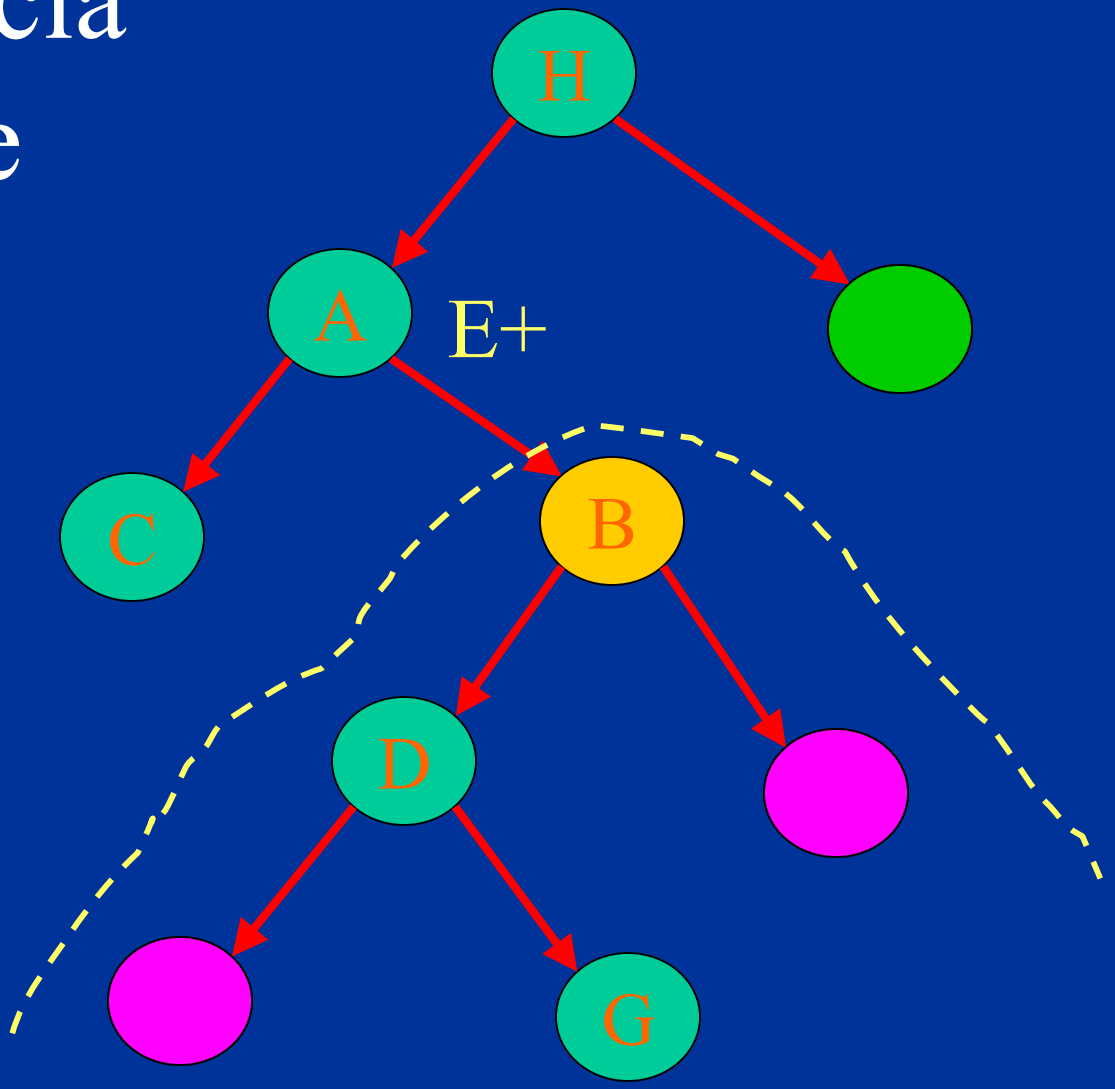
Evidencia de los demás nodos (π)

- Condicionando sobre los diferentes valores del nodo padre (A):

$$\pi(B_i) = P(B_i \mid E^+) = \sum_j P(B_i \mid E^+, A_j)\, P(A_j \mid E^+)$$

- Donde A_j corresponde a los diferentes valores del nodo padre de B

Evidencia padre



Evidencia de los demás nodos (p)

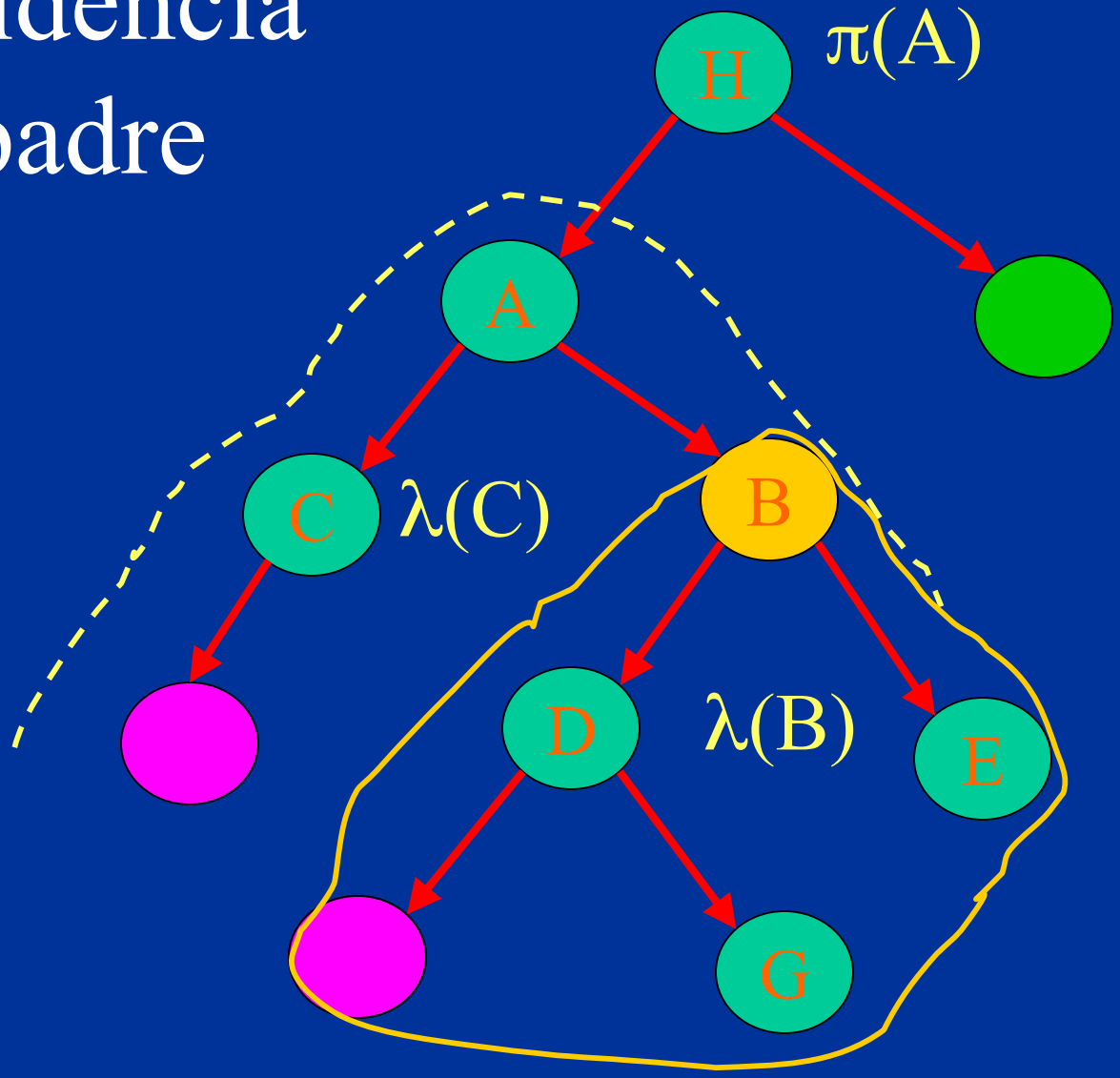
- Dado que B es independiente de la evidencia "arriba" de A, dado A:

$$\Pi\,(B_i) = \Sigma_j\, P\,(B_i \mid A_j)\, P(A_j \mid E^+)$$

- La $P(A_j \mid E^+)$ corresponde a la P posterior de A dada toda la evidencia excepto B y sus hijos, por lo que se puede escribir como:

$$P(A_j \mid E^+) = \alpha\, \pi\,(A_i)\, \Pi_{k^1 B}\, \lambda_k\,(A_i)$$

Evidencia padre



Evidencia de los demás nodos (π)

- Substituyendo $P(A_j \mid E^+)$ en la ecuación de π :

$$\pi(B_i) = \sum_j P(B_i \mid A_j) \left[\alpha \, \pi(A_i) \prod_{k^1 B} \lambda_k(A_i) \right]$$

- De forma que se obtiene combinando la π de del nodo padre con la λ de los demás hijos

Evidencia de los demás nodos (p)

- Dado que también π es un vector, lo podemos ver en forma matricial (donde P_A es el producto de la evidencia de padre y otros hijos):

$$\pi = P(B \mid A) \, P_A$$

Algoritmo

- Mediante estas ecuaciones se integra un algoritmo de propagación de probabilidades en árboles.
- Cada nodo guarda los valores de los vectores π y λ, así como su matriz de probabilidad condicional (CPT), P.
- La propagación se hace por un mecanismo de paso de mensajes, en donde cada nodo envía los mensajes correspondientes a su padre e hijos

Mensaje al padre (hacia arriba) – nodo B a su padre A:

$$\lambda_B(A_i) = \sum_j P(B_j|A_i)\lambda(B_j)$$

Mensaje a los hijos (hacia abajo) - nodo B a su hijo S_k :

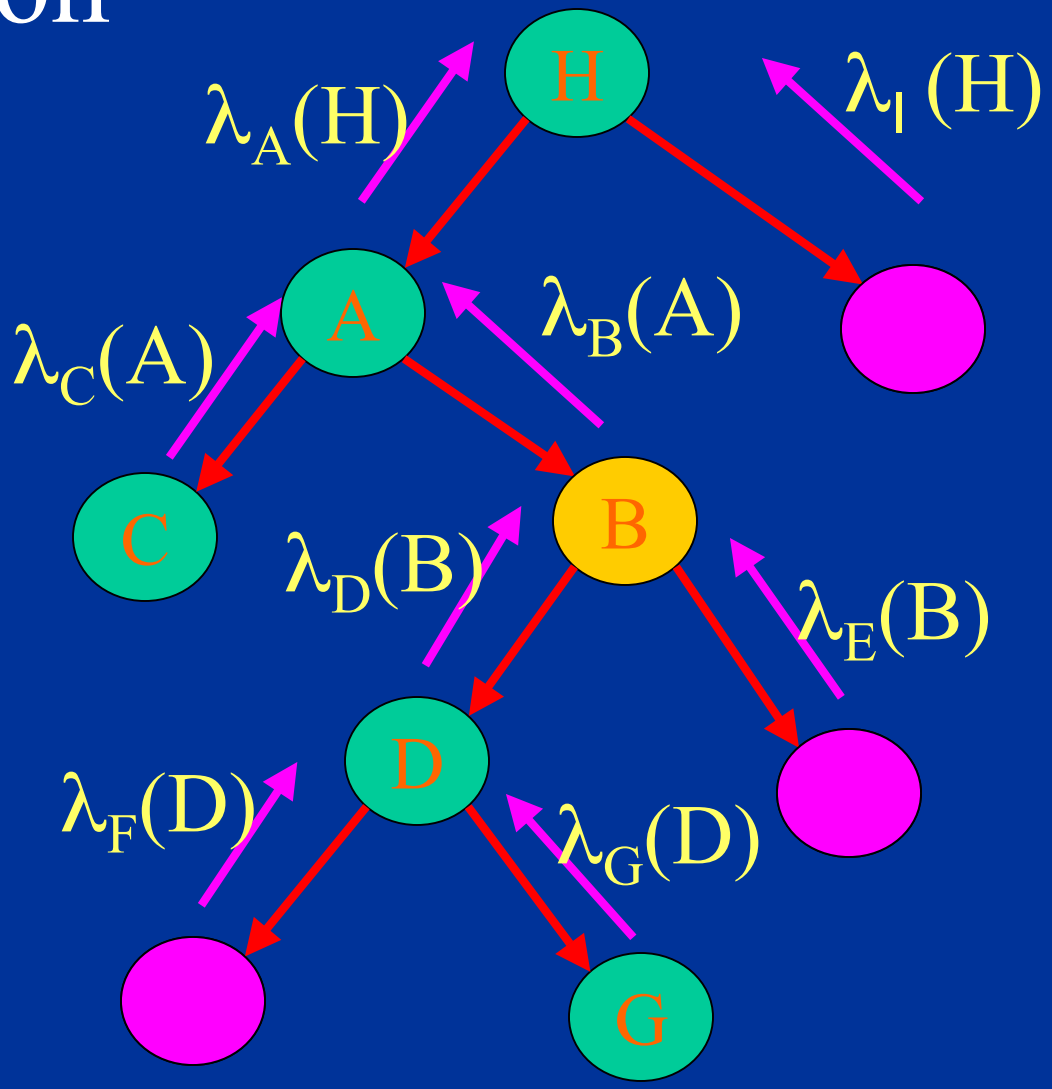
$$\pi_k(B_i) = \alpha\pi(B_j)\prod_{I\neq k}\lambda_I(B_j)$$

Algoritmo

- Al instanciarse ciertos nodos, éstos envían mensajes a sus padres e hijos, y se propagan hasta a llegar a la raíz u hojas, o hasta encontrar un nodo instanciado.

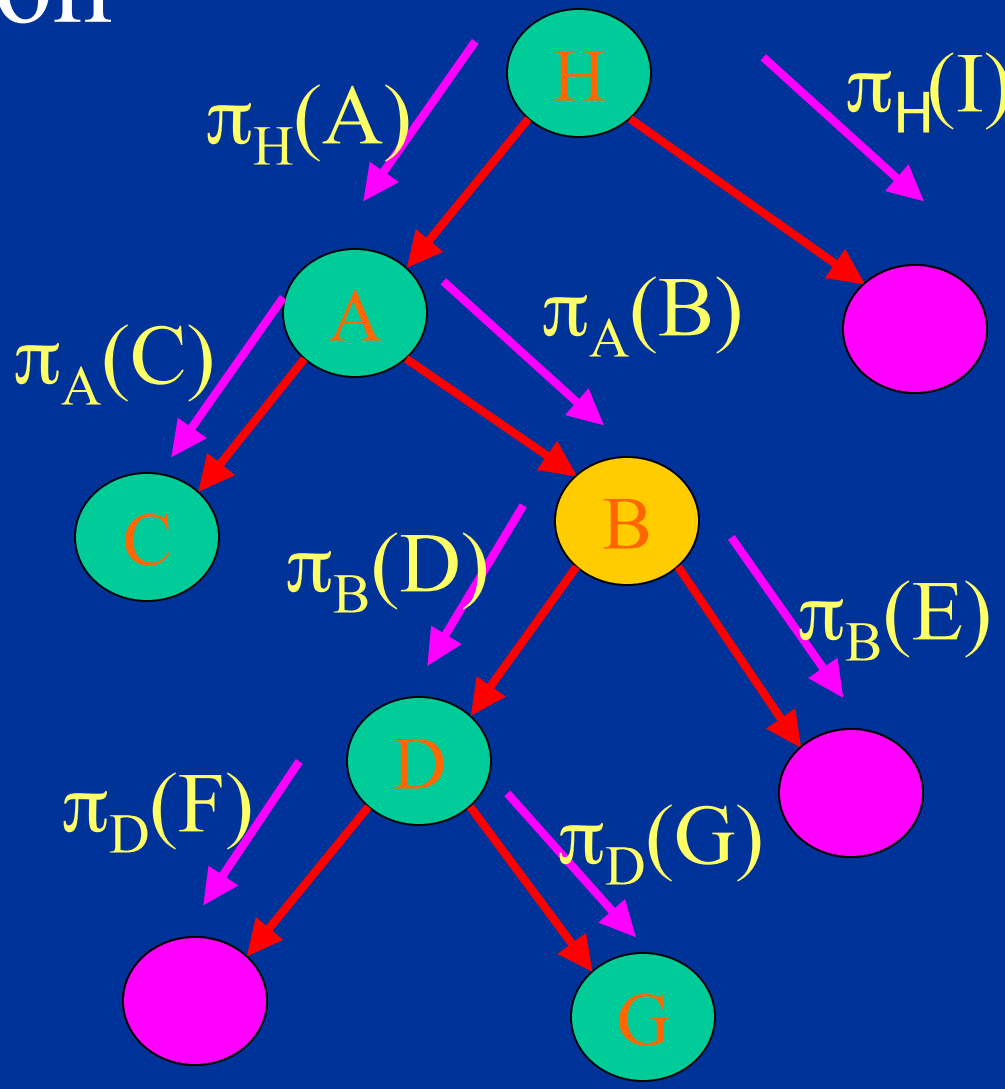
- Así que la propagación se hace en un solo paso, en un tiempo proporcional al diámetro de la red.

Propagación λ



Propagación

π



Condiciones Iniciales

- Nodos hoja no conocidos:

  $\lambda(B_i) = [1,1, \ldots]$

- Nodos asignados (conocidos):

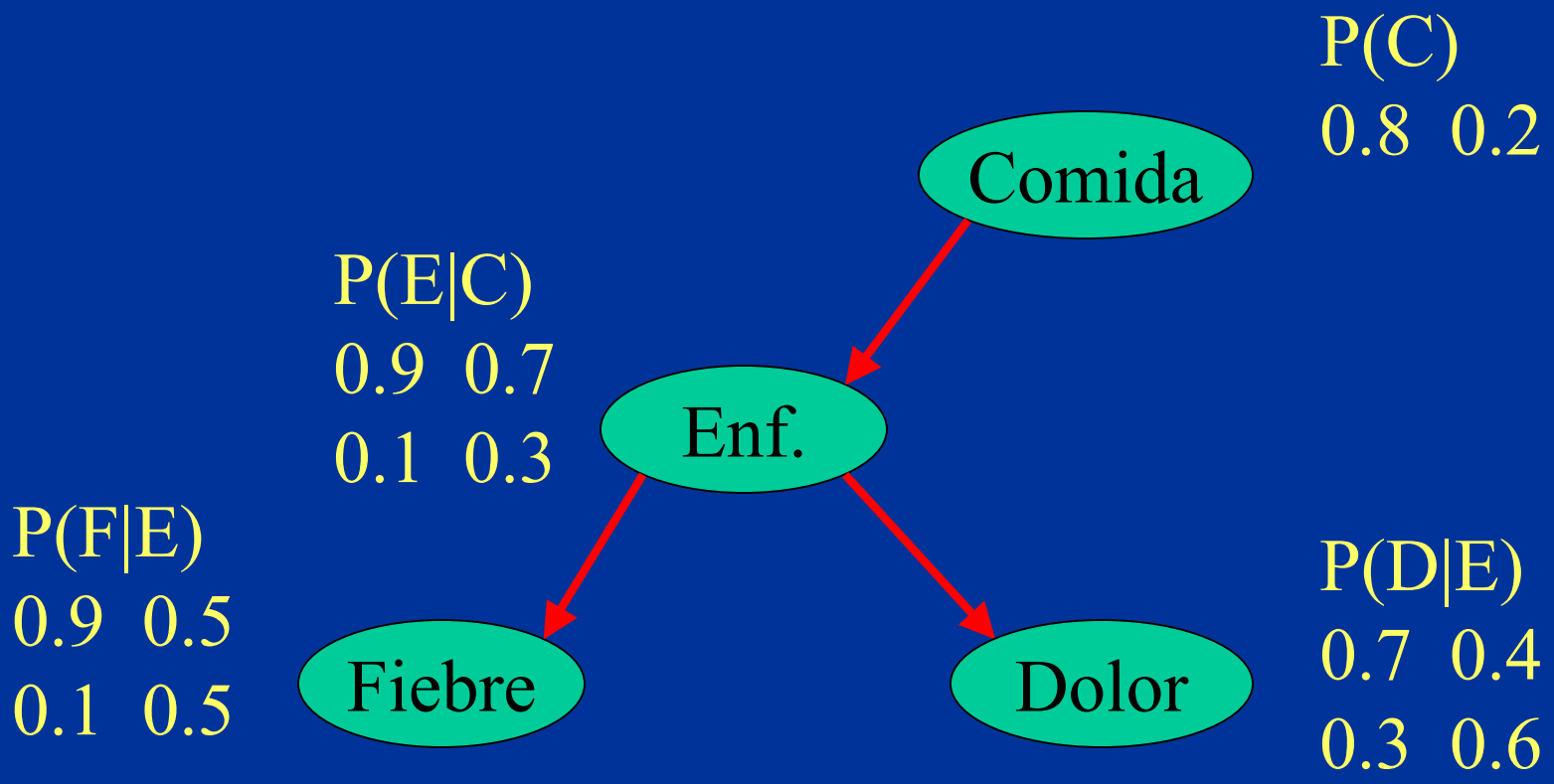
  $\lambda(B_i) = [0,0, ..1, 0, \ldots, 0]$ (1 para valor asignado)

  $\pi(B_i) = [0,0, ..1, 0, \ldots, 0]$ (1 para valor asignado)

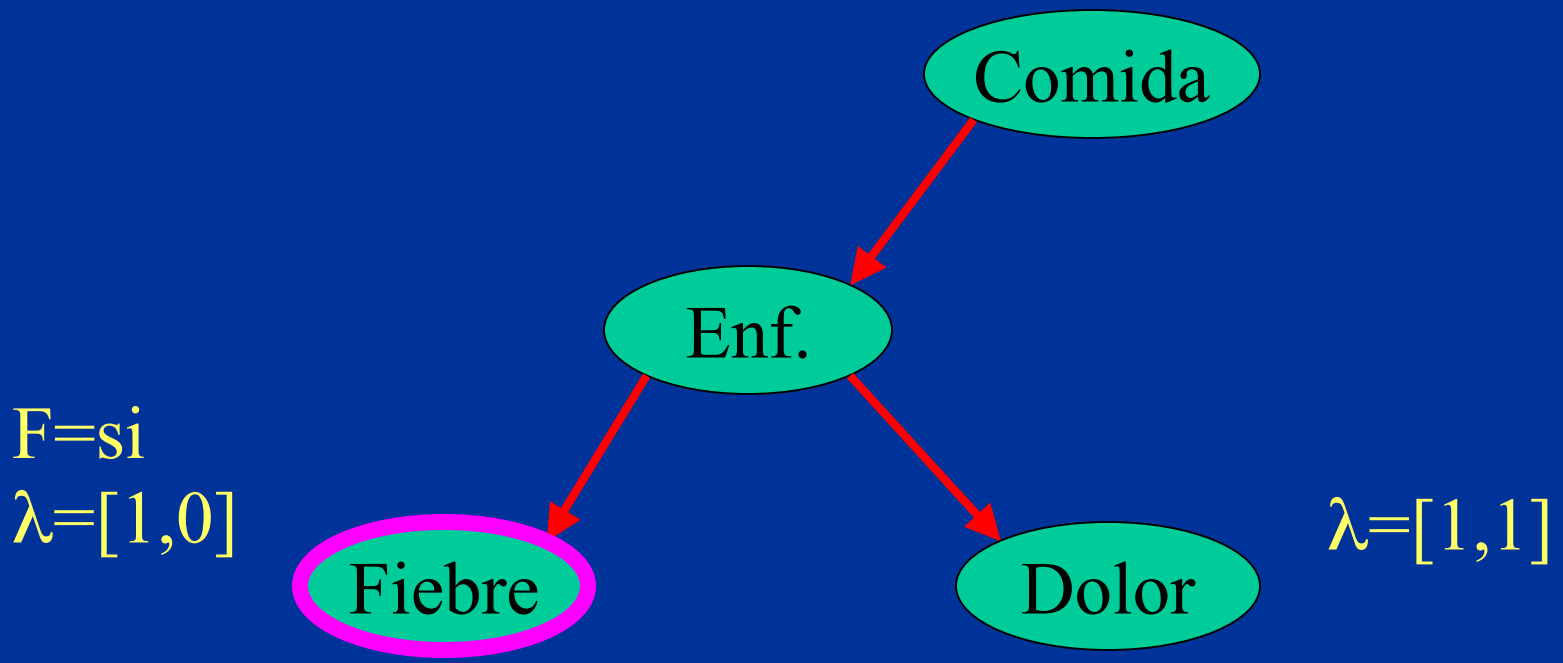
- Nodo raíz no conocido:

