Mô hình đồ thị xác suất

Trần Quốc Long

Nội dung

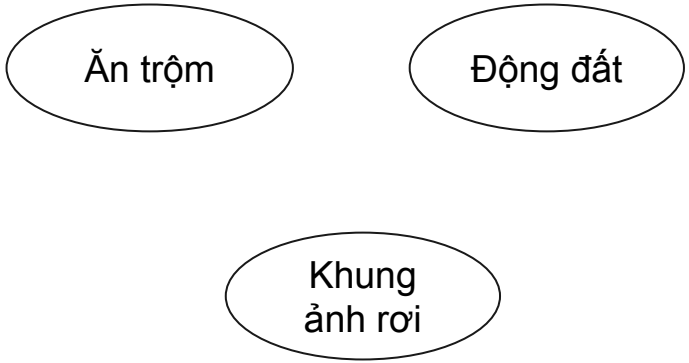
1. Giới thiệu
2. Các mô hình đồ thị
3. Suy diễn
4. Ứng dụng

Giới thiệu

- Mô hình hoá
- Các đặc trưng cần mô hình hoá
- Công cụ xác suất

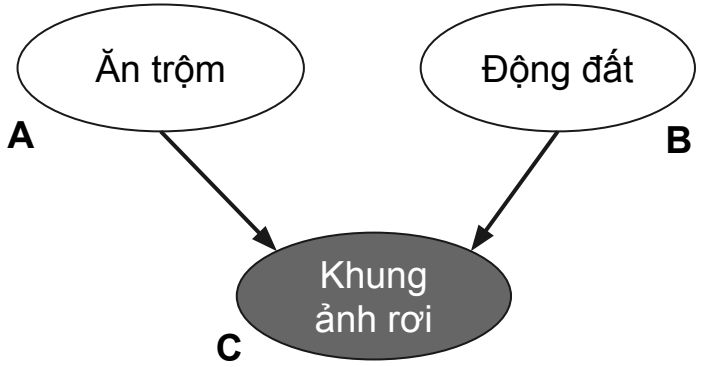
Mô hình hóa: tại sao ?

- Đơn giản hóa



Mô hình hóa: tại sao ?

- Đơn giản hóa
- Trực quan hóa



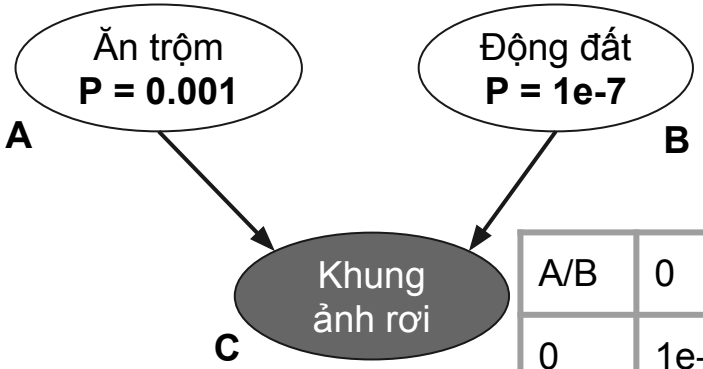
Ô rỗng: chưa biết

Ô đặc: đã biết

Mô hình hóa: tại sao ?

- Đơn giản hóa
- Trực quan hóa
- Định lượng hóa

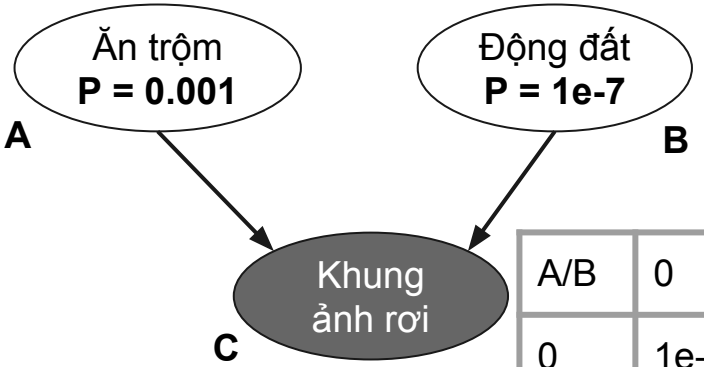
$$P(A, B, C)$$
$$= P(A) \times P(B) \times P(C|A, B)$$



A/B	0	1
0	1e-4	0.8
1	0.4	0.99

Mô hình hóa: tại sao ?

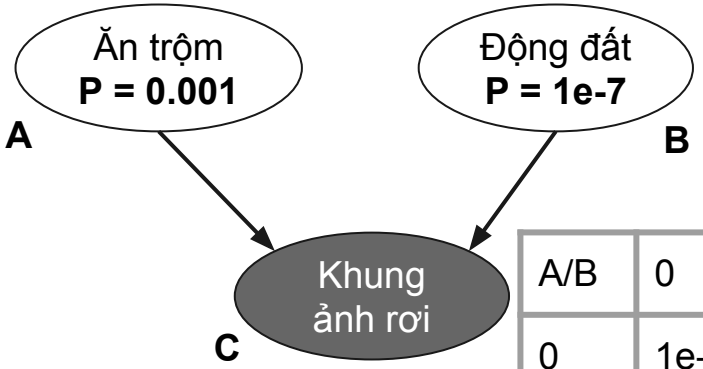
- Đơn giản hóa
- Trực quan hóa
- Định lượng hóa
- Mô phỏng



A/B	0	1
0	1e-4	0.8
1	0.4	0.99

Mô hình hóa: tại sao ?

- Đơn giản hóa
- Trực quan hóa
- Định lượng hóa
- Mô phỏng
- Truy vấn / Suy diễn



A/B	0	1
0	1e-4	0.8
1	0.4	0.99

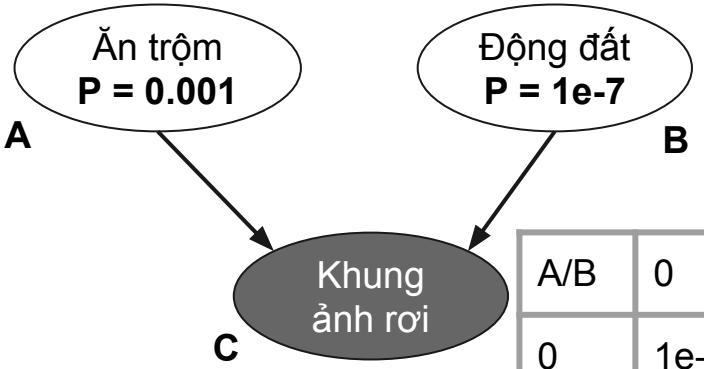
Nếu khung ảnh rơi, khả năng ăn trộm = ? **80%**

Nếu khung ảnh rơi và động đất, khả năng ăn trộm = ? **0.1%**

Các đặc trưng cần mô hình hóa

Tính không chắc chắn

- Sự vật, hiện tượng có thể xảy ra hoặc không xảy ra
  - Tất nhiên / ngẫu nhiên