  $\pi(A) = P(A)$, (probabilidad marginal inicial)

Ejemplo

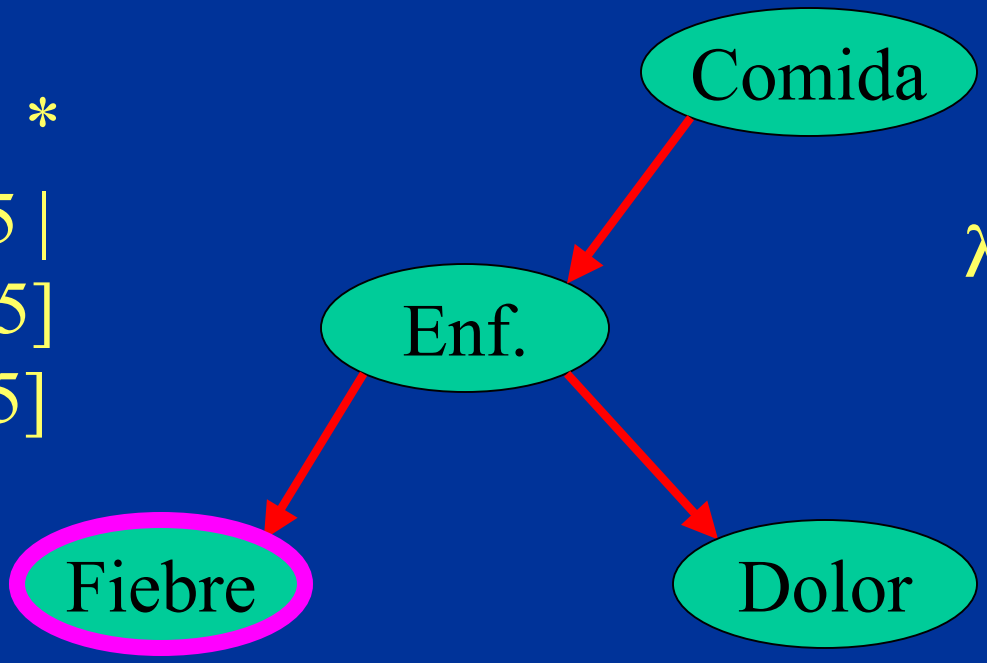

P(C)
0.8 0.2
Comida
P(E|C)
0.9 0.7
0.1 0.3
Enf.
P(F|E)
0.9 0.5
0.1 0.5
Fiebre
Dolor
P(D|E)
0.7 0.4
0.3 0.6

Ejemplo



Ejemplo

λ_F= [1,0] *
[.9 .5 |
.1 .5]
= [.9 .5]



λ_D= [1,1] *
[.7 .4 |
.3 .6]
= [1 1]

P(F|E)
0.9 0.5
0.1 0.5

P(D|E)
0.7 0.4
0.3 0.6

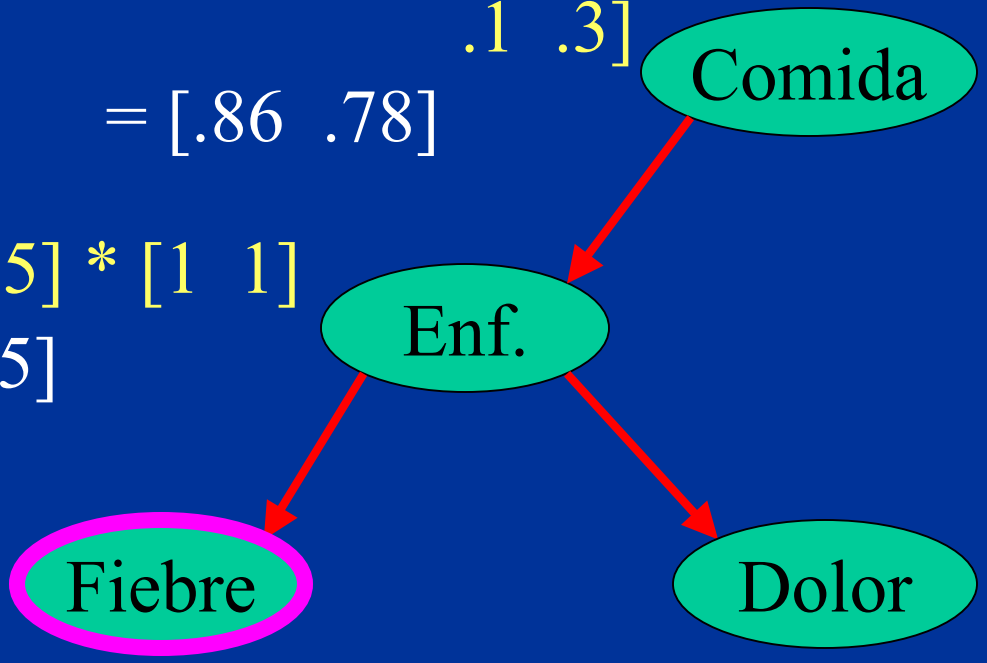
Ejemplo

$\lambda(C) = [.9\ .5] * [.9\ .7 \mid .1\ .3]$

$= [.86\ .78]$

$\lambda(E) = [.9\ .5] * [1\ 1]$

$= [.9\ .5]$



P(E|C)

0.9	0.7
0.1	0.3

P(F|E)

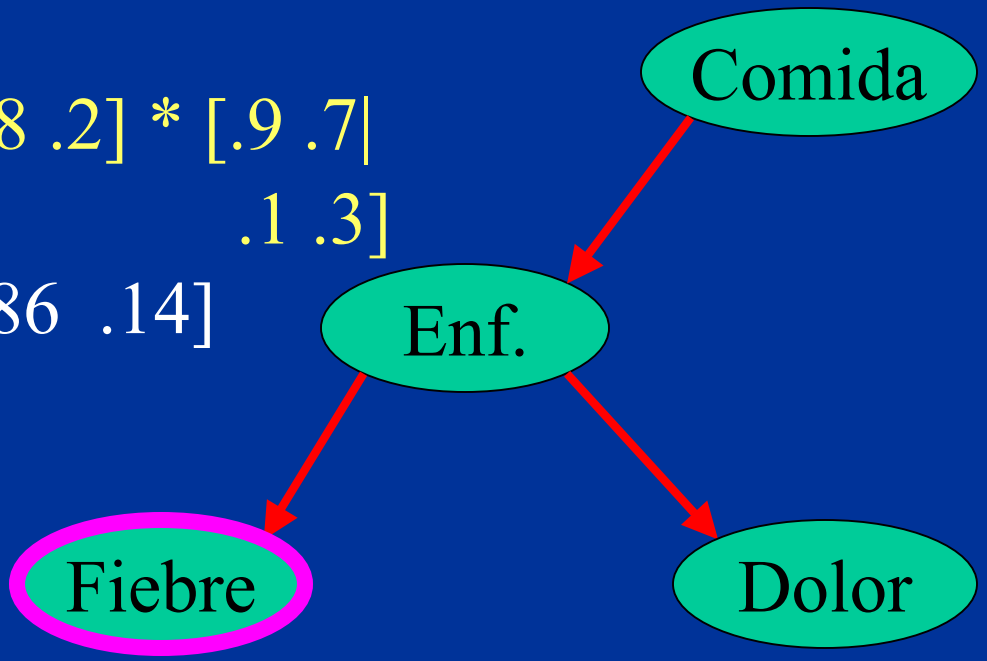
0.9	0.5
0.1	0.5

P(D|E)

0.7	0.4
0.3	0.6

Ejemplo

$\pi(C) = [.8 \quad .2]$



$\pi(E) = [.8 \ .2] * [.9 \ .7 | .1 \ .3]$

$= [.86 \quad .14]$

P(E|C)

0.9	0.7
0.1	0.3

P(F|E)

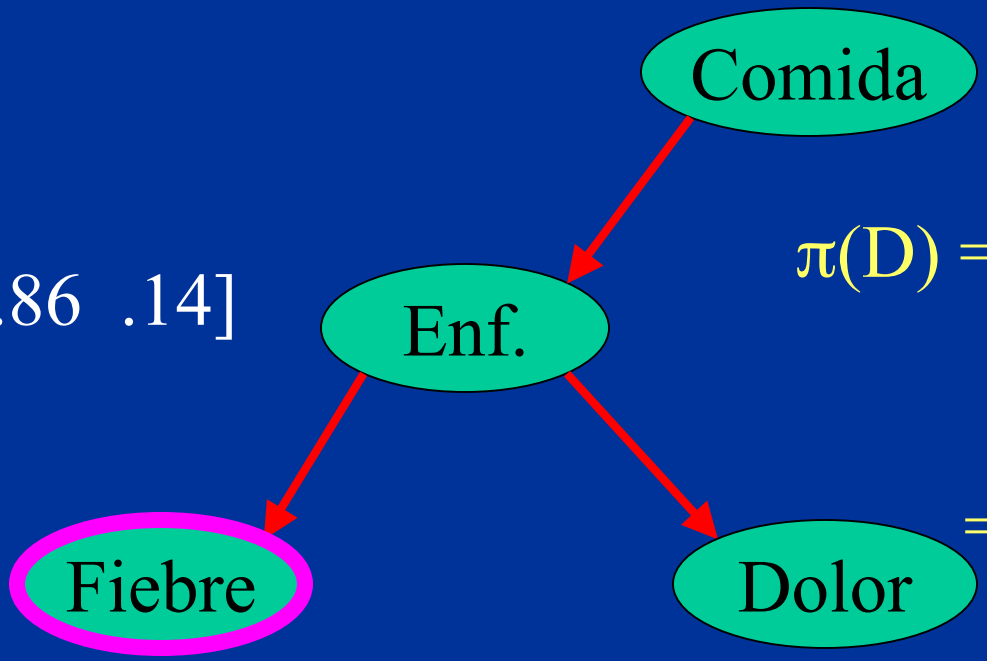
0.9	0.5
0.1	0.5

P(D|E)

0.7	0.4
0.3	0.6

Ejemplo

$\pi(C) = [.8 \quad .2]$



$\pi(E) = [.86 \quad .14]$

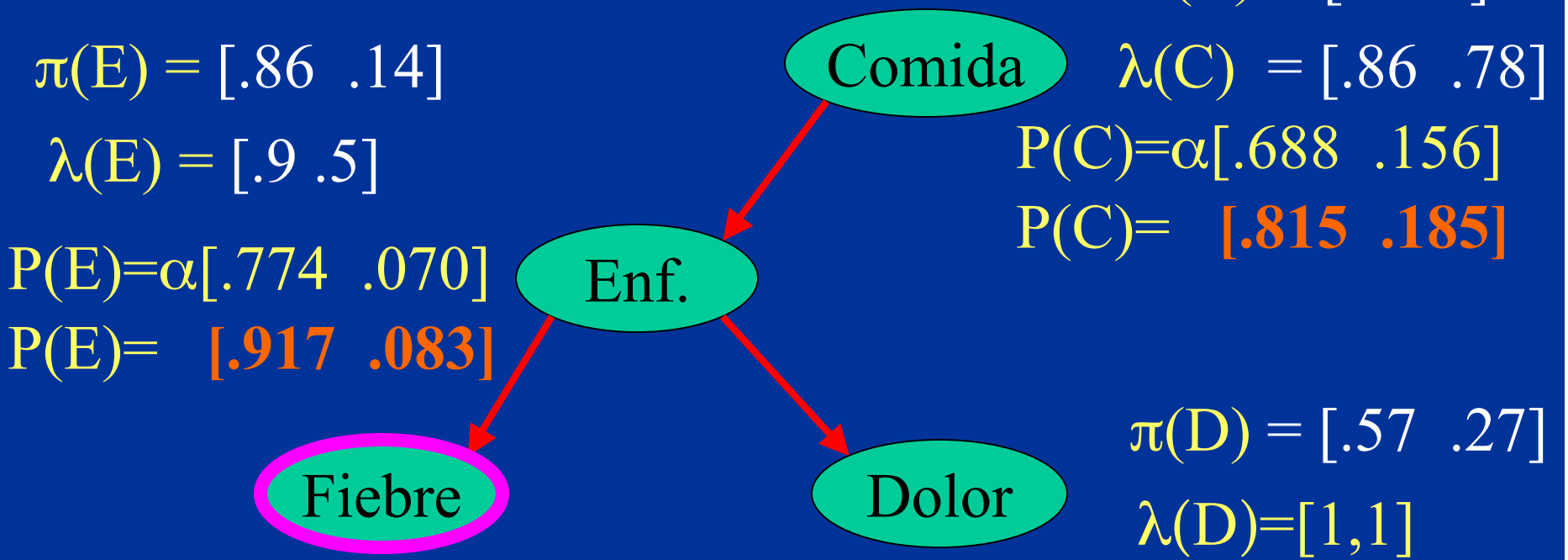
$\pi(D) = [.86 \ .14] * \begin{bmatrix} .9 & .5 \\ .7 & .4 \\ .3 & .6 \end{bmatrix}$

$= [.5698 \quad .2742]$

$P(D|E)$

$\begin{matrix} 0.7 & 0.4 \\ 0.3 & 0.6 \end{matrix}$

Ejemplo


π(C) = [.8 .2]
Comida
λ(C) = [.86 .78]
P(C)=α[.688 .156]
P(C)= [.815 .185]
π(E) = [.86 .14]
λ(E) = [.9 .5]
P(E)=α[.774 .070]
P(E)= [.917 .083]
Enf.
Fiebre
Dolor
π(D) = [.57 .27]
λ(D)=[1,1]
P(D)=α[.57 .27]
P(D)= [.67 .33]

Demo 1

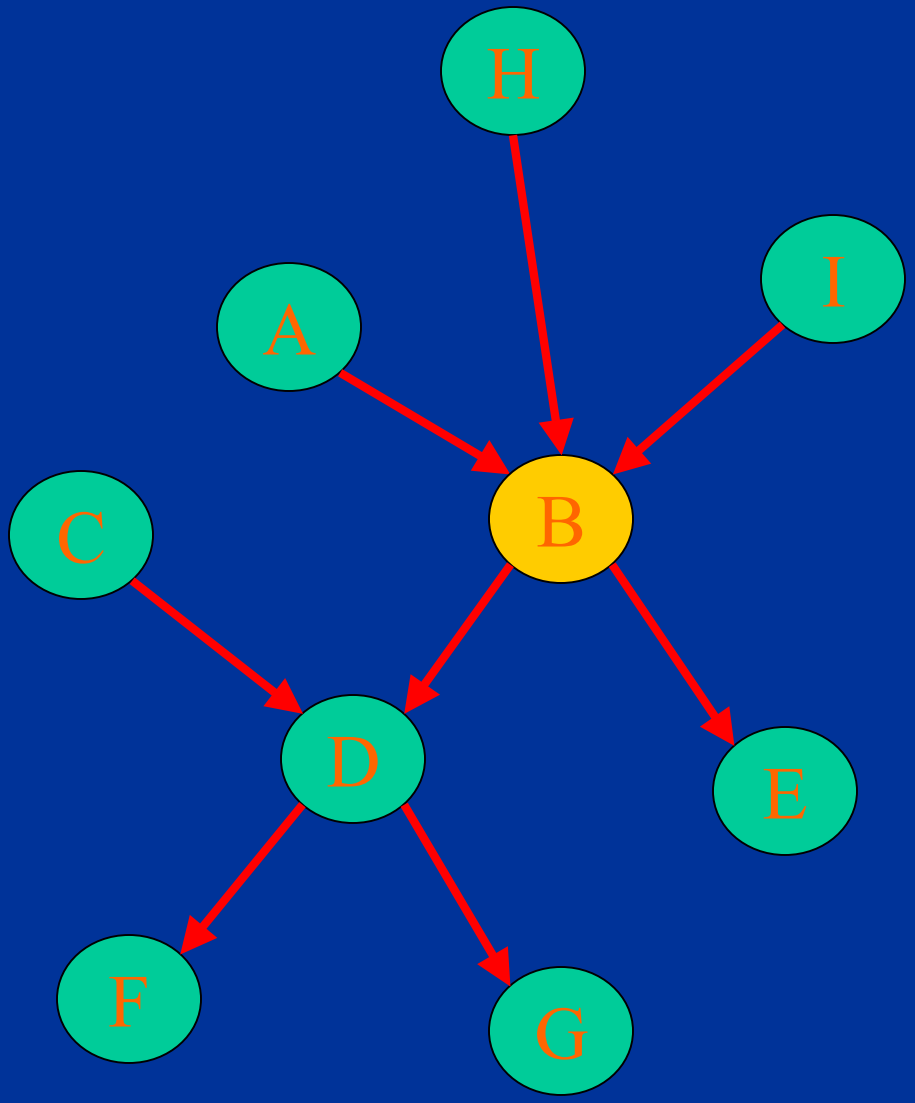
- Ejemplo en HUGIN

Propagación en poliárboles

- Un poliárbol es una red conectada en forma sencilla, pero en la que un nodo puede tener varios padres:

$$P(B \mid A1, A2, \ldots, An)$$

Propagación en Poliárboles



Algoritmo

- El método es muy similar al de árboles, con algunas consideraciones adicionales:
  - Considerar la probabilidad condicional del nodo dados todos sus padres para el cálculo de π y λ
  - Enviar los mensajes λ a cada uno de los padres de un nodo

Propagación en redes multiconectadas

- Una red multiconectada es un grafo no conectado en forma sencilla, es decir, en el que hay múltiples trayectorias entre nodos.
- Para este tipo de redes existen varios tipos de técnicas de inferencia:
  - **Propagación "Loopy"**
  - **Condicionamiento**
  - **Simulación estocástica**
  - **Agrupamiento**

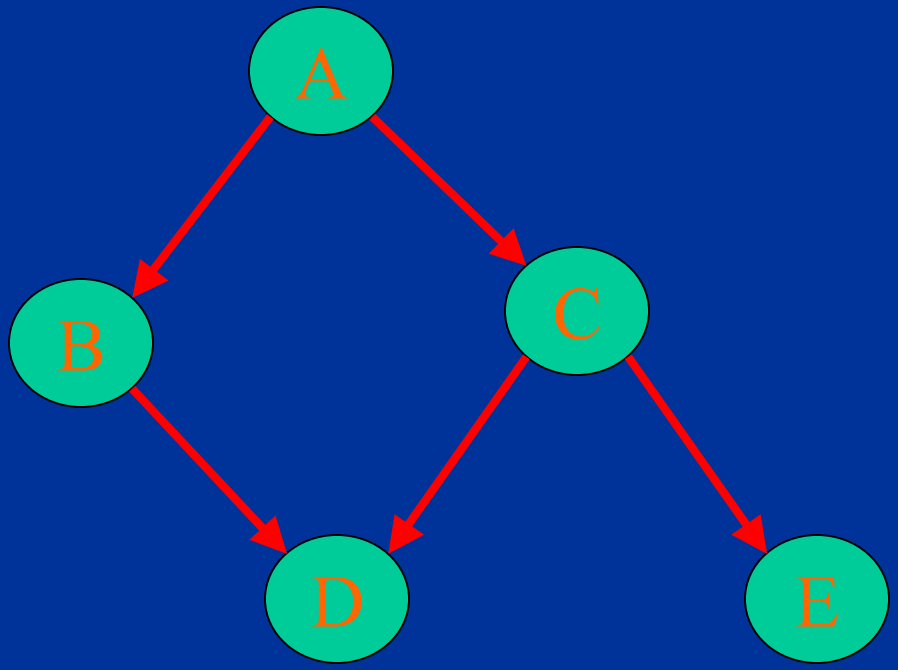
Agrupamiento

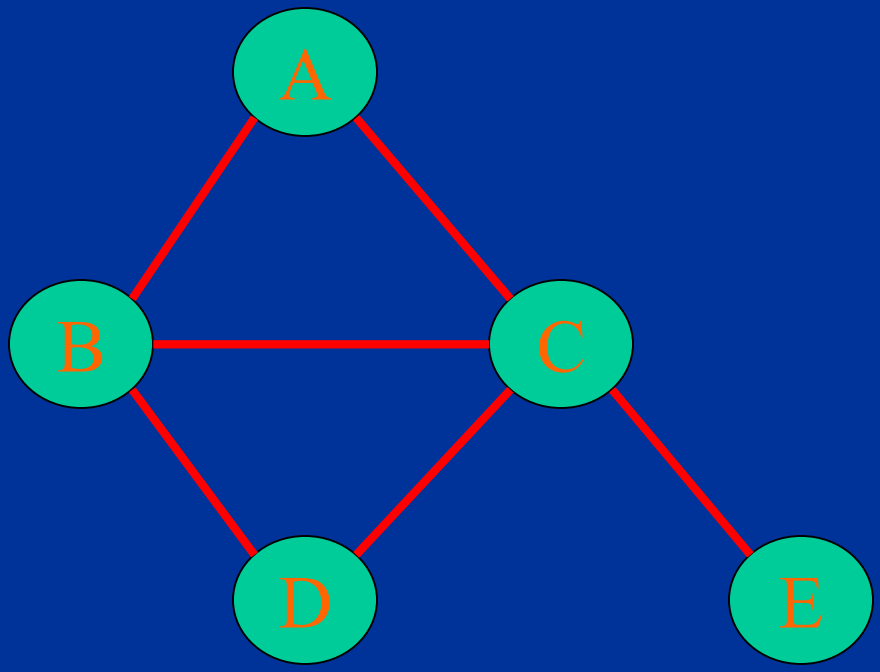
- El método de agrupamiento consiste en transformar la estructura de la red para obtener un árbol, mediante agrupación de nodos usando la teoría de grafos.
- La propagación se realiza sobre el árbol de macro-nodos obtenido, donde cada macro-nodo corresponde a un clique o *unión* de la RB original (*junction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

Agrupamiento

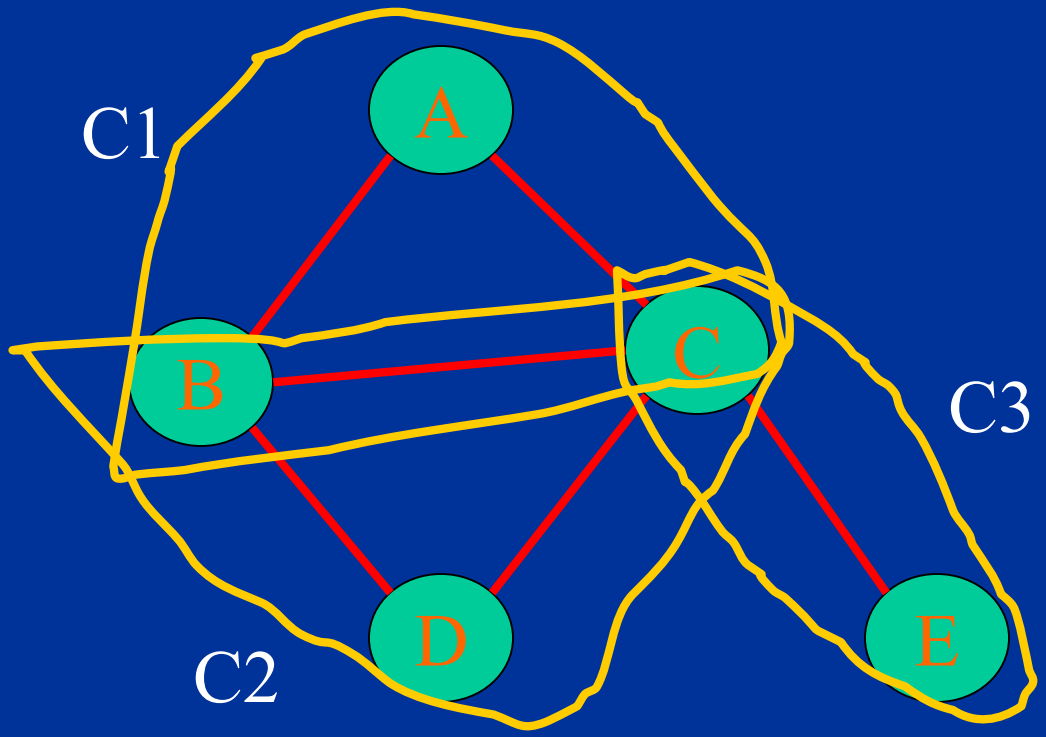
- Transformación:
  - Eliminar direccionalidad de los arcos
  - Ordenamiento de los nodos por máxima cardinalidad
  - Moralizar el grafo (arco entre nodos con hijos comunes)
  - Triangular el grafo
  - Obtener los *cliques* y ordenar
  - Construir árbol de *cliques*

Ejemplo

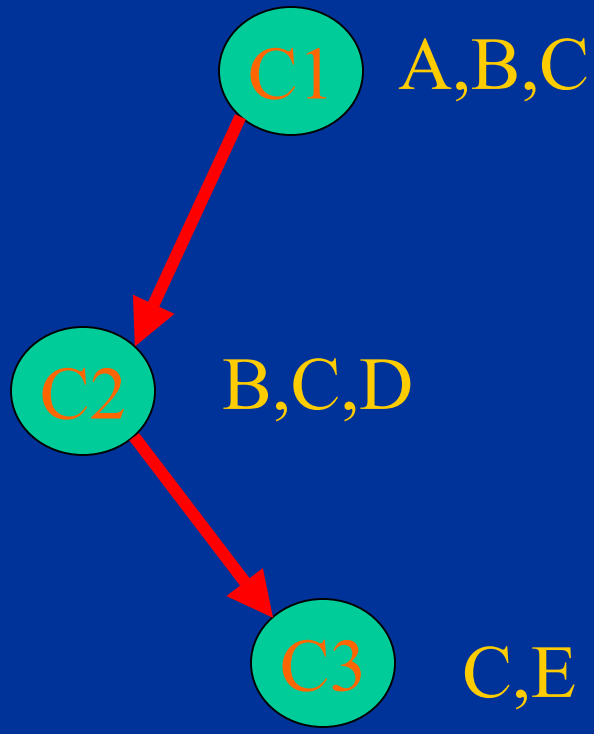




Ordenamiento de Cliques



Árbol de Cliques



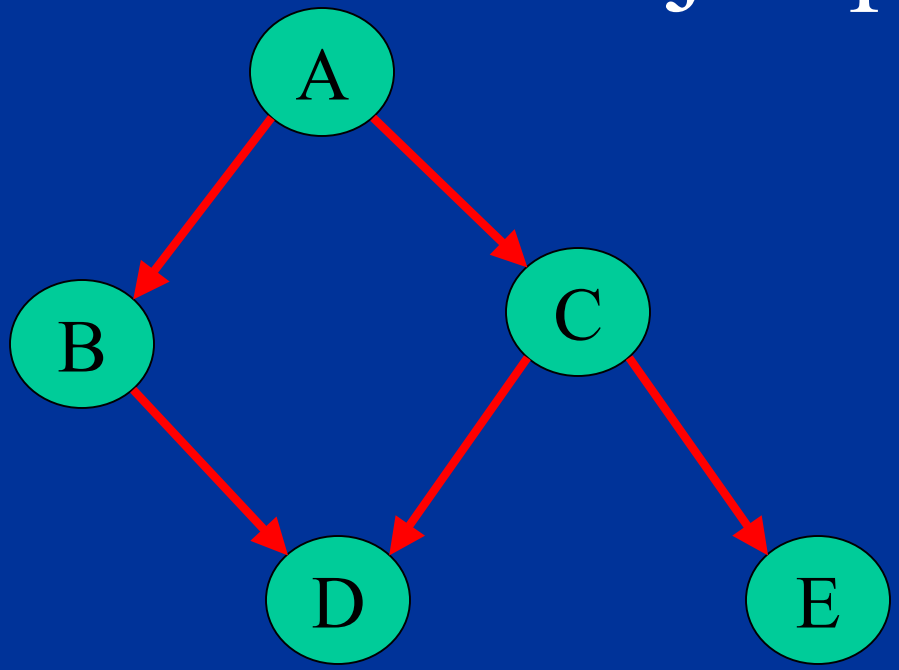
Propagación

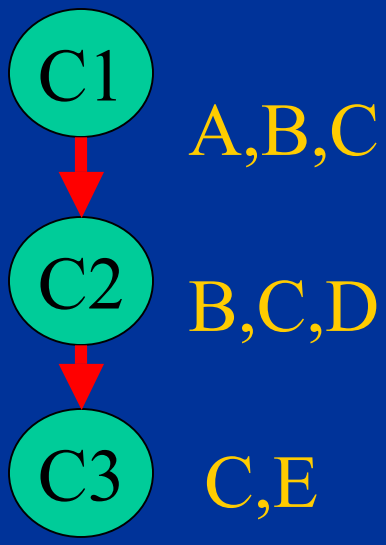
- La propagación es mediante el envío de mensajes en el árbol de *cliques* (en forma similar a árboles)
- Inicialmente se calcula la probabilidad conjunta (potencial) de cada *clique*, y la condicional dado el padre
- Dada cierta evidencia se recalculan las probabilidades de cada *clique*
- La probabilidad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





A,B,C

B,C,D

C,E

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

$$\lambda(Ci) = \Sigma_R \, Y(Ci)$$

- Actualización:

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

- Enviar λ a padre

Propagación sin evidencia:

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

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

- 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":

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

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

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

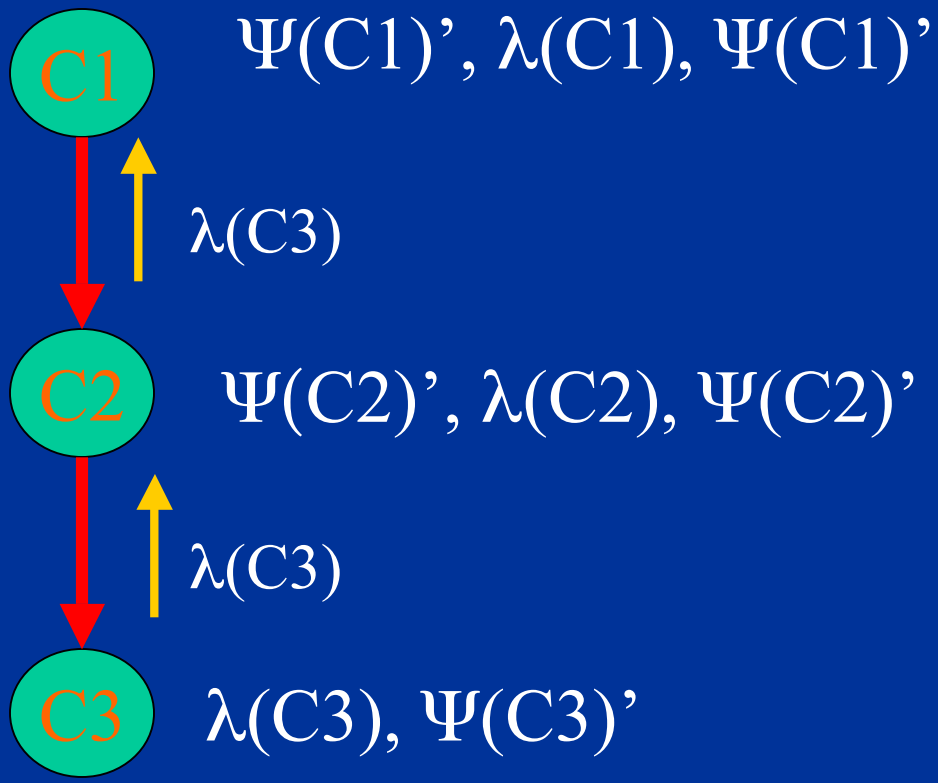
Propagación sin evidencia:

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

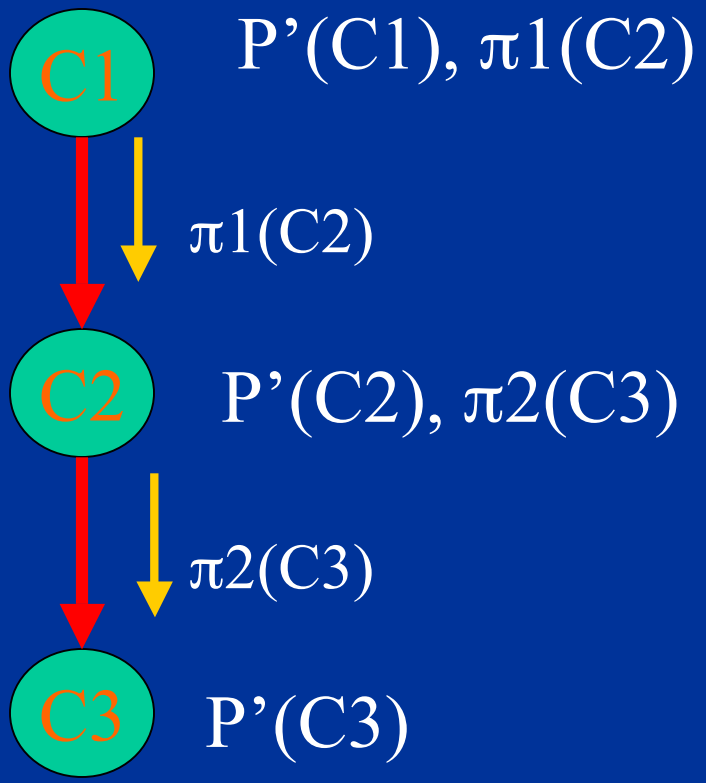
$$P'(Ci) = \pi(Ci)\ \Psi'(Ci)$$

  - 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:
  - 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} { Æ}, { Æ}
- Potenciales:
  - Ψ({A,B,C})
  - Ψ({B,C}) ← D=di
  - Ψ({C}) ← 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

Referencias

- Pearl 88 – Cap. 4,5
- Neapolitan 90 – Cap. 6,7,8
- Jensen 01 – Cap. 5
- Darwiche 09 – Cap. 4-8

I.5:
Aprendizaje de Redes Bayesianas

"Preferiría descubrir una ley causal que ser rey de Persia" [Democritus]

Aprendizaje

El aprendizaje inductivo consiste en obtener conocimiento a partir de datos.

En redes bayesianas se divide en 2 aspectos:

- Obtener la estructura de la red – aprendizaje estructural
- Obtener las probabilidades asociadas – aprendizaje paramétrico

Aprendizaje Paramétrico

- Datos completos - se estiman las probabilidades a partir de frecuencias

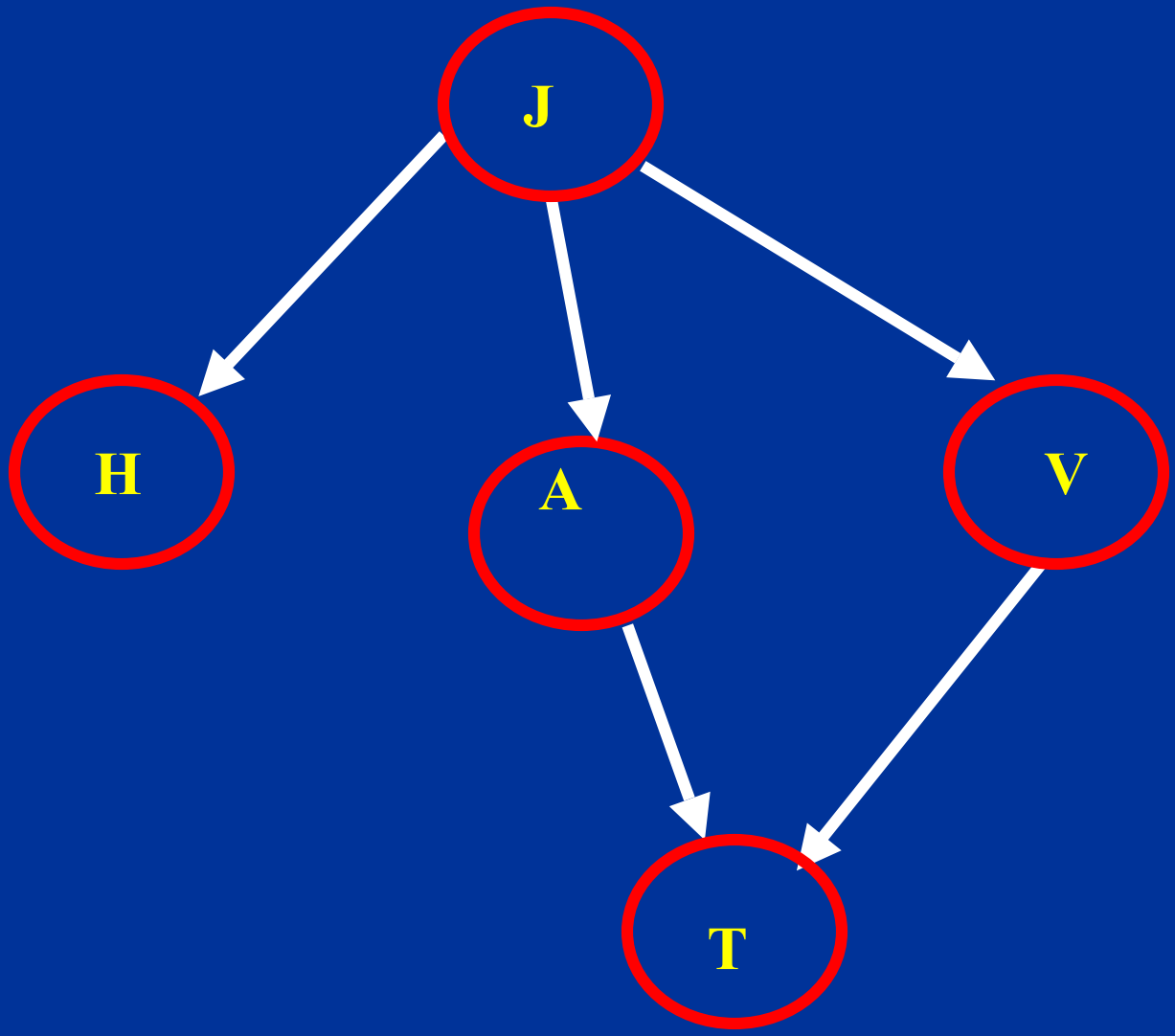
$$P(A) \sim N_a / N_t$$

$$P(B|A1, ..., An) \sim N_{a1, ..., an, b} / N_{a1, ..., an}$$

Ejemplo - ¿Cuándo jugar golf?

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	no	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvia	media	alta	si	N

Ejemplo – estructura



Ejemplo

- P(J)
  - P(N) = 5/14
  - P(P) = 9/14
- P(V|J)
  - P(si|N)=3/5, P(si|P)=3/9
  - P(no|N)=2/5, P(no|P)=6/9
- Etc.

Suavizado

- Cuando se tienen pocos datos (o muchas variables-valores) se pueden tener probabilidades igual a cero, lo que ocasiona problema
- Para ello se pueden "suavizar" las estimaciones de las probabilidades
- Existen varios métodos de suavizado, el más sencillo y común el de *Laplace*

Suavizado

- El suavizado de Laplace consiste en inicializar todas las probabilidades en forma uniforme, y después *incrementarlas* con los datos
- Ejemplo:
  - P(A), A={a1, a2, a3)
  - Inicial: P(a1) = 1/3, P(a2) = 1/3, P(a3) = 1/3
  - Dato A=a2 → P(a1) = 1/4, P(a2) = 2/4, P(a3) = 1/4
  - Dato A=a3 → P(a1) = 1/5, P(a2) = 2/5, P(a3) = 2/5
  - Dato A=a3 → P(a1) = 1/6, P(a2) = 2/2, P(a3) = 3/6

Información incompleta

- En la práctica, en muchas ocasiones los datos no están completos
- Dos tipos básicos de información incompleta:
  - Faltan algunos valores de una de las variables en algunos casos – **datos incompletos**
  - Faltan todos los valores de una variable – **nodos ocultos**

Información incompleta

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	xxx	alta	--	N
soleado	alta	alta	--	N
nublado	alta	alta	--	P
lluvia	media	alta	--	P
lluvia	baja	normal	--	P
lluvia	baja	normal	--	N
nublado	baja	normal	--	P
soleado	media	alta	--	N
soleado	xxx	normal	--	P
lluvia	media	normal	--	P
soleado	media	normal	--	P
nublado	media	alta	--	P
nublado	alta	normal	--	P
lluvia	media	alta	--	N

Datos incompletos

Existen varias alternativas:

1. Considerar un nuevo valor “desconocido”
2. Tomar el valor más probable (promedio) de la variable
3. Considerar el valor más probable en base a las otras variables
4. Considerar la probabilidad de los diferentes valores en base a las otras variables

Datos incompletos

Valor más probable:

1. Asignar todas las variables observables.
2. Propagar su efecto y obtener las probabilidades posteriores de las no observables.
3. Para las variables no observables, asumir el valor con probabilidad mayor como observado.
4. Actualizar las probabilidades previas y condicionales de acuerdo a las fórmulas anteriores.
5. Repetir 1 a 4 para cada observación.

Datos incompletos

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	xxx	alta	--	N
soleado	alta	alta	--	N
nublado	alta	alta	--	P
lluvia	media	alta	--	P
lluvia	baja	normal	--	P
lluvia	baja	normal	--	N
nublado	baja	normal	--	P
soleado	media	alta	--	N
soleado	xxx	normal	--	P
lluvia	media	normal	--	P
soleado	media	normal	--	P
nublado	media	alta	--	P
nublado	alta	normal	--	P
lluvia	media	alta	--	N

P(T|sol,alta,N)

P(T|sol,nor,P)

Datos incompletos

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	media	alta	--	N
soleado	alta	alta	--	N
nublado	alta	alta	--	P
lluvia	media	alta	--	P
lluvia	baja	normal	--	P
lluvia	baja	normal	--	N
nublado	baja	normal	--	P
soleado	media	alta	--	N
soleado	media	normal	--	P
lluvia	media	normal	--	P
soleado	media	normal	--	P
nublado	media	alta	--	P
nublado	alta	normal	--	P
lluvia	media	alta	--	N

P(T|sol,alta,N)

P(T|sol,nor,P)

Nodos ocultos – algoritmo EM

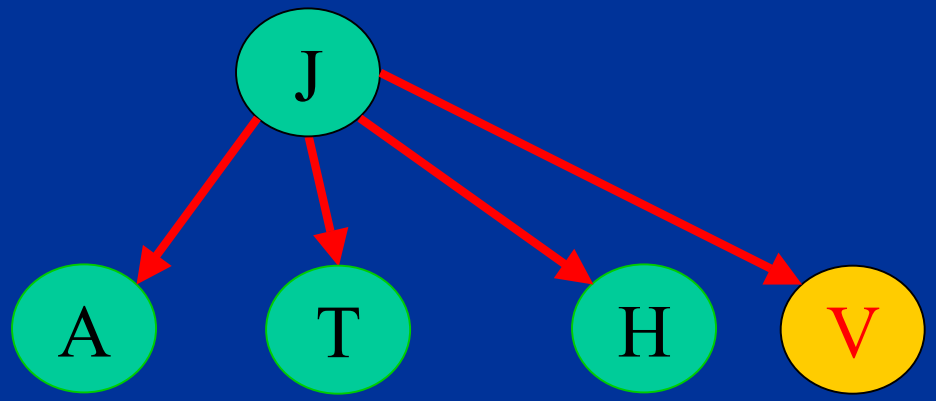
- El algoritmo EM es un método estadístico muy utilizado para estimar probabilidades cuando hay variables no observables (un caso especial es el algoritmo de Baum-Welch en HMM)
- Consiste básicamente de 2 pasos que se repiten en forma iterativa:

1. Paso E: se estiman los datos faltantes en base a los parámetros (P) actuales
2. Paso M: se estiman las probabilidades (parámetros) considerando los datos estimados

EM para RB con nodos ocultos

1. Iniciar los parámetros desconocidos (CPTs) con valores aleatorios (o estimaciones de expertos)
2. Utilizar los datos conocidos con los parámetros actuales para estimar los valores de la variable(s) oculta(s)
3. Utilizar los valores estimados para completar la tabla de datos
4. Re-estimar los parámetros con los nuevos datos
5. Repetir 2→4 hasta que no haya cambios significativos en las probabilidades

Ejemplo



- V es un nodo oculto
- Se seleccionan valores aleatorios para P(V|J)
- Se calcula la probabilidad de V para cada caso, dados los valores de A, T, H, J
- Cada caso se “pesa” de acuerdo a las probabilidades posteriores de V (un caso puede representar “n” datos)
- Se recalculan los parámetros (P(V|J)) en base a los casos obtenidos
- Se repite el proceso hasta que converja

EM: inicio

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	media	alta	--	N
soleado	alta	alta	--	N
nublado	alta	alta	--	P
lluvia	media	alta	--	P
lluvia	baja	normal	--	P
lluvia	baja	normal	--	N
nublado	baja	normal	--	P
soleado	media	alta	--	N
soleado	media	normal	--	P
lluvia	media	normal	--	P
soleado	media	normal	--	P
nublado	media	alta	--	P
nublado	alta	normal	--	P
lluvia	media	alta	--	N

"Adivinar"
P(V | J):

V\J	N	P
no	0.5	0.5
si	0.5	0.5

EM: paso E

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	media	alta	no	N
soleado	alta	alta	no	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	si	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	media	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	si	P
lluvia	media	alta	si	N

Estimar valores de V en base a P(V | J) y los datos

EM: paso M

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	media	alta	no	N
soleado	alta	alta	no	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	si	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	media	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	si	P
lluvia	media	alta	si	N

Re-estimar P(V | J) con los Nuevos datos:

V\J	N	P
no	0.6	0.44
si	0.4	0.66

EM

- Limitaciones:
  - Puede caer en máximos locales (depende del valor inicial)
  - Complejidad computacional

Aprendizaje Estructural

Diversos métodos:

- Aprendizaje de árboles
- Aprendizaje de poliárboles
- Aprendizaje de redes multiconectadas
  - Métodos basados en medidas
  - Métodos basados en relaciones de dependencia

Aprendizaje de árboles

- Algoritmo desarrollado por Chow y Liu para aproximar una distribución de probabilidad por un producto de probabilidades de segundo orden (árbol).
- La probabilidad conjunta de n variables se puede representar como:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{j(i)})$$

- donde $X_{j(i)}$ es la causa o padre de X_i.

Aprendizaje de árboles

- Se plantea el problema como uno de optimización - obtener la estructura que más se aproxime a la distribución "real".
- Medida de la diferencia de información entre la distribución real (P) y la aproximada (P*):

$$I(P,P^*)=\sum_x P(X)\log\frac{P(X)}{P^*(X)}$$

- El objetivo es minimizar I.

Aprendizaje de árboles

- Se puede definir dicha diferencia en función de la información mutua entre pares de variables, que se define como:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(X_i, X_j) \log \frac{P(X_i, X_j)}{P(X_i)P(X_j)}$$

- Se puede demostrar (Chow 68) que la diferencia de información es una función del negativo de la suma de las informaciones mutuas (pesos) de todos los pares de variables que constituyen el árbol
- Encontrar el árbol más próximo equivale a encontrar el árbol con mayor peso.

Aprendizaje de árboles - algoritmo

1. Calcular la información mutua entre todos los pares de variables (n(n - 1)/2).
2. Ordenar las informaciones mutuas de mayor a menor.
3. Seleccionar la rama de mayor valor como árbol inicial.
4. Agregar la siguiente rama mientras no forme un ciclo, si es así, desechar.
5. Repetir (3-4) hasta que se cubran todas las variables (n -1 ramas).

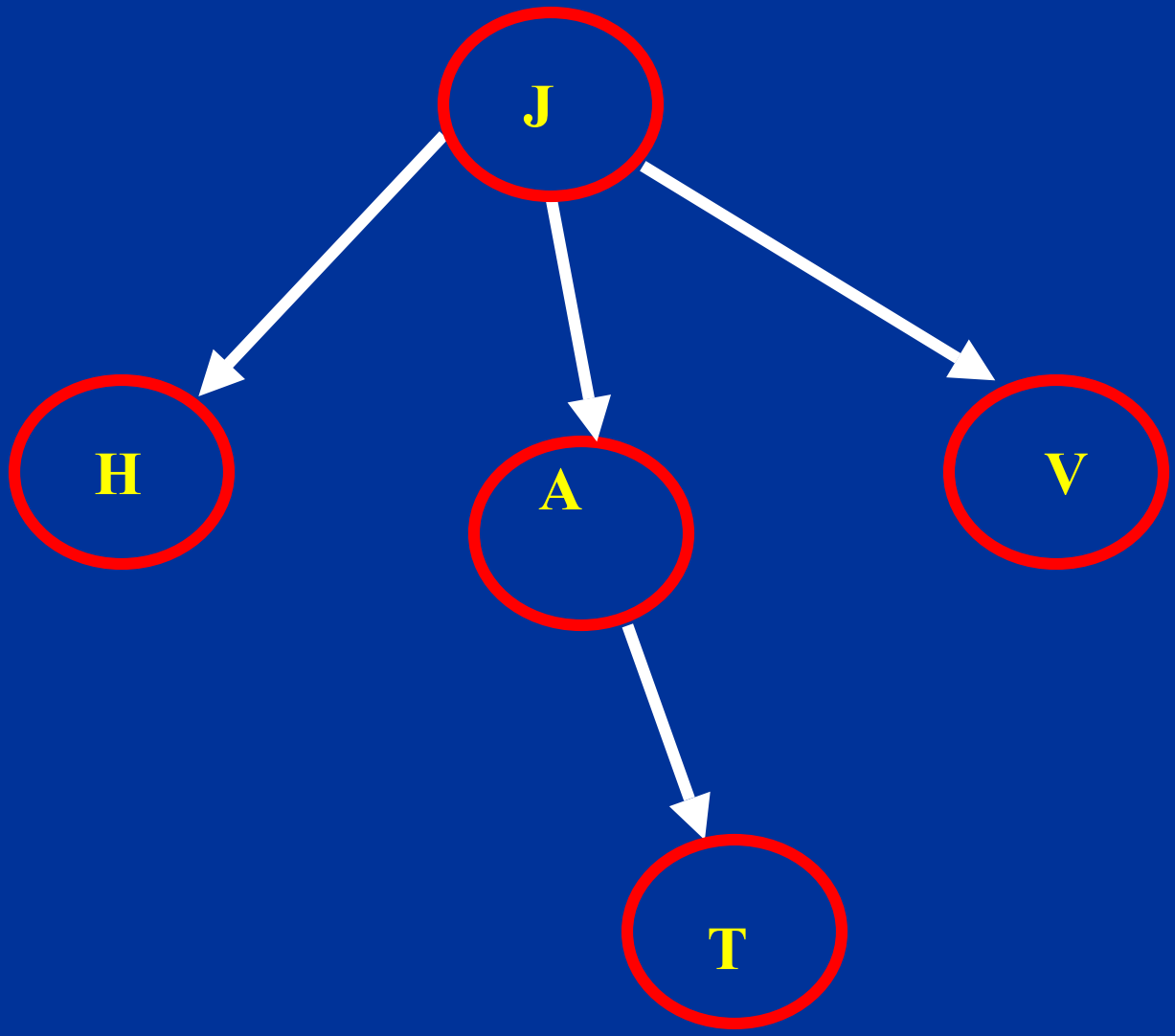
- El algoritmo NO provee la dirección de los arcos, por lo que ésta se puede asignar en forma arbitraria o utilizando semántica externa (experto).

Ejemplo (golf)

- Informaciones mutuas ordenadas

No.	Var 1	Var 2	I.M.
1	temp.	ambiente	.2856
2	juega	ambiente	.0743
3	juega	humedad	.0456
4	juega	viento	.0074
5	humedad	ambiente	.0060
6	viento	temp.	.0052
7	viento	ambiente	.0017
8	juega	temp.	.0003
9	humedad	temp.	0
10	viento	humedad	0

Ejemplo (golf)



Aprendizaje de poliárboles

- Parte del esqueleto (estructura sin direcciones) obtenido con el algoritmo anterior
- Determina la dirección de los arcos utilizando pruebas de dependencia entre tripletas de variables.
- Dadas 3 variables, existen 3 casos posibles:
  - Arcos divergentes
  - Arcos secuenciales
  - Arcos convergentes
- Los primeros dos casos son indistinguibles, pero el tercero es diferente, ya que las dos variables "padre" son marginalmente independientes.

Prueba de Tripletas

- Tripleta de variables:

$$X - Z - Y$$

- Si X – Y son independientes dado Z, entonces pueden ser secuenciales o divergentes

$$X \rightarrow Z \rightarrow Y;\ X \leftarrow Z \rightarrow Y$$

- Si X – Y no son independientes dado Z, entonces son arcos convergentes

$$X \rightarrow Z \leftarrow Y$$

Aprendizaje de poliárboles - algoritmo

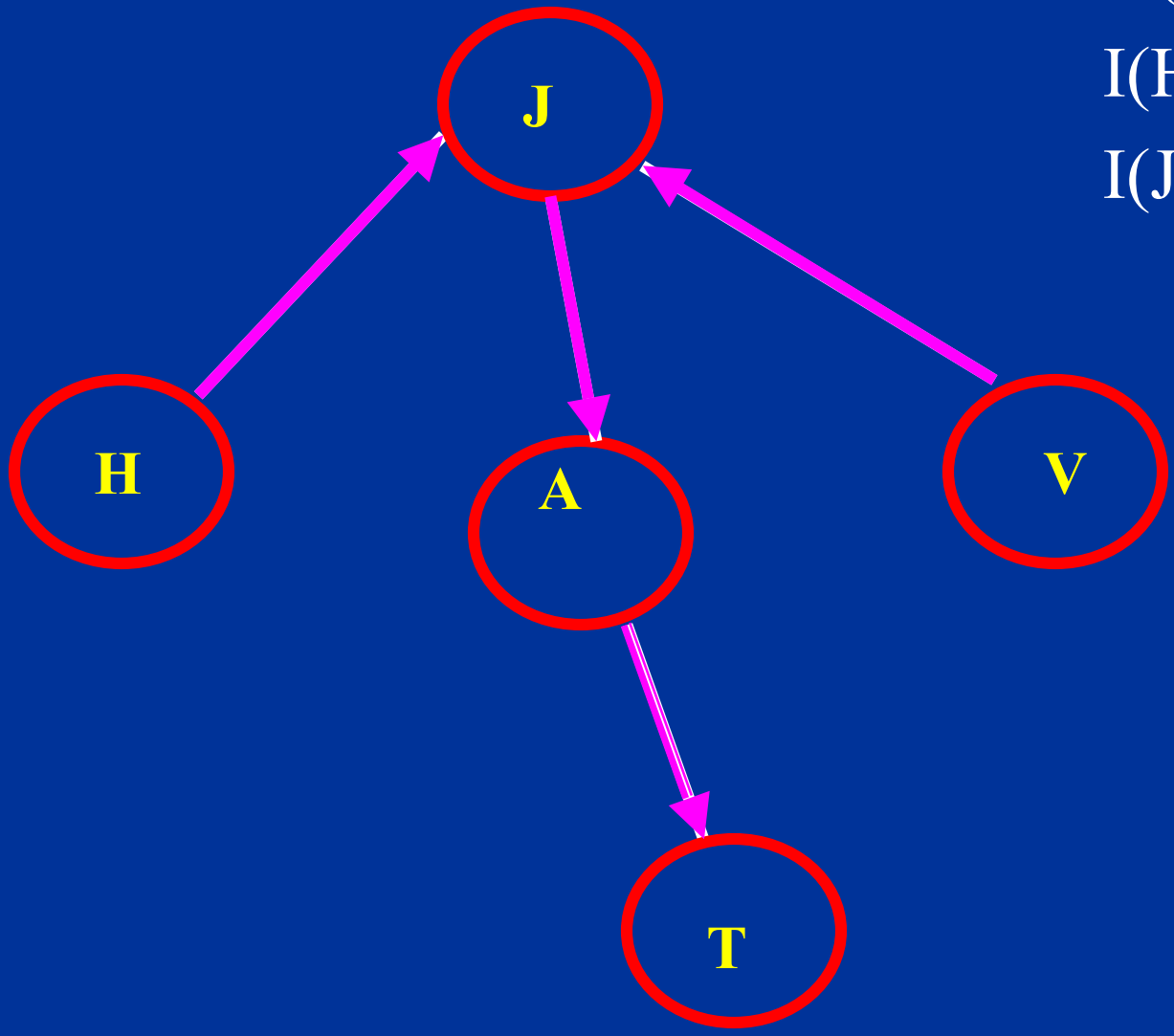
1. Obtener esqueleto utilizando el algoritmo de Chow y Liu
2. Recorrer la red hasta encontrar una tripleta de nodos que sean convergentes (tercer caso) - nodo multipadre-
3. A partir de un nodo multipadre determinar las direcciones de los arcos utilizando la prueba de tripletas hasta donde sea posible (base causal).
4. Repetir 2-3 hasta que ya no se puedan descubrir más direcciones.
5. Si quedan arcos sin direccionar, utilizar semántica externa para obtener su dirección (o fijar direcciones).

Ejemplo

~I(H,J,V)

I(H,J,A)

I(J,A,T)



Aprendizaje de redes multiconectadas

Existen dos tipos de métodos para el aprendizaje genérico de redes bayesianas:

1. Métodos basados en medidas de ajuste y búsqueda
2. Métodos basados en pruebas de independencia

Métodos basados en medidas

Se generan diferentes estructuras y se evalúan respecto a los datos utilizando alguna medida

Dos aspectos principales:

- Medida de "ajuste" de la estructura a los datos
- Búsqueda de la "mejor" estructura

Medidas

- Evalúan que tan “buena” es una estructura respecto a los datos
- Hay varias posibles medidas, las dos más comunes son:
  - Medida bayesiana
  - Medida basada en el principio de longitud de descripción mínima (MDL)

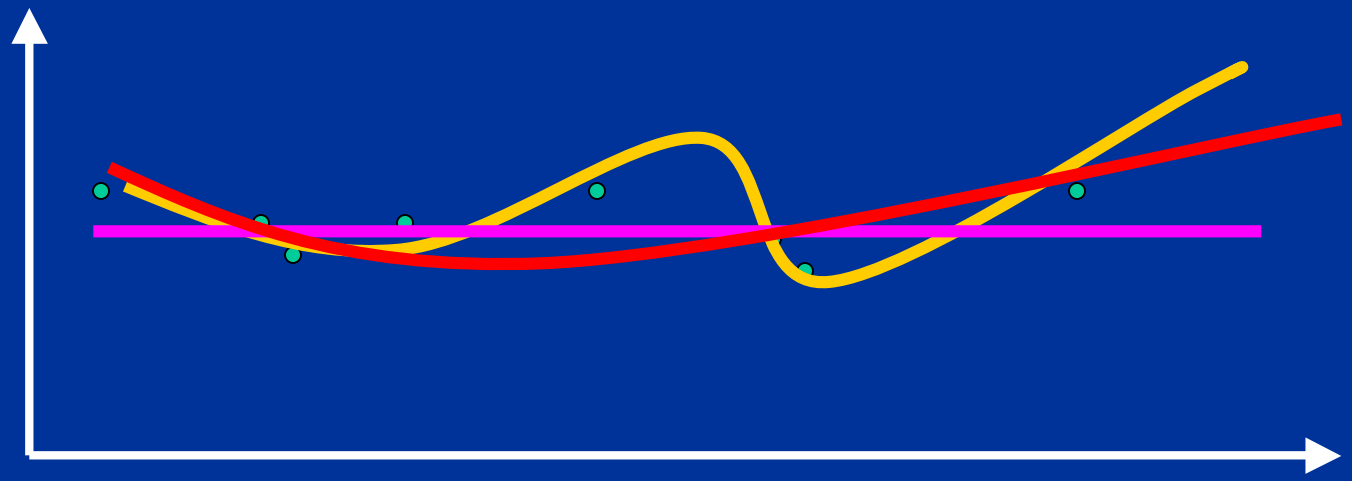
MDL

- La “calidad” de la estructura se basa en el principio de “descripción de longitud mínima” (MDL):
  - Tamaño de la descripción de la red (complejidad)
  - Tamaño de error de predicción de los datos por la red (exactitud)
- Se hace una búsqueda heurística de la estructura en base al MDL

MDL

Compromiso entre exactitud y complejidad-minimizar: ***long. de descripción del modelo +***
descripción de lo datos dado el modelo

Ejemplo – ajustar un polinomio a un conjunto de puntos:



MDL

Para redes bayesianas:

Complejidad:

$$L = \sum_i [\, k_i \log_2 n + d(S_i - 1) \prod_{Fi} s_i]$$

n-# de nodos, k-# padres por nodo, Si-# de valores por variable, Fi-conj. de padres, d-# de bits

Exactitud:

$$w(x_i, Fx_i) = \sum P(x_i, Fx_i) \log_2 [P(x_i,Fx_i)/P(x_i)P(Fx_i)]$$

$$W = \sum_i w(x_i, Fx_i)$$

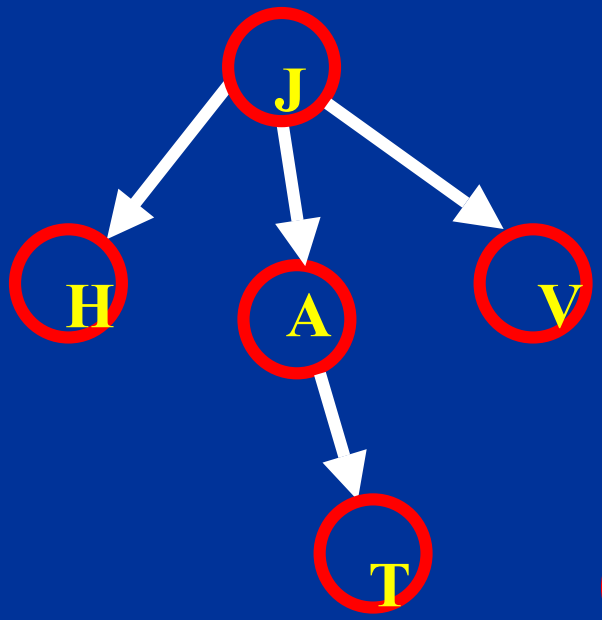
Buscando la mejor estructura

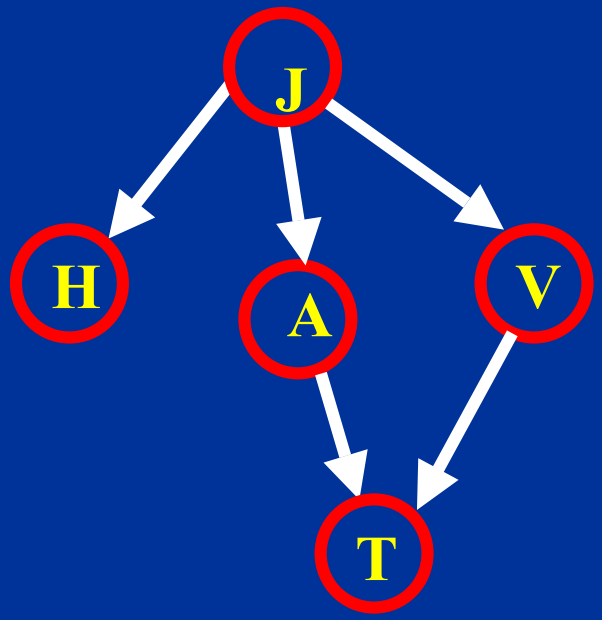
- Búsqueda de ascenso de colinas (*hill climbing*)
- Se inicia con una estructura simple (árbol) y se van agregando arcos hasta llegar a un *mínimo local*

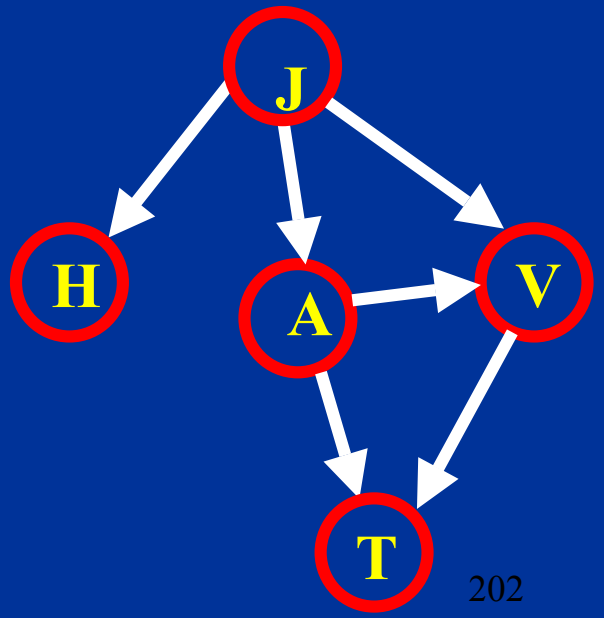
Ejemplo - ¿Cuándo jugar golf?

Ambiente	Temp.	Humedad	Viento	Jugar
soleado	alta	alta	no	N
soleado	alta	alta	si	N
nublado	alta	alta	no	P
lluvia	media	alta	no	P
lluvia	baja	normal	no	P
lluvia	baja	normal	si	N
nublado	baja	normal	si	P
soleado	media	alta	no	N
soleado	baja	normal	no	P
lluvia	media	normal	no	P
soleado	media	normal	si	P
nublado	media	alta	si	P
nublado	alta	normal	no	P
lluvia	media	alta	si	N

Ejemplo





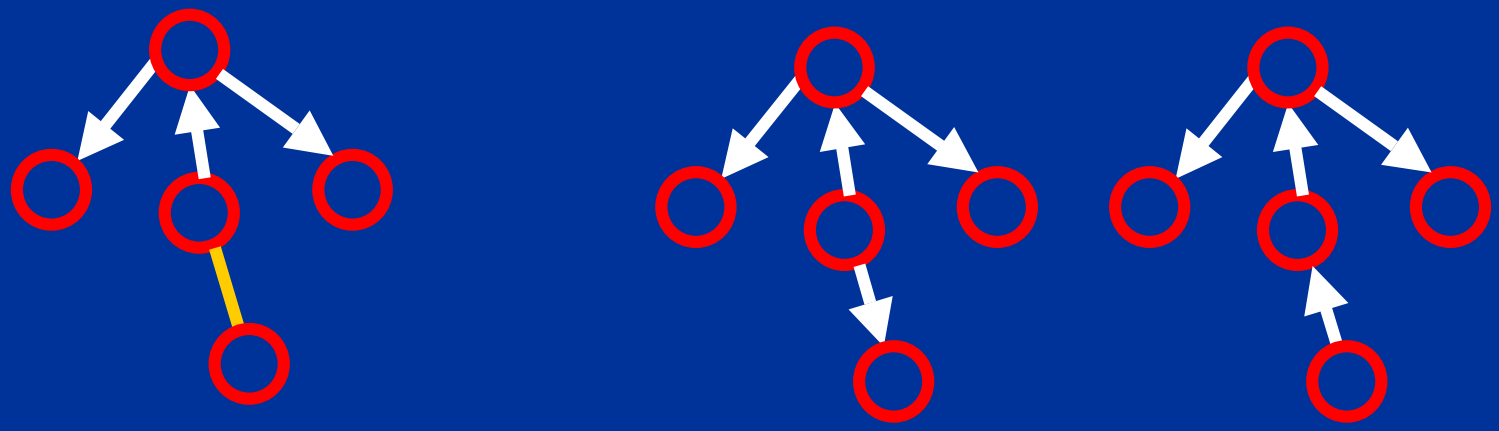


Variantes

- Utilizar otros métodos de búsqueda:
  - Algoritmos genéticos
  - “Beam search”
  - Etc.
- Considerar sólo estructuras que sean diferentes estadísticamente, buscando sobre estructuras equivalentes (se llega a una estructura parcial)

Estructuras Equivalentes

- Cuando ciertos arcos no se pueden determinar por pruebas estadísticas, por ejemplo:



Métodos basados en medidas

- Se genera la estructura en base a ir agregando/eliminando arcos de acuerdo a medidas de dependencia entre variables
- Ejemplos:
  - Árboles – método de Chow y Liu
  - Poliárboles – método de Rebane y Pearl
  - Multiconectadas – existen varios algoritmos basados en diferentes medidas

Algoritmo PC

- Se basa en pruebas de independencia entre variables:

$$I\,(Xi, Xj \mid A)$$

- Donde A es un subconjunto de variables
- Asume que:
  - Se tienen suficientes datos
  - Las pruebas estadísticas no tienen errores

Prueba de Independencia

- Para probar si X, Y son independientes dado A se utiliza la entropía cruzada condicional:

$$CE(X,Y \mid Z) = \Sigma_z P(z) \Sigma_{x,y} P(x,y|z) \log [P(x,y|z) / P(x|z) P(y|z)]$$

- Si es cero o cercana a cero, quiere decir que son independientes (se puede usar un umbral o una prueba estadística con cierto nivel de significancia)

Algoritmo

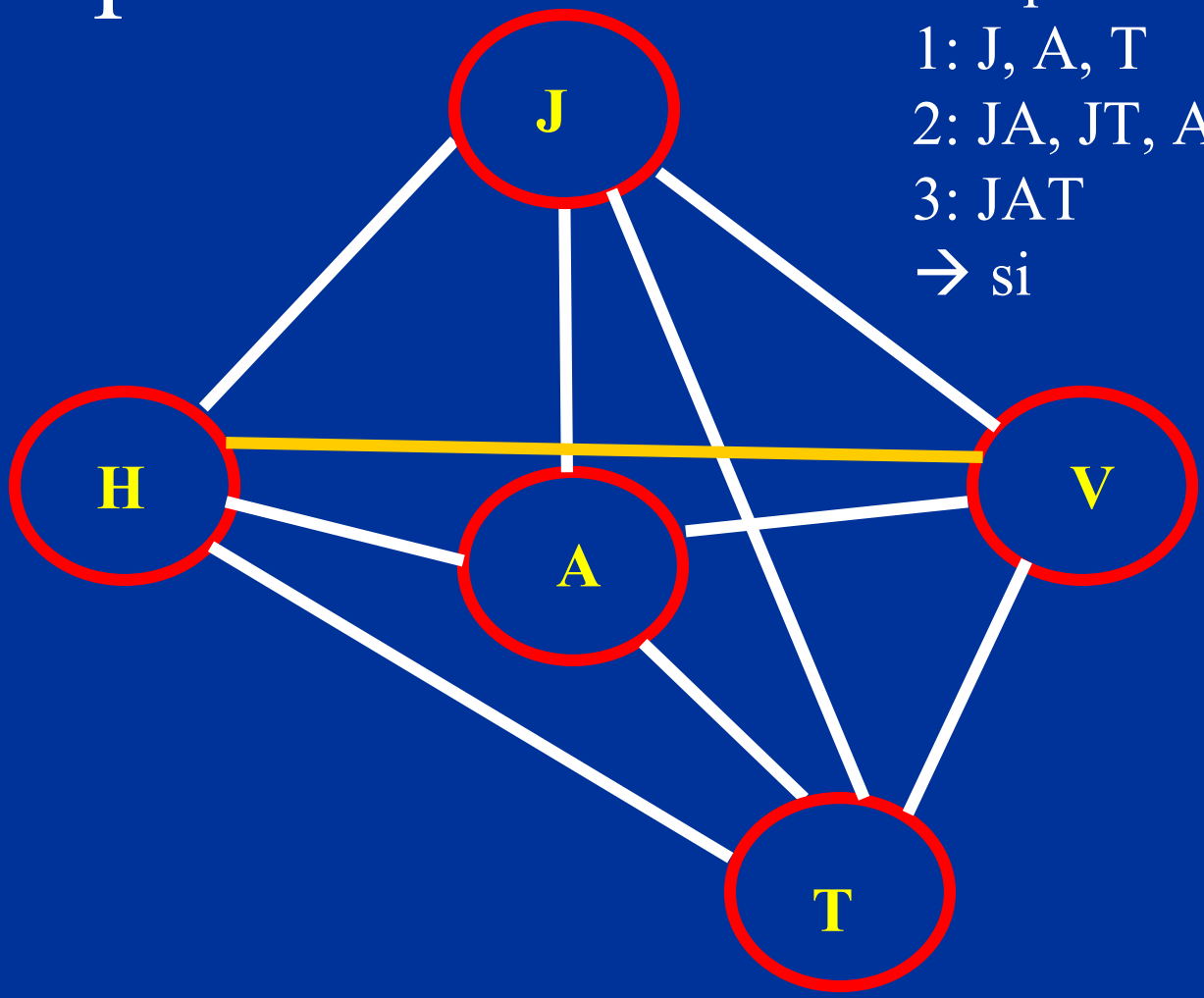
1. Encontrar un "esqueleto" (grafo no dirigido)
2. Encontrar arcos convergentes en tripletas de variables por pruebas de independencia
3. Orientar el resto de las ligas de forma que no se produzcan ciclos

Esqueleto

- La idea básica para determinar el esqueleto es iniciar con un grafo completo (conectando todos vs. todos los nodos) y eliminar el arco entre X – Y si hay un subconjunto de nodos en G (excepto X, Y) que los hace independientes
- En principio se consideran todos los posibles subconjuntos de variables, de tamaño 1 hasta de tamaño N-1 (N es el número de nodos adyacentes a X)
- El considerar todos los posibles subconjuntos es muy ineficiente, y normalmente se limita a considerar sólo subconjuntos de 1, 2, …, k nodos

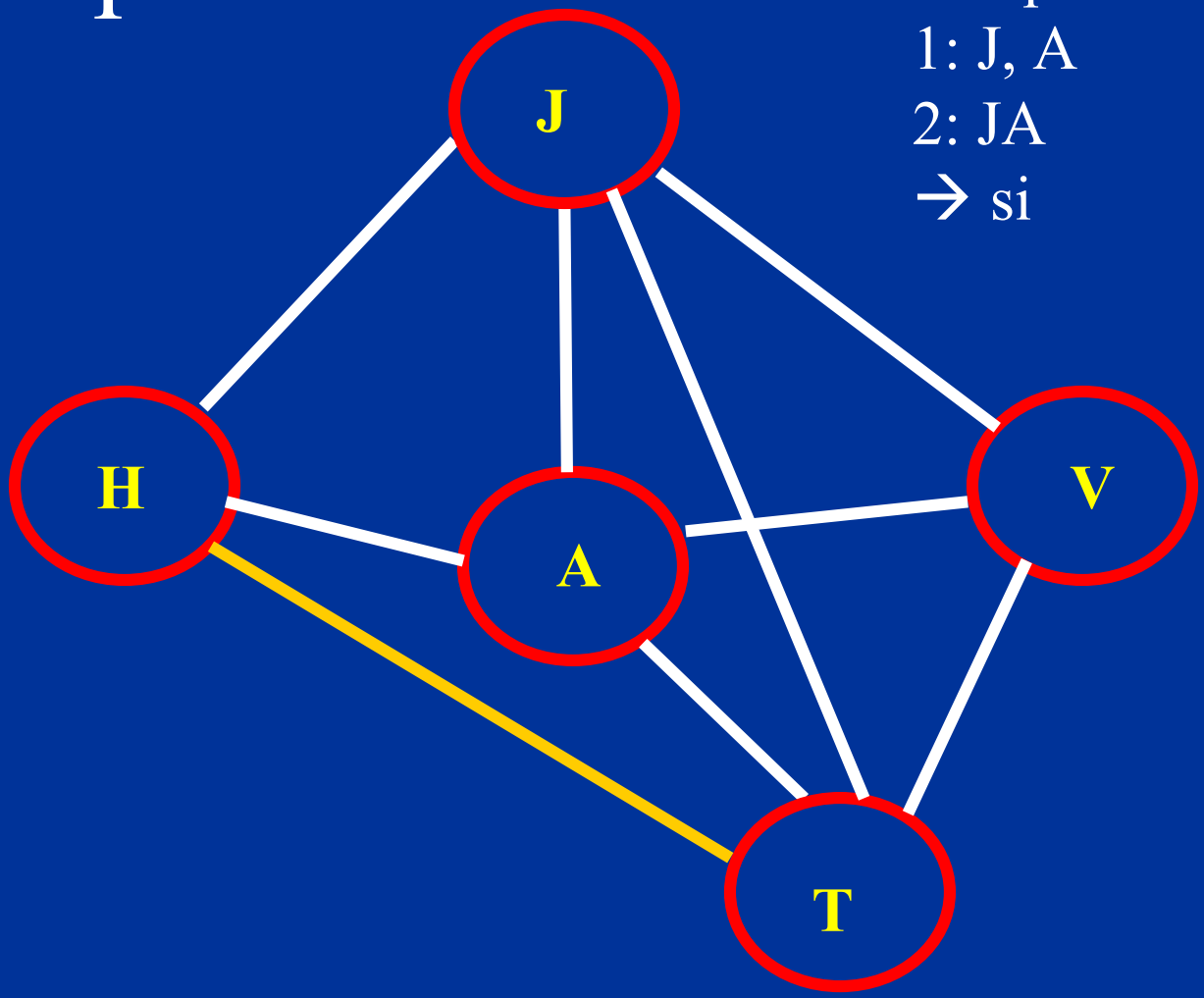
Ejemplo

Probar si H,V son Independientes dados:
1: J, A, T
2: JA, JT, AT
3: JAT
→ si



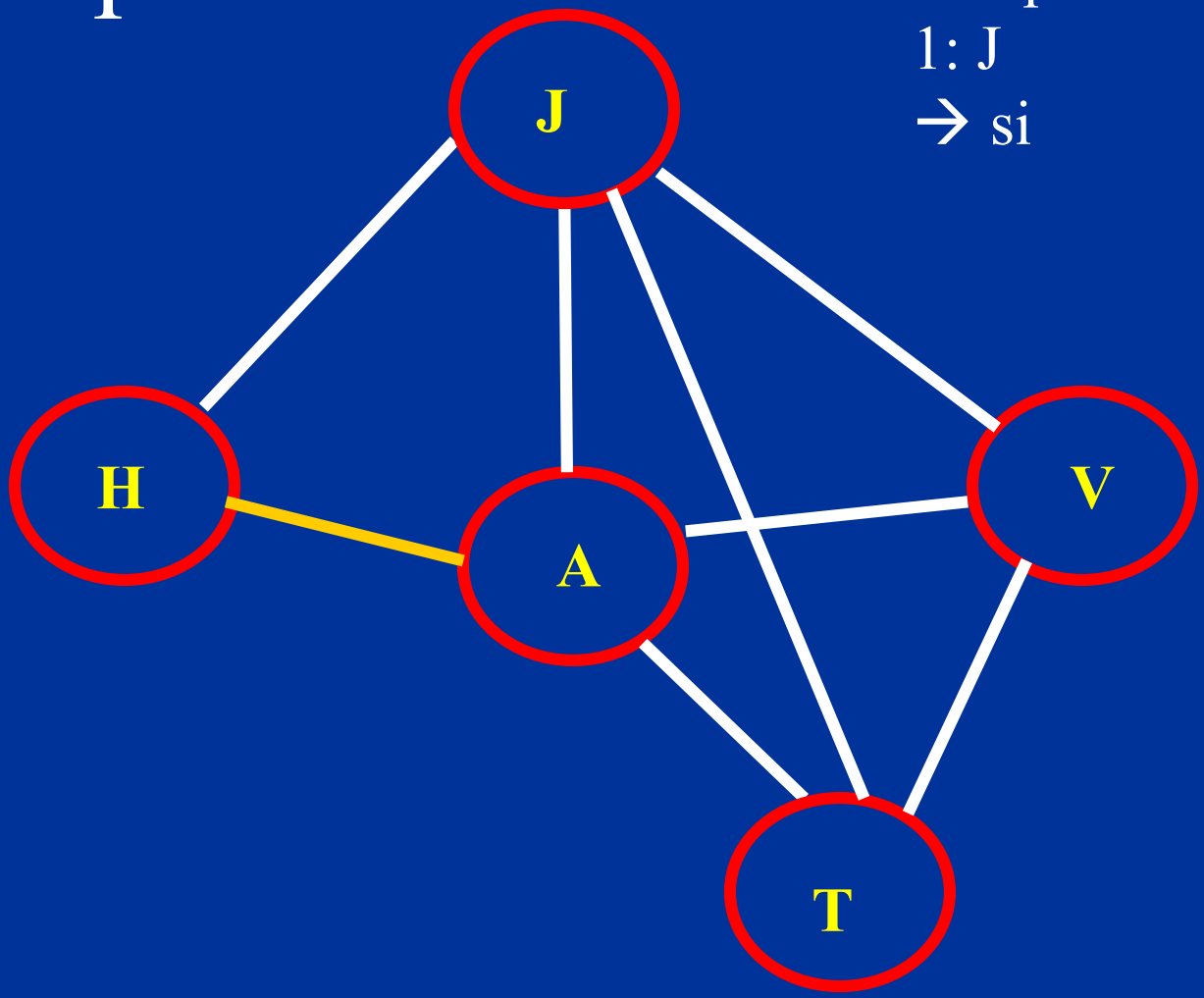
Ejemplo

Probar si H,T son Independientes dados:
1: J, A
2: JA
→ si



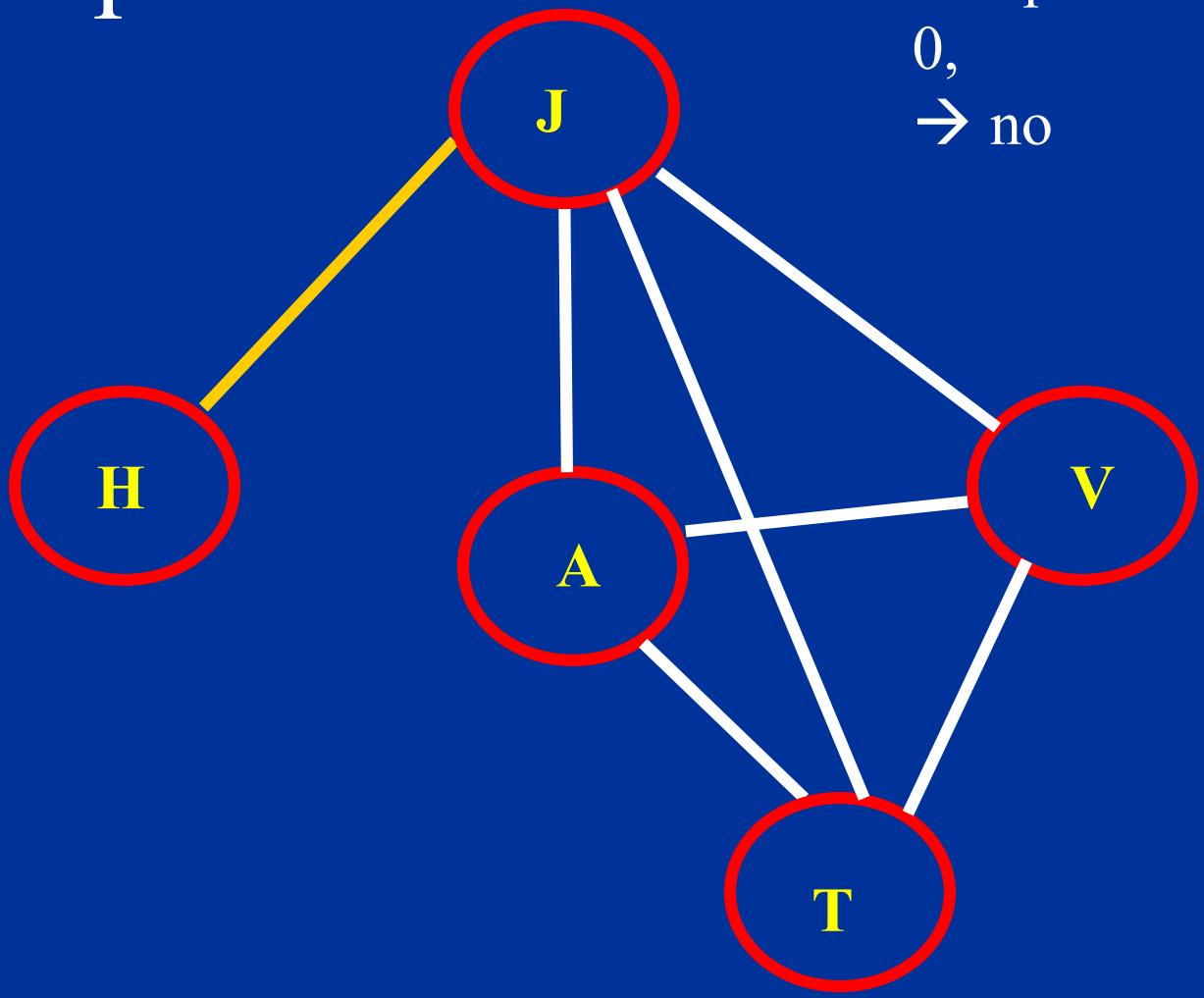
Ejemplo

Probar si H,A son Independientes dados:
1: J
→ si



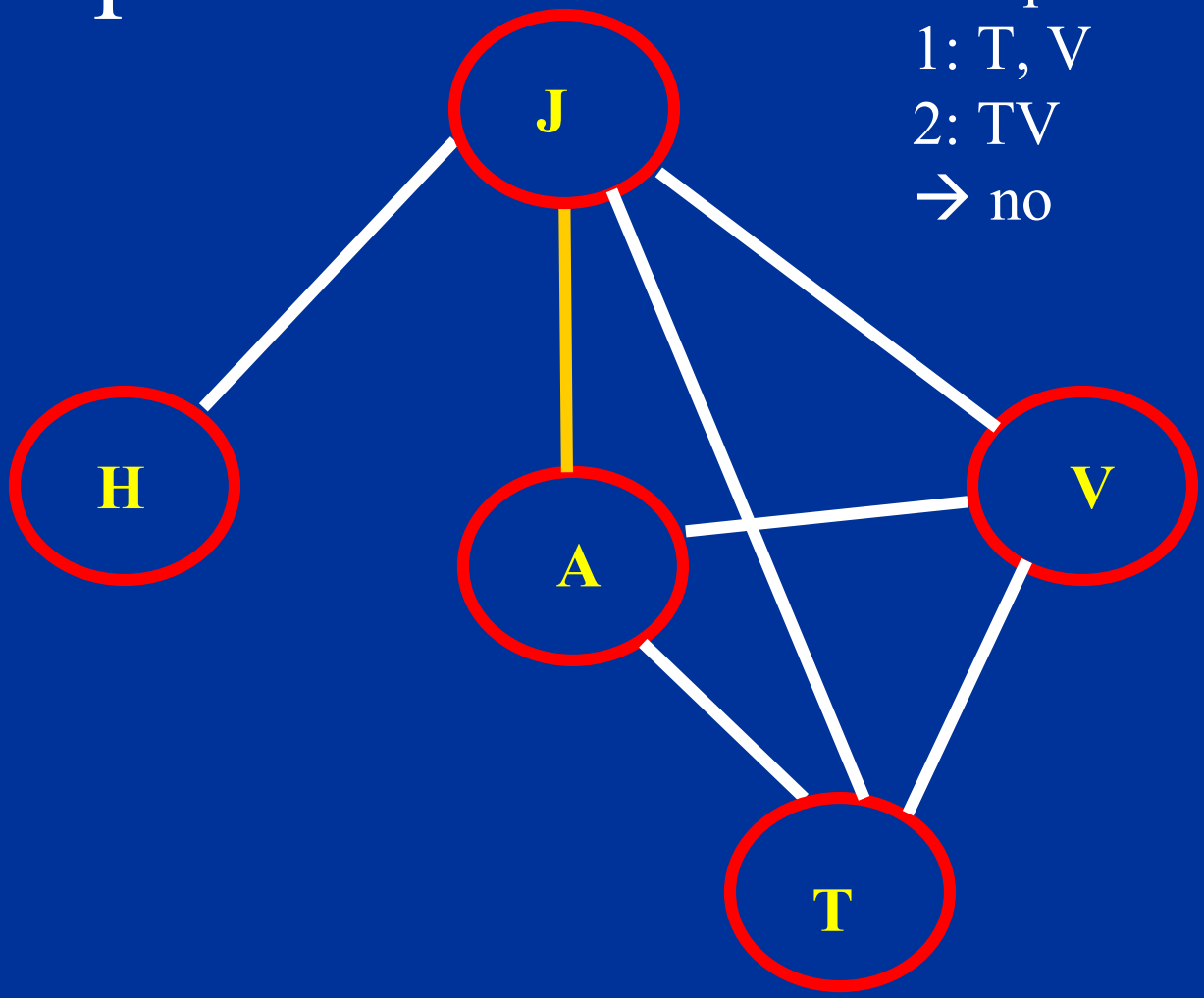
Ejemplo

Probar si H,J son Independientes dados: 0,
→ no



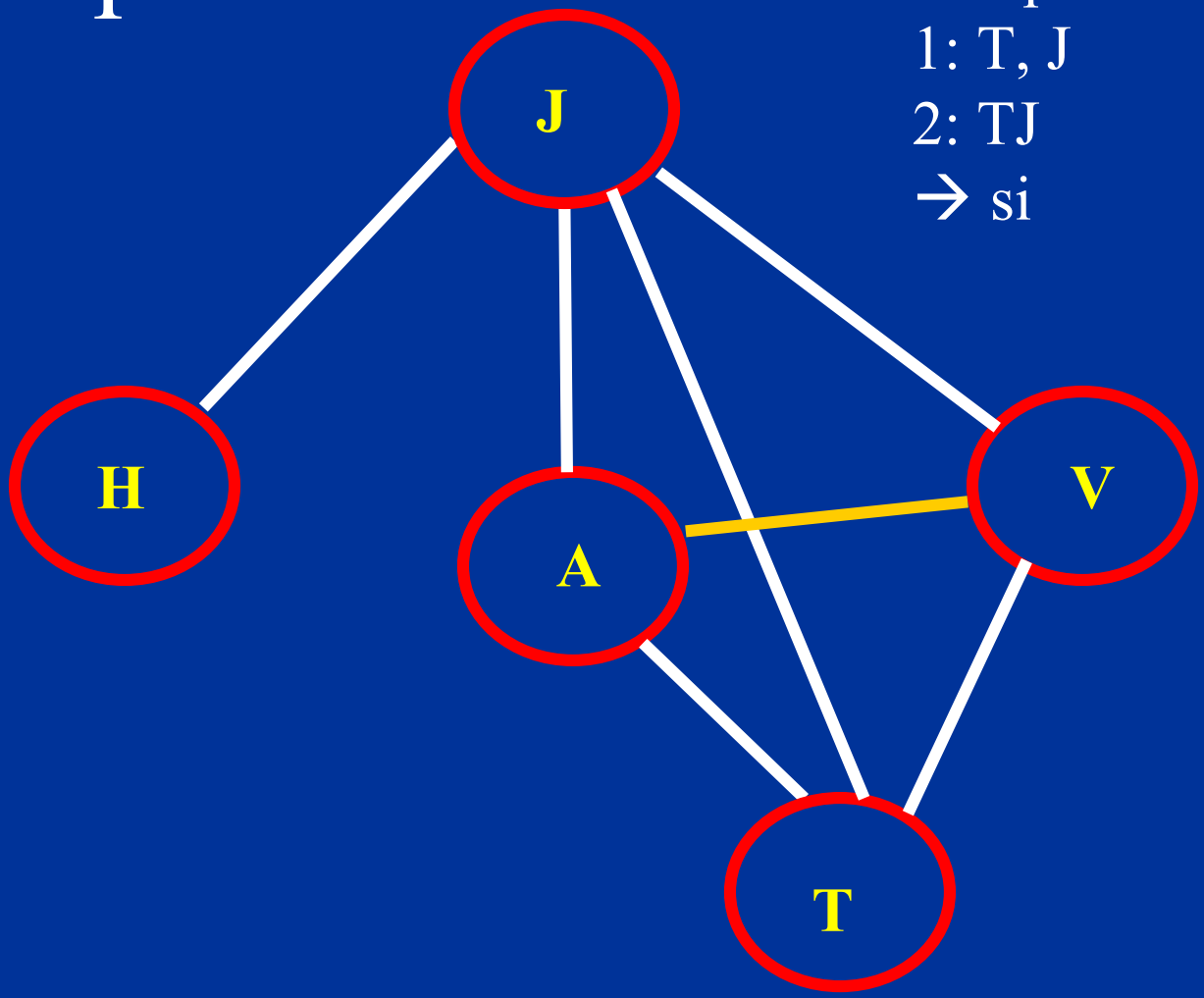
Ejemplo

Probar si A,J son Independientes dados:
1: T, V
2: TV
→ no



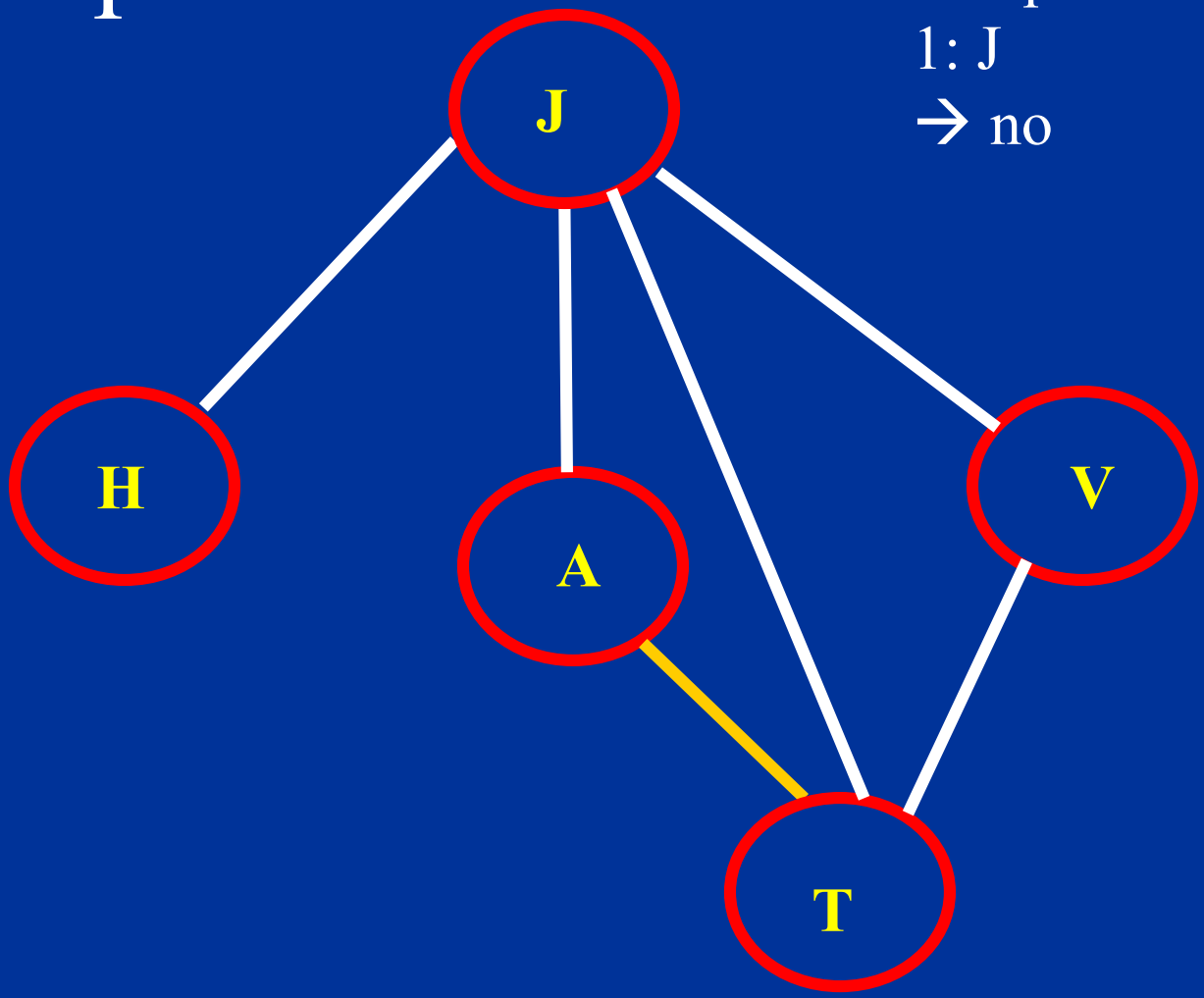
Ejemplo

Probar si A,V son Independientes dados:
1: T, J
2: TJ
→ si



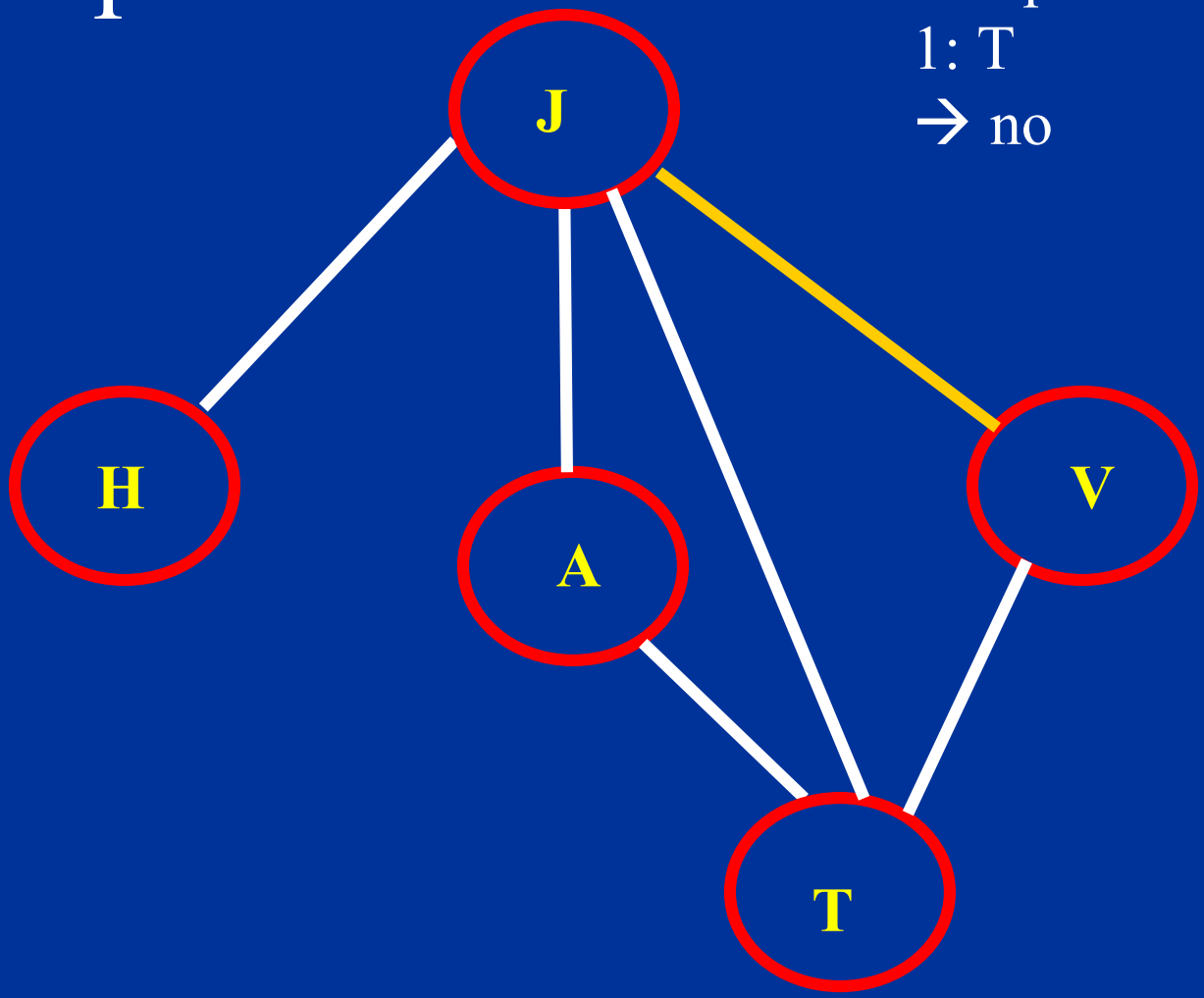
Ejemplo

Probar si A,T son Independientes dados:
1: J
→ no



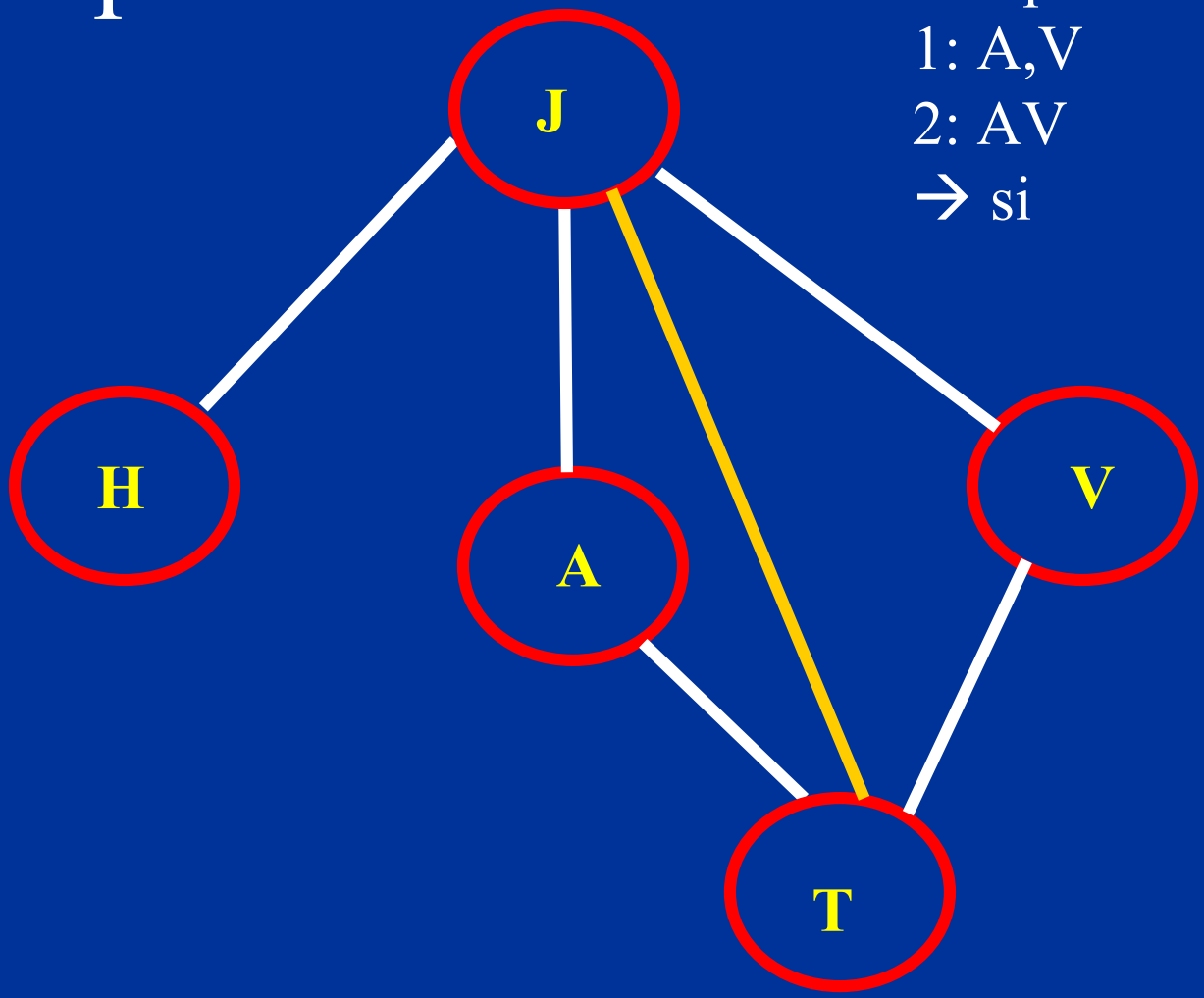
Ejemplo

Probar si J,V son Independientes dados:
1: T
→ no



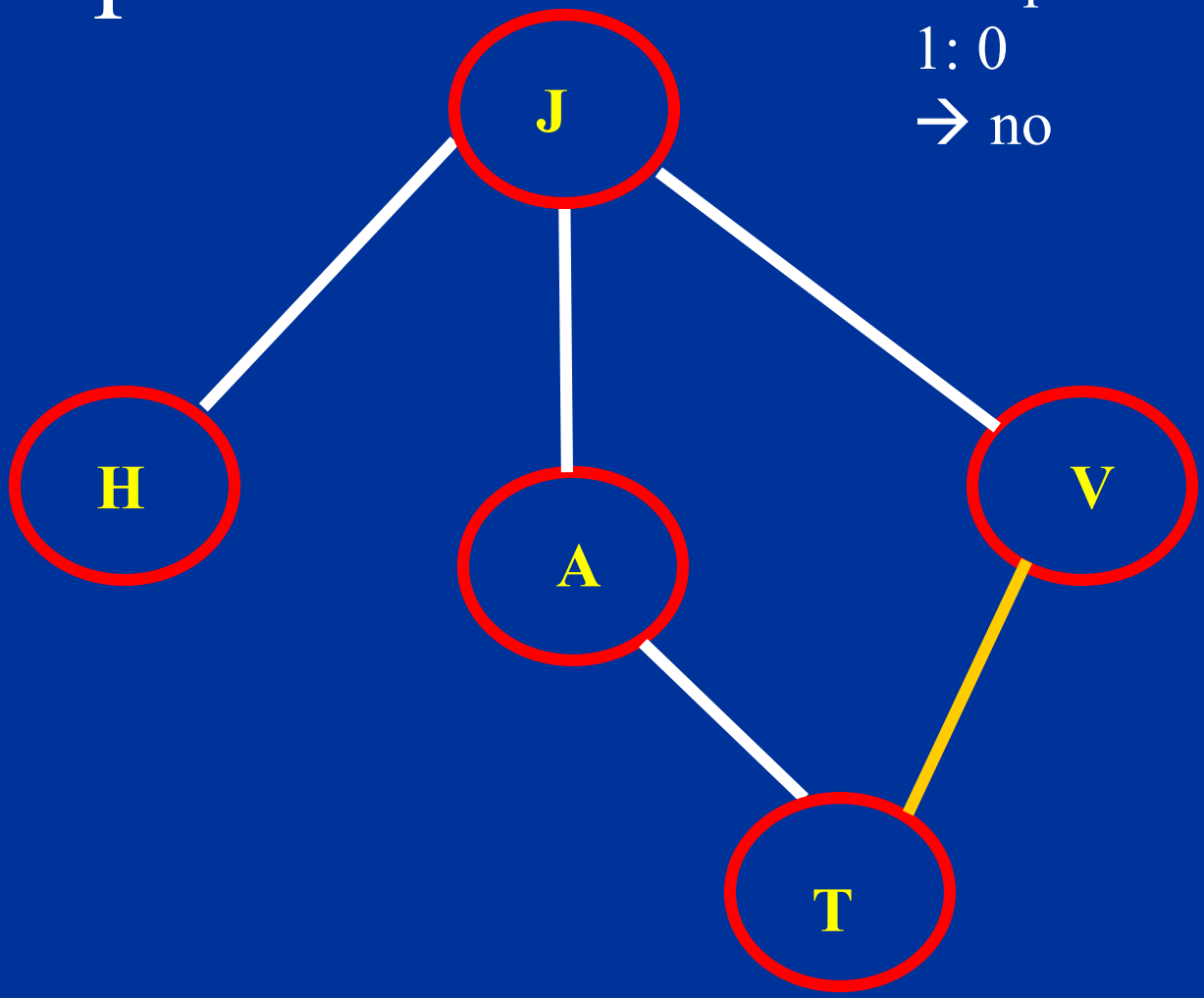
Ejemplo

Probar si J,T son Independientes dados:
1: A,V
2: AV
→ si



Ejemplo

Probar si V,T son Independientes dados:
1: 0
→ no



Arcos convergentes

- Se verifica cada tripleta de variables para encontrar arcos convergentes mediante pruebas de independencia:

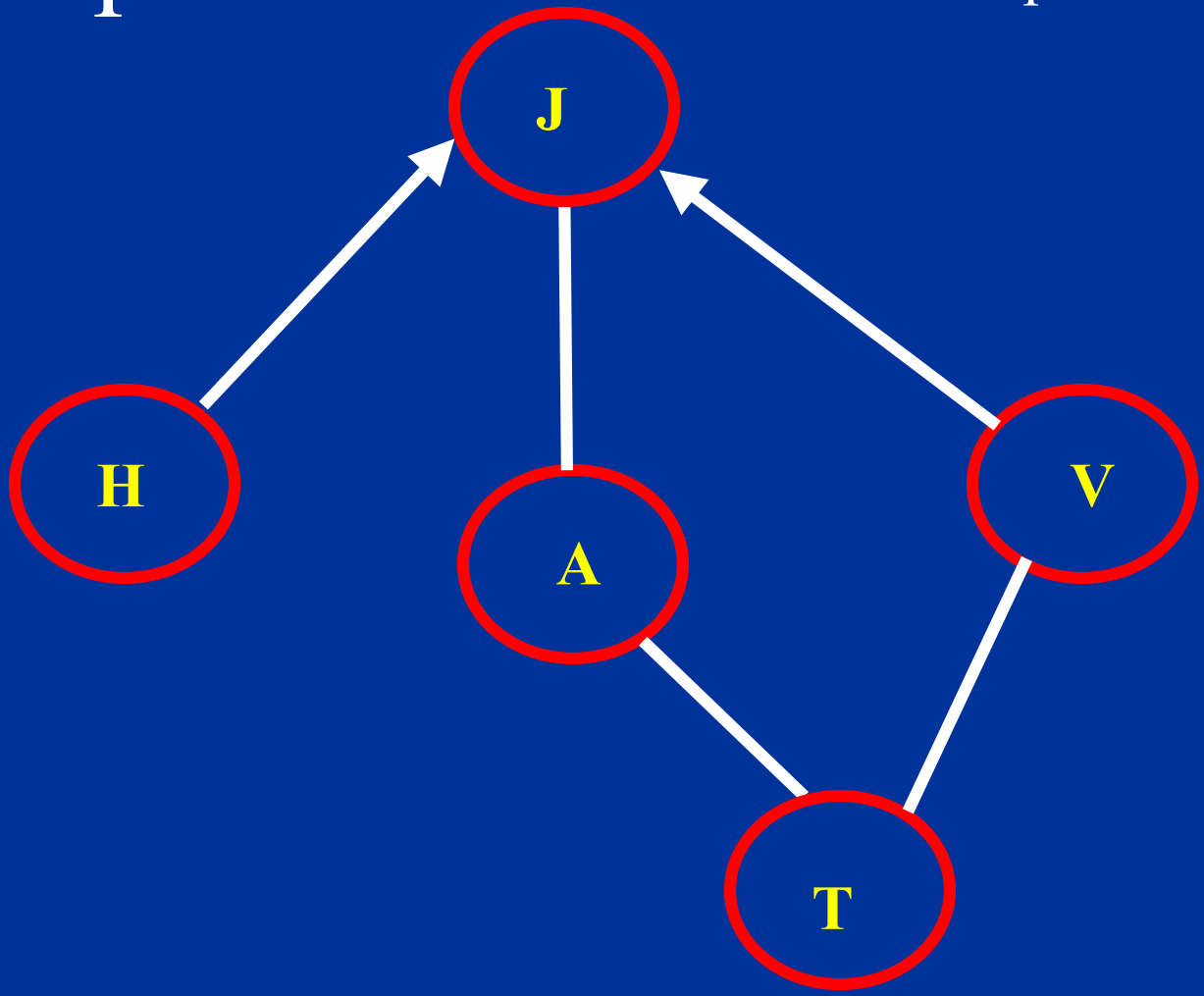
$$X - Z - Y$$

- Si X – Y no son independientes dado Z, entonces son arcos convergentes

$$X \rightarrow Z \leftarrow Y$$

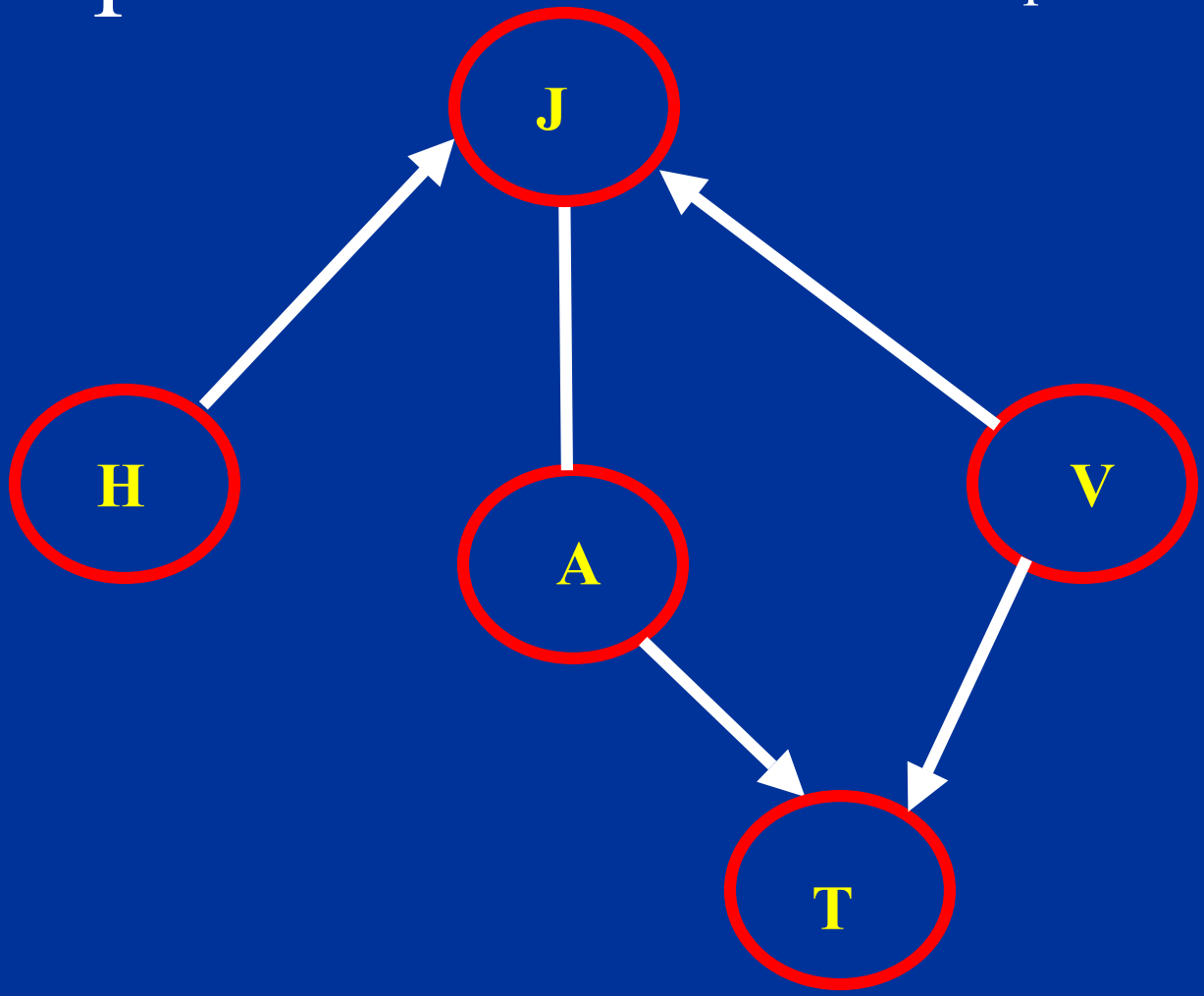
Ejemplo

H,V no son
Independientes dado J



Ejemplo

A,V no son
Independientes dado T

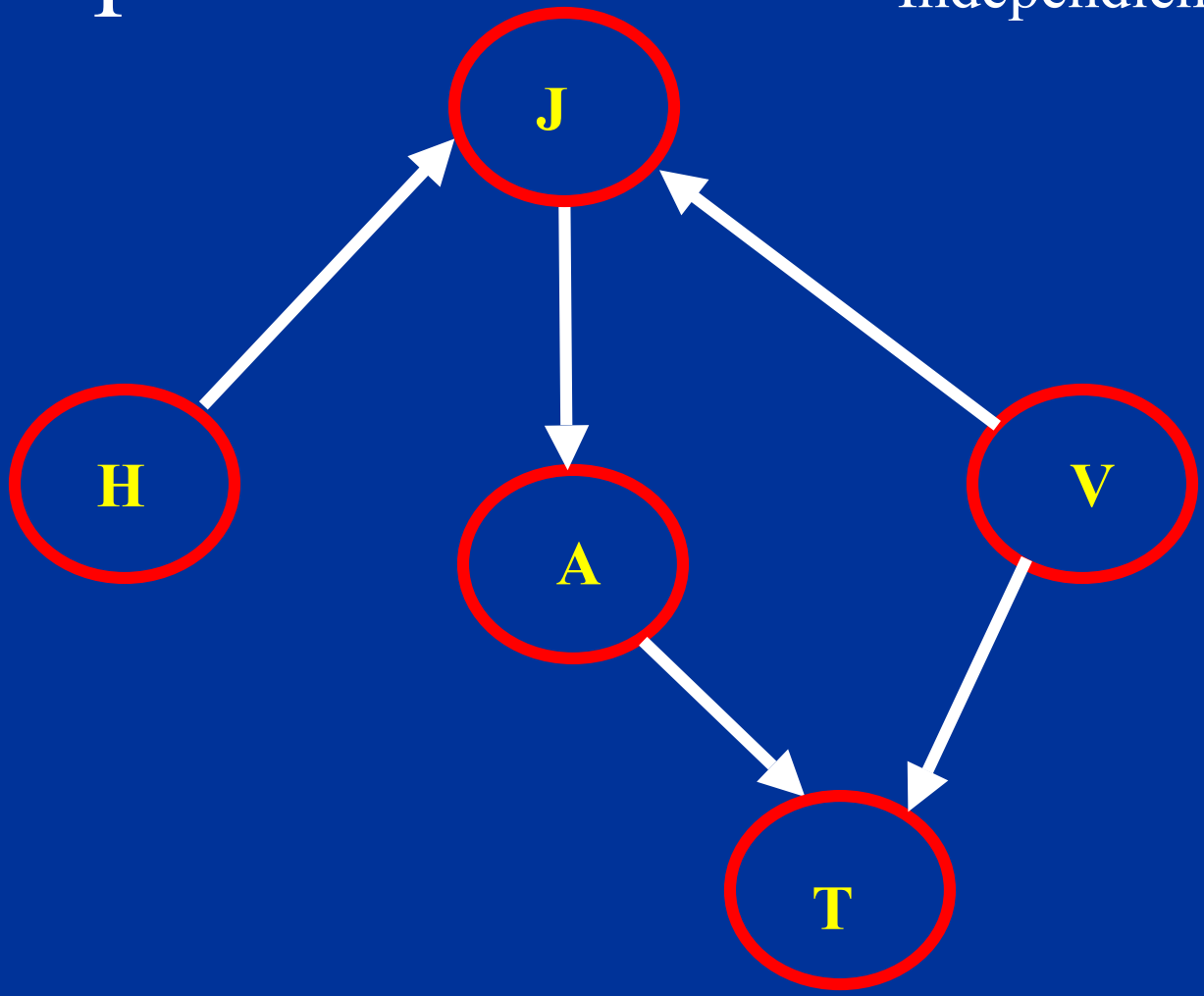


Otras orientaciones

- En base a los arcos existentes, se orientan los demás con pruebas de independencia, evitando crear ciclos
- Si quedan al final arcos sin orientar, se direccionan en forma aleatoria, evitando ciclos

Ejemplo

H, A son
Independientes dado J



HUGIN

Aprendizaje de RB

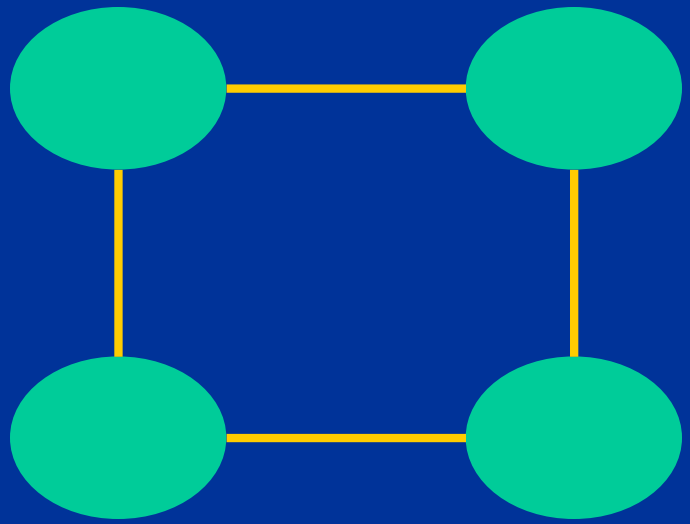
Referencias

- Pearl 88 – Cap. 8
- Neapolitan 90 – Cap. 10
- Darwiche 09 - Caps. 17,18
- T. Mitchell, Machine Learning, McGraw-Hill, 1997 – Cap. 6

Referencias

- W. Lam, F. Bacchus, "Learning Bayesian Belief Networks: An Approach based on the MDL Princlple", Computational Intelligence, Vol. 10 (1994) 269-293.
- G. Cooper, E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data”, Machine Learning, Vol 9, 1992.
- G. Cooper, E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data”, Machine Learning, Vol 9, 1992.
- L. E. Sucar, D. F. Gillies, D. A. Gillies, "Objective Probabilities in Expert Systems", Artificial Intelligence Journal, Vol. 61 (1993) 187-208.
- W. Buntine, “A guide to the literature on learning probabilistic networks form data”, IEEE TKDE.

I.6: Campos de Markov



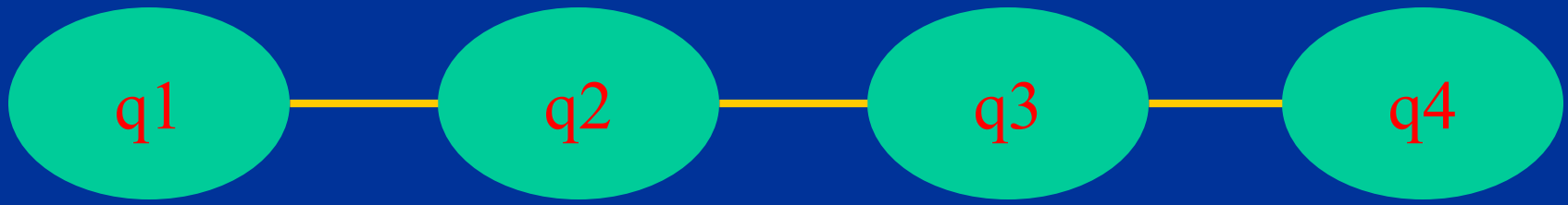
Campos de Markov (CAM)

- Ciertos procesos, como un campo magnético o una imagen, se pueden ver como una serie de estados en el espacio o una malla de estados.
- A este tipo de modelos se les conoce como *Campos de Markov.*
- Estos procesos son una extensión de las cadenas de Markov en los cuales el índice de tiempo se substituye por un índice espacial

Modelo de Ising

- Surgen del problema de modelar materiales ferromagnéticos en lo que se conoce como el Modelo *Ising*.
- Se tiene una serie de dipolos en una línea que pueden estar orientados hacia “arriba” (+) o hacia “abajo” (-).
- El estado de cada dipolo se ve influenciado por los dipolos cercanos - probabilidad para cada estado depende de los estado de los puntos vecinos.

Modelo de Ising



Posibles configuraciones:

+ + + +

+ + + -

+ + - +

....

Modelo de Ising

- Un campo de Markov asigna probabilidad a cada configuración en el espacio de posibles configuraciones.
- Se considera que la probabilidad del estado de una variable es independiente de los demás dados sus 2 vecinos (para una cadena), es decir que tiene la propiedad Markoviana

$$P(S_n = q_i \mid S_1 = q_j, S_2 = q_{k....}) = P(S_n = q_i \mid S_{n-1} = q_j, S_{n+1} = q_k)$$

Configuración más probable

- Dadas las probabilidades locales, el problema central en es encontrar la probabilidad de cada una las posibles configuraciones, y en particular cual es la configuración más probable.
  - + + + +
  - + + + -
  - + + - +
  - ...
  - - - + +
  - ...
  - - - - -

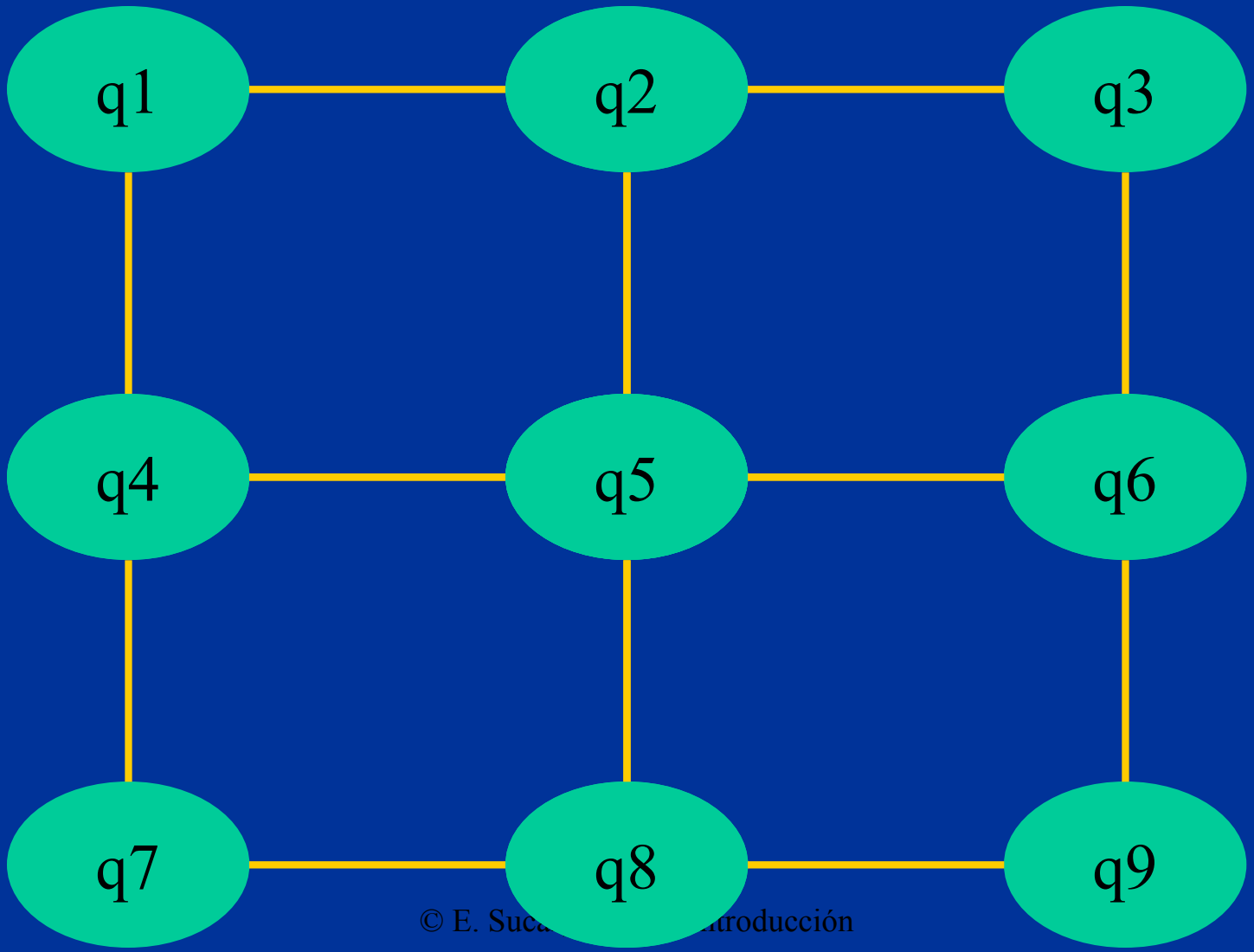
Probabilidades

- Podemos distinguir dos factores que determinan la probabilidad de una configuración:
  - la P a *priori* de cada estado,
  - la P conjunta con sus vecinos.
- En el modelo de *Ising*, estos corresponden a la influencia de un campo magnético externo, y a las interacciones entre los dipolos vecinos.

Campos en 2-D

- Esto lo podemos extender de una dimensión a dos dimensiones. En este caso tenemos una malla de puntos, donde el estado de cada punto depende del estado de sus vecinos (4, 8, etc).

Ejemplo



Representación

- Un campo aleatorio es una colección de variables aleatorias indexadas por sitios.
- Se considera un conjunto de variables aleatorias F = $\{F_1,\ldots..,\ F_M\}$, asociadas a cada sitio del sistema de sitios S. Cada variable toma un valor f_i de un un conjunto de posibles valores L. Entonces F es un *campo aleatorio*.
- Un campo *aleatorio de Markov* (CAM) es un campo aleatorio con la propiedad de "localidad".

Propiedades

Un CAM debe satisfacer las siguientes propiedades:

- $P(f) > 0 - positivo$
- $P(f_i \mid f_{s-i}) = P(f_i \mid vec(f_i))$

Donde *vec(f_i)* son los vecinos de f_i

Vecindad

Un sistema de vecindad para S se define como:

$$V = \{V_i \mid \forall_i \in S\}$$

- Cumple con las siguientes dos propiedades:
  1. Un sitio no es vecino de si mismo.
  2. La relación de vecindad es mutua.
- Se pueden tener diferentes “vecindades” (primer orden, segundo orden, etc.).

Vecindad

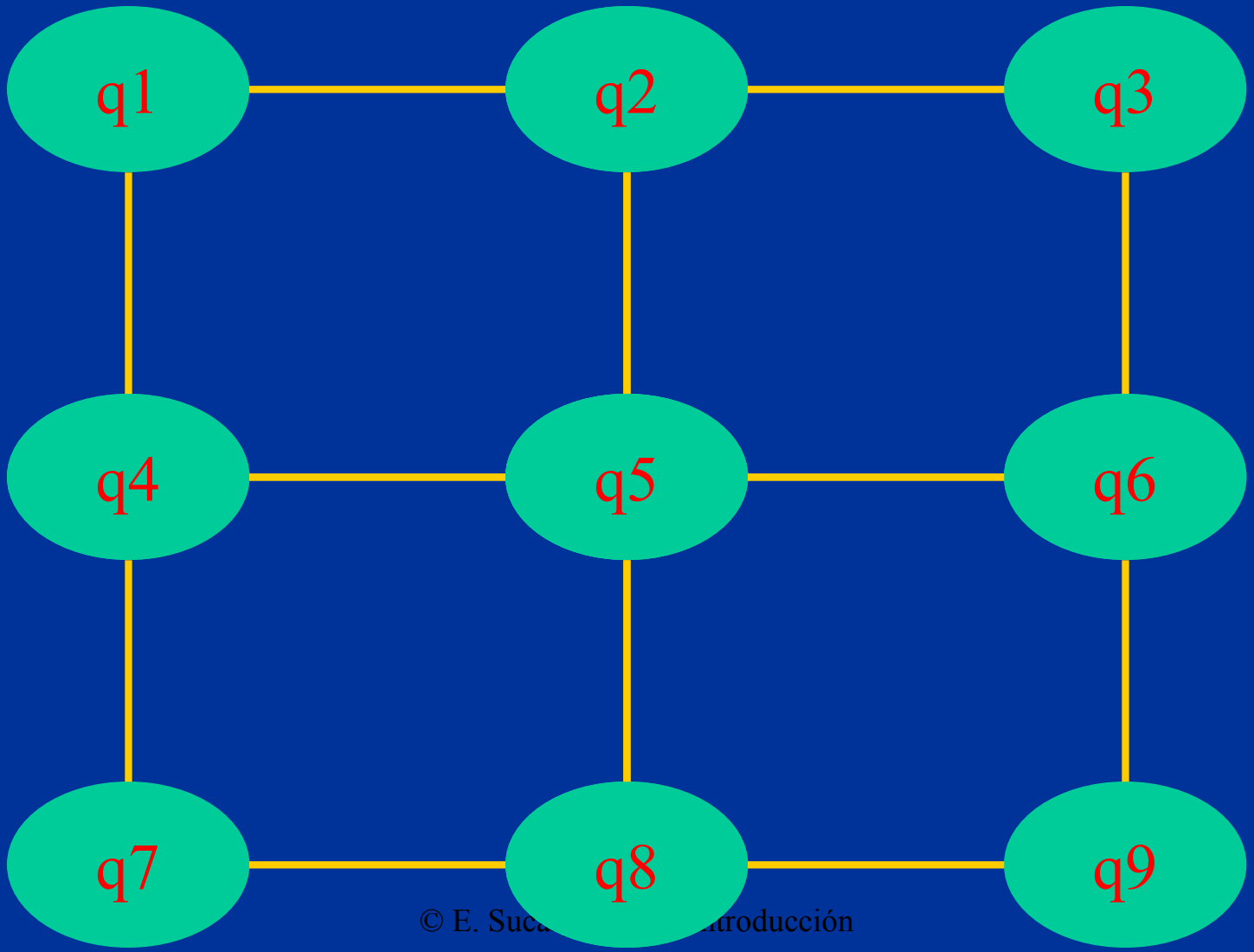
Para una malla regular, la vecindad de orden *i* con radio *r* se define como:

$$V_i = \left\{ v_{(i)} \in S \mid dist\left(i, v\left(i\right)\right) \leq r \right\}$$

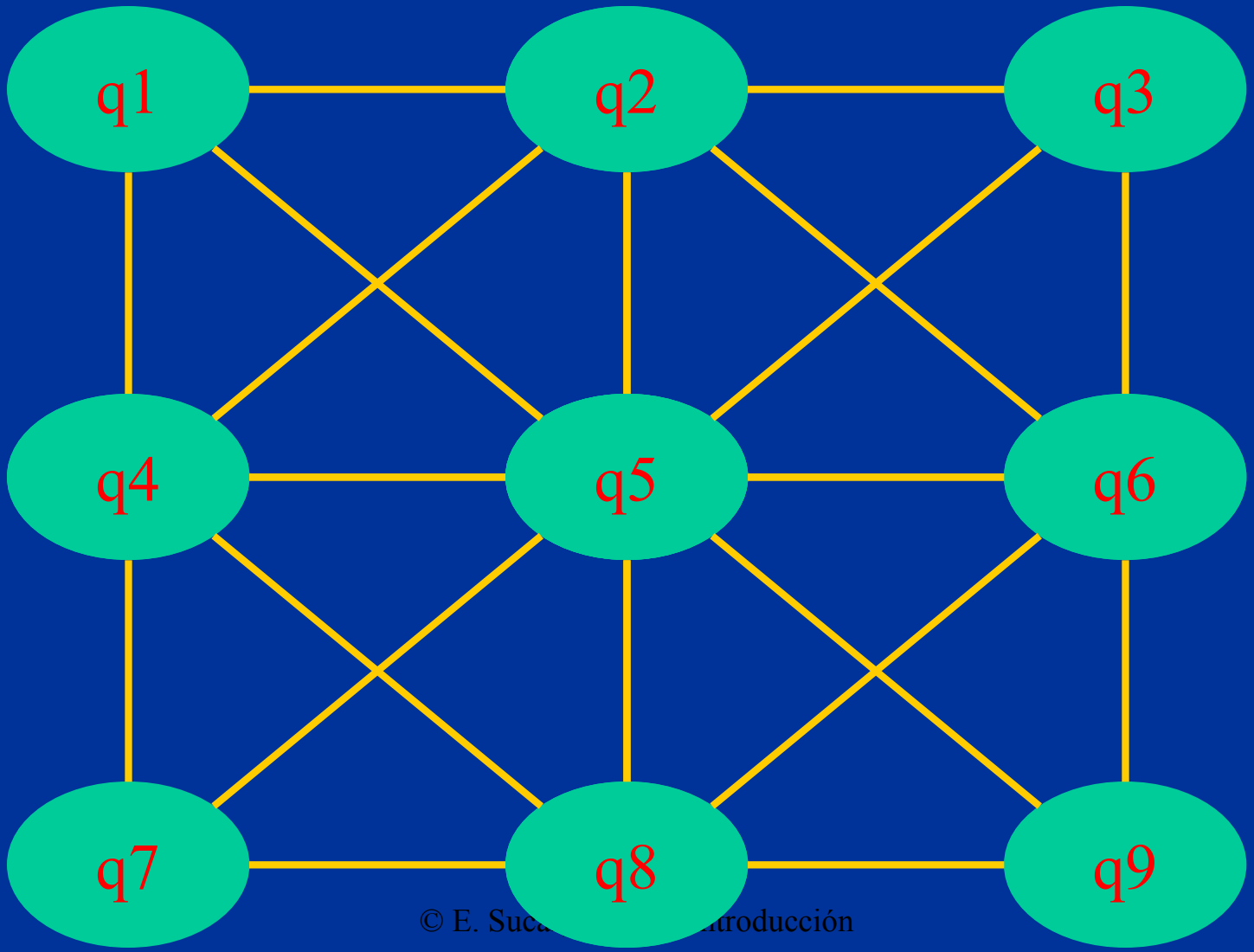
Donde *dist(x,y)* es la distancia euclidiana entre *x* y *y*.

En un sistema de vecindad de primer orden, cada, sitio (interior) tiene 4 vecinos; en uno de segundo orden, 8 vecinos: en uno de tercer orden, 12 vecinos, etc.

Ejemplo – 1er orden



Ejemplo – 2do orden



Conjuntos Completos

El conjunto de sitios junto con las vecindades nos definen un grafo no-dirigido: G=*(S, V)*.

Un *conjunto completo (C)* se define como un subconjunto de sitios de forma que están todos conectado contra todos

Un *conjunto completo* puede ser un solo sitio, C_1, un par de sitios vecinos, C_2, una tripleta de vecinos, C_3, y así sucesivamente. De forma que la colección de *C* para un grafo esta dada por:

$$C = C_1 \cup C_2 \cup C_3 \cup \ldots$$

Configuración más probable

Una aplicación típica de CAMs es para obtener la configuración más probable dadas ciertas restricciones representadas por las probabilidades locales (potenciales)

Podemos expresar la probabilidad conjunta, como el producto de las probabilidades de las vecindades:

$$P_{F/G}(f) = k \prod_c P_c$$

Configuración más probable

Dichas probabilidades de las vecindades (potenciales) se pueden ver como "restricciones" que van a *favorecer* o *desfavorecer* ciertas configuraciones.

De esta forma, la configuración más probable se puede ver como aquella que tiene una mayor *compatibilidad* con las probabilidades locales.

Configuración más probable

Podemos expresar los potenciales en forma de exponenciales (de esta forma pueden ser cualquier número real):

$$\text{Pot}(x_c) = \exp\{-U_c(x_c)\}$$

Así que el producto se vuelve una suma:

$$\text{Up}(f) = \Sigma\, U_c(x_c)$$

Por lo que la probabilidad conjunta se puede expresar como:

$$P_{F/G}(f) = (1/Z)\exp[-U_p(f)]$$

Campo de Gibbs

Lo anterior también se puede obtener mediante una analogía entre los CAM y los Campo *Aleatorio de Gíbbs* (CAG).

Una distribución de Gibbs tiene la siguiente forma (Z es una constante de normalización):

$$P(f) = \left(1/Z\right) \exp\left[-U(f)/T\right]$$

$$Donde:$$

$$Z = \Sigma_{f \in F} \exp\left[-U(f)/T\right]$$

Campo de Gibbs

U(f) se conoce como la función de energía y se obtiene como la suma de los *potenciales* de todos los C:

$$U(f) = \Sigma c\, Vc(f)$$

La configuración más probable corresponde a la de energía mínima.

La función de energía se puede expresar en términos de los *C* de cada tamaño:

$$U(f) = \Sigma_{c1} V_1(f_i) + \Sigma_{c2} V_2(f_i, f_j) + \Sigma_{c3} V_3(f_i, f_j, f_k) + \ldots..$$

Campo de Gibbs

Se puede demostrar que un CAM y el correspondiente CAG son equivalentes.

Esto permite que se pueda expresar la probabilidad conjunta especificando los potenciales de los *C*. En la práctica se seleccionan los potenciales de acuerdo al comportamiento deseado – con los potenciales se codifica el conocimiento a *priori* del problema.

Configuración más probable

- Entonces, para especificar un CAM se requiere:
  - Definir el esquema de vecindad
  - Especificar las probabilidades (potenciales) para cada uno de los conjuntos completos de nodos

Para el caso de vecindad de primer orden:

$$U_p(f) = \Sigma_c V_c(f) + \lambda \Sigma_O V_O(f)$$

Configuración más probable

- V_c corresponde a P_F o la información del dominio dada por los vecinos y V_O corresponde a $P_{G/F}$ o la información de las observaciones; λ es una constante que da el peso relativo entre ambas.

- Bajo este enfoque, la solución a un problema particular corresponde en encontrar la configuración del CAM de mayor probabilidad o de "energía" (U_P) mínima. La función que se logre depende de la forma de las funciones para V_C y V_0.

Ejemplo

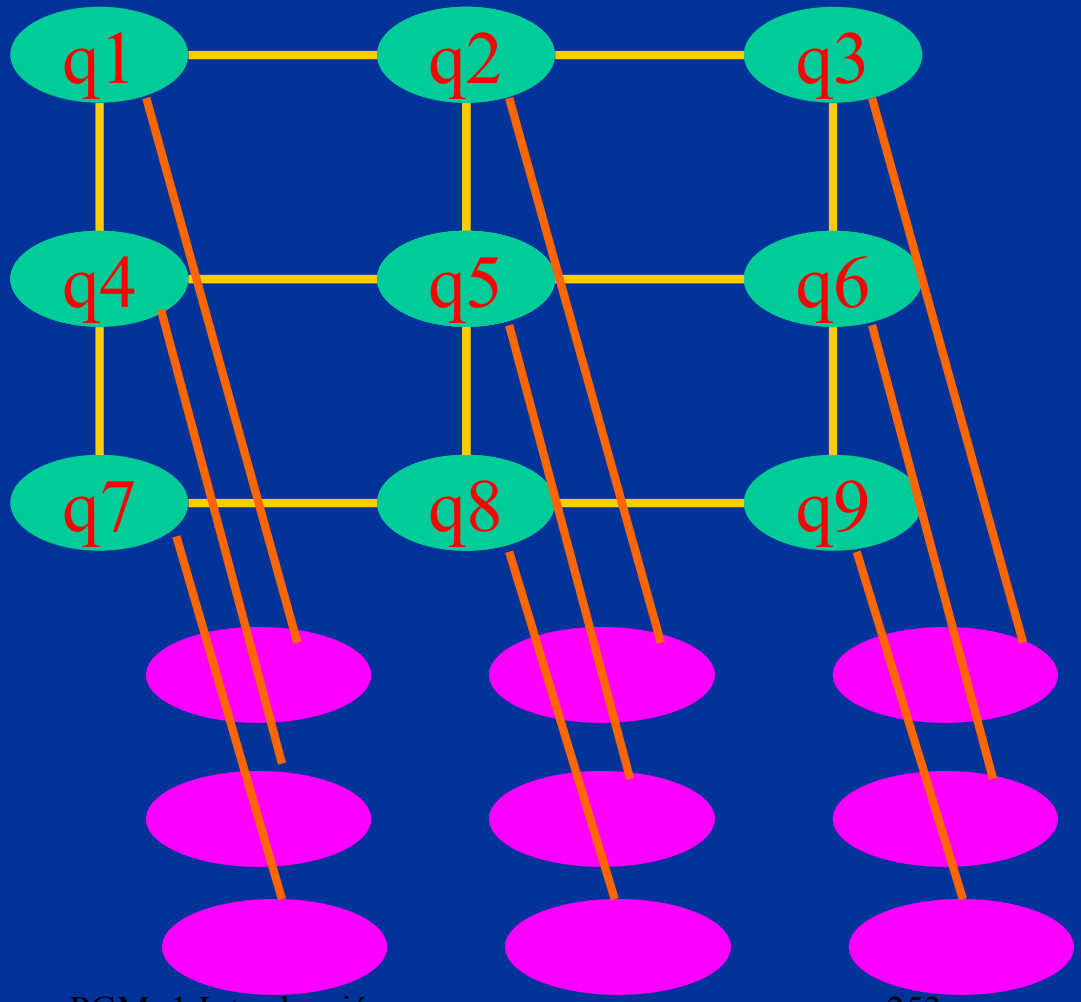
- Por ejemplo, podemos querer “suavizar” una imagen; es decir, minimizar la “distancia” de cada pixel a sus vecinos, pero también mantenerlo cercano a su valor en la imagen (observación):

$$V_c(f) = (f - u)^2$$
$$V_0(f) = (f - g)^2$$

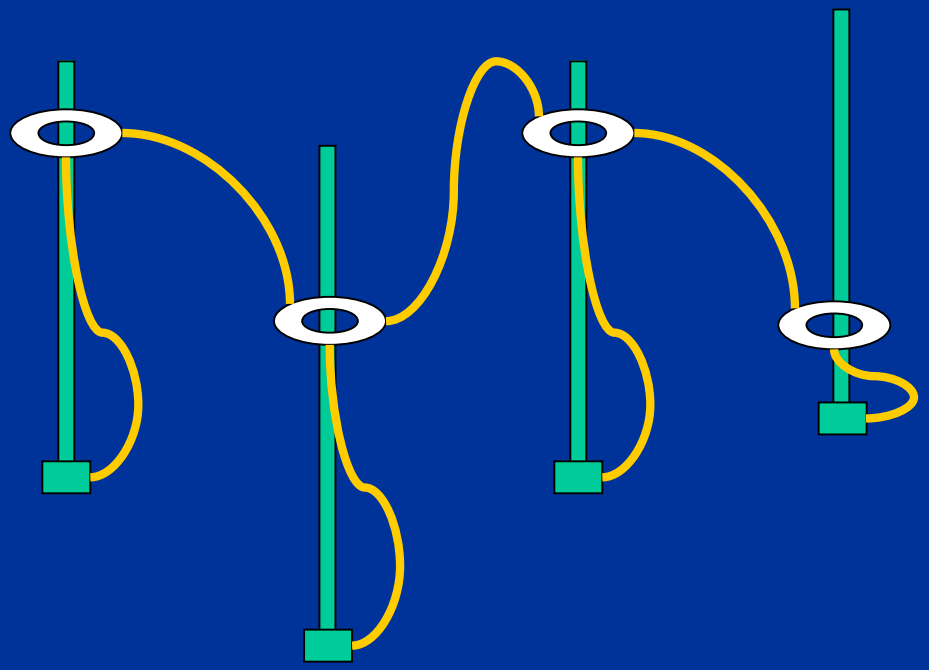
Ejemplo

F
imagen “suavizada”

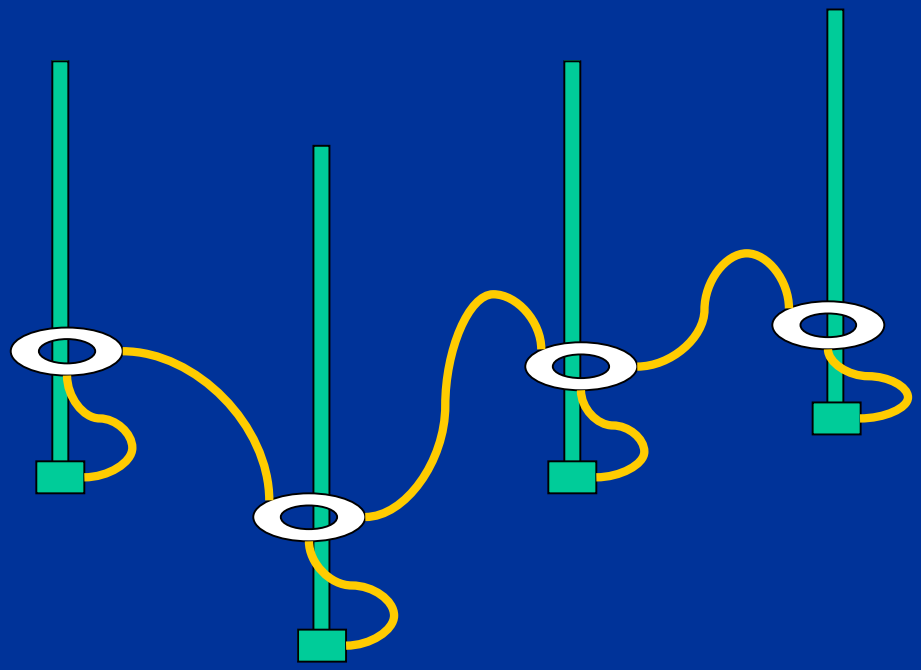
G
imagen “original”



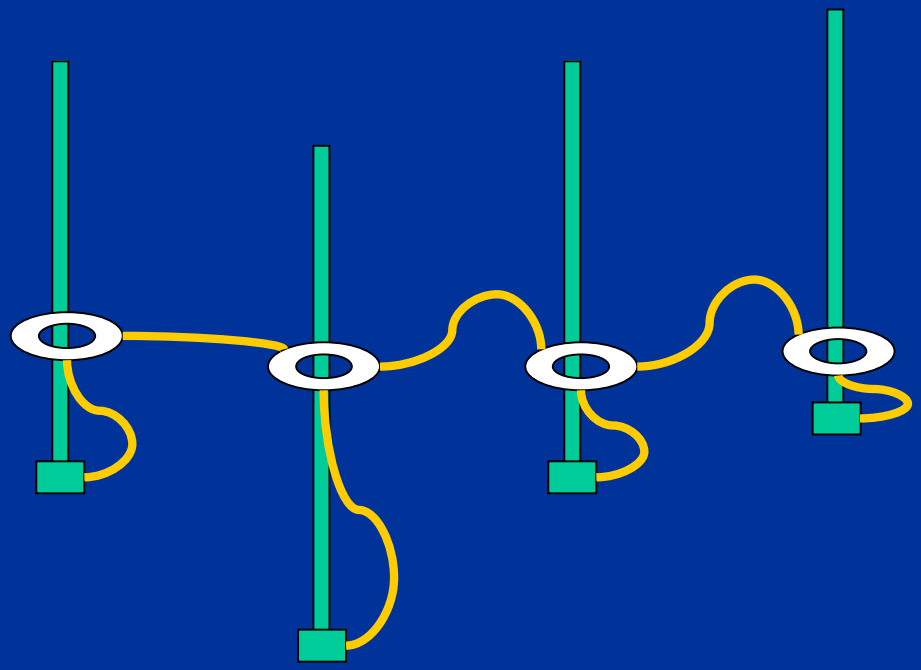
Analogía Física



Analogía Física
- más peso a las observaciones -



Analogía Física
- más peso a los vecinos -



Algoritmos

- El obtener la configuración de mayor probabilidad (mínima energía) en forma directa (exhaustiva) prácticamente imposible (excepto para problemas muy pequeño), por lo que se plantea como un problema de *búsqueda*. Se busca la configuración de mayor probabilidad, sin tener que calcular directamente las probabilidades de cada configuración.

Algoritmos

El problema incluye 3 aspectos:

1. Representación: CAM con un esquema de vecindad y los potenciales asociados
2. Función objetivo: Función de energía a minimizar.
3. Algoritmo de optimización: simulación estocástica (Metropolis)

Algoritmo básico

- Inicializar con un valor aleatorio cada variable.
- Repetir para cada variable en el campo:
  - Calcular el valor de energía (potencial) de cada variable en base a la función deseada y los valores de los vecinos.
    - Si el valor de energía es menor al anterior cambiar de valor.
    - Si no, con cierta probabilidad también cambiar de valor.
- Hasta que se cumplan N iteraciones o ya no haya cambios (convergencia)
- Obtener configuración “óptima”

Variantes

Cálculo del óptimo:

- MAP: se toma el valor para cada variable al final de las iteraciones.
- MPM: se toma el valor de mayor frecuencia de acuerdo a su ocurrencia durante la simulación.

Variantes

Forma de optimización:

- Metrópolis: con probabilidad fija se puede pasar a estados de mayor energía.
- Recocido simulado: se va disminuyendo la probabilidad de pasar a estados de mayor energía (temperatura).
- ICM: tomar siempre el estado de menor energía.

Probabilidades de transición

- Probabilidad de transición a estado de mayor energía:

$$P = e^{-\delta V / T}$$

- Donde δV es la diferencia de energía y *T* es la *temperatura*

Ejemplo

- Dada la siguiente imagen:
- Consideramos:
  - Configuración inicial de ceros
  - Vecindad de primer orden
  - Potenciales de suavizamiento con $l = 4$

0	0	0	0
0	1	1	1
0	1	0	1
0	1	1	1

Ejemplo

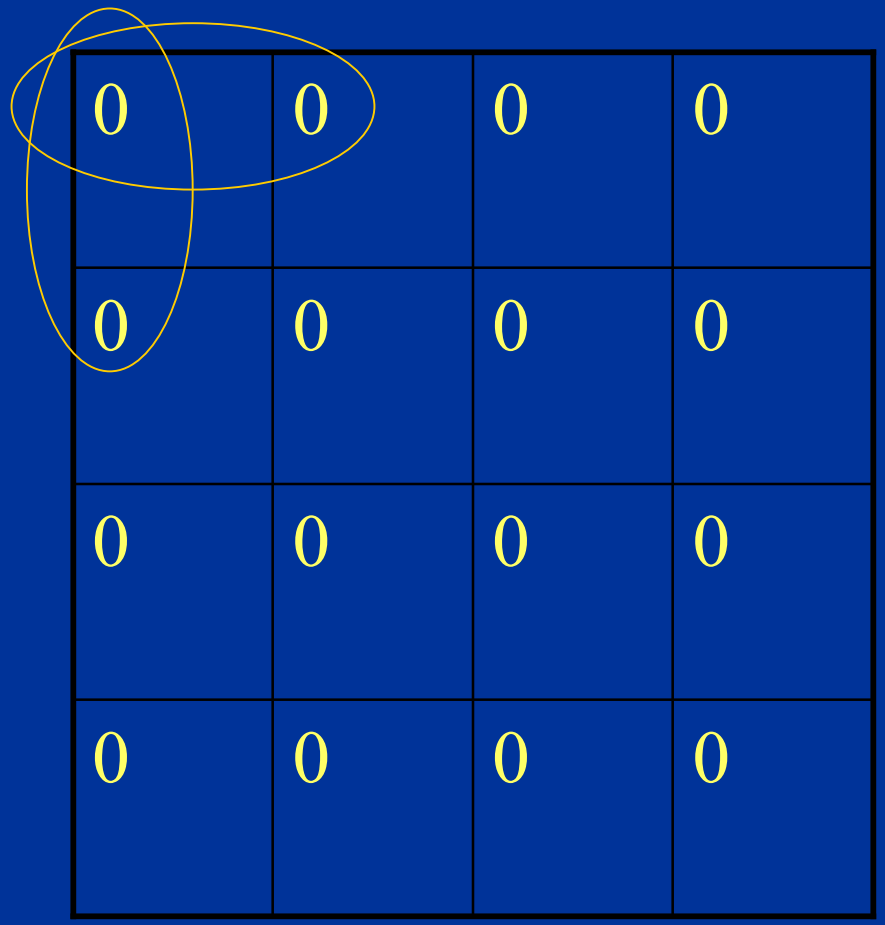
1era iteración:

- 1er sitio:

  V1 (0) = 0

  V1(1) = 2 + 4 (1) = 6

- …
- …

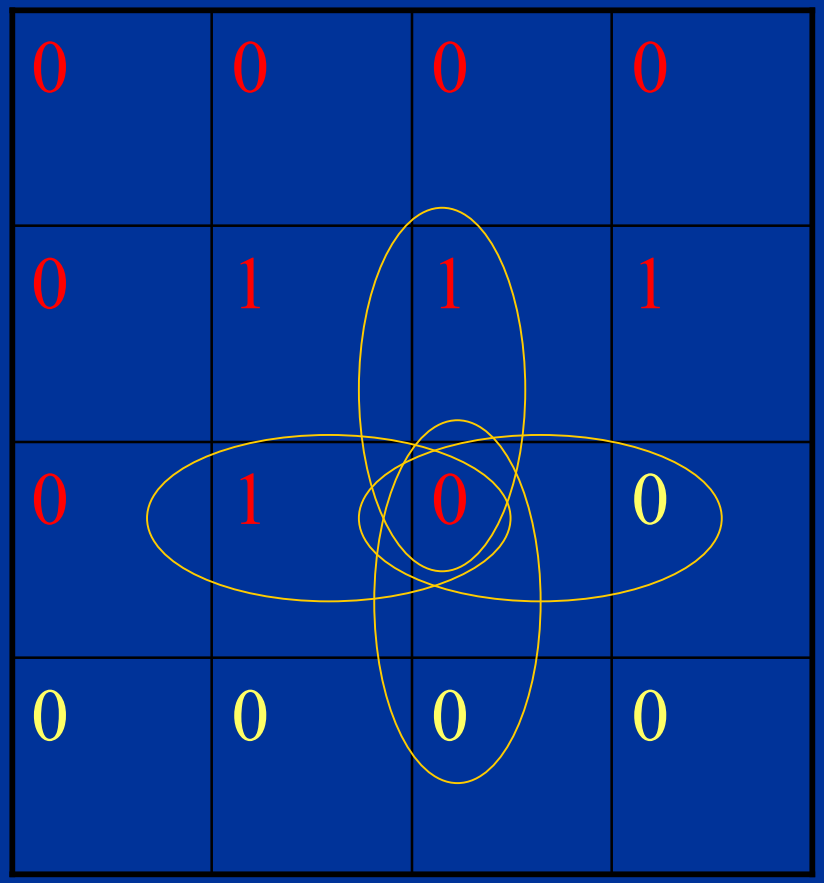


Ejemplo

- …
- …
- 11vo sitio:

  V11(0) = 2

  V11(1) = 6
- …



Ejemplo

2da iteración

- …
- 11ovo sitio:

  V11(0) = 4

  V11(1) = 4
- …

0	0	0	0
0	1	1	1
0	1	1	1
0	1	1	1

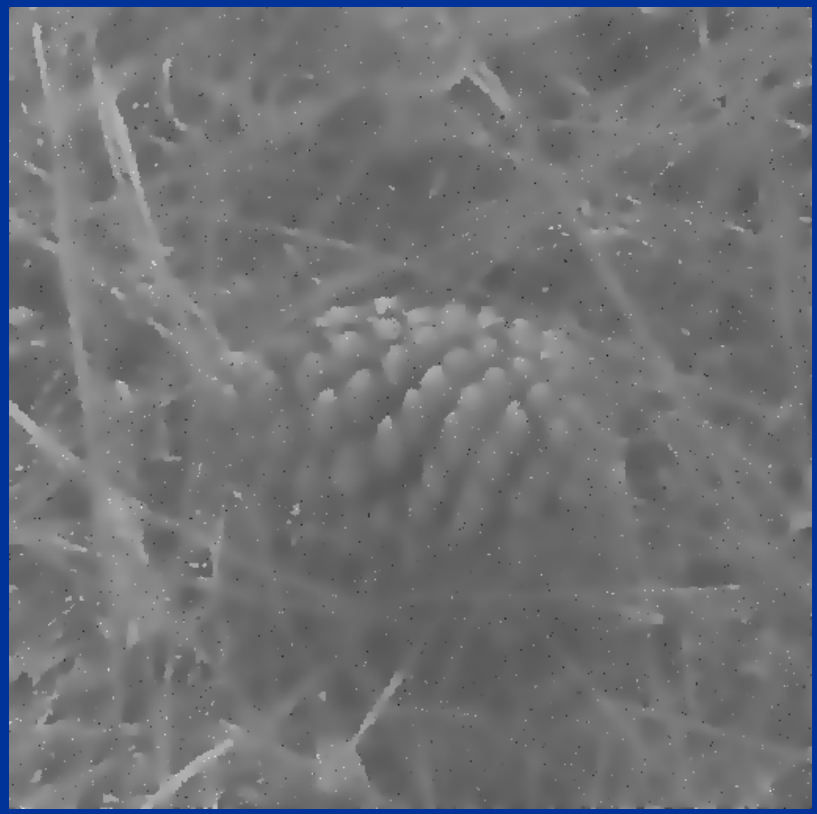
Aplicaciones

- Procesamiento de imágenes
  - Eliminación de ruido
  - Filtrado de una imagen
  - Segmentación de texturas
  - Visión estereoscópica
  - Recuperación de forma (3-D)
  - Reconocimiento de objetos
  - Etiquetado de imágenes

Ejemplo – eliminación de ruido

Ejemplo – eliminación de ruido

Ejemplo – eliminación de ruido



Referencias

- Li, “Markov Random Fields Models in Computer Vision”, Springer-Verlag
- Chellapa, Jain, “Markov Random Fields: Theory and Models”, Academic Press.

Sumario

- Los modelos gráficos probabilistas permiten representar una distribución de probabilidad en forma compacta y más *entendible*
- A partir de estas representaciones podemos obtener la probabilidad de ciertas variables dadas otras, mediante lo que se conoce como *razonamiento probabilista*

Sumario

- Existen varios tipos de modelos: clasificadores bayesianos, redes bayesianas, campos de Markov, etc.
- El tipo del modelo más adecuado depende de la aplicación
- Hay diversas técnicas para aprender estos modelos a partir de datos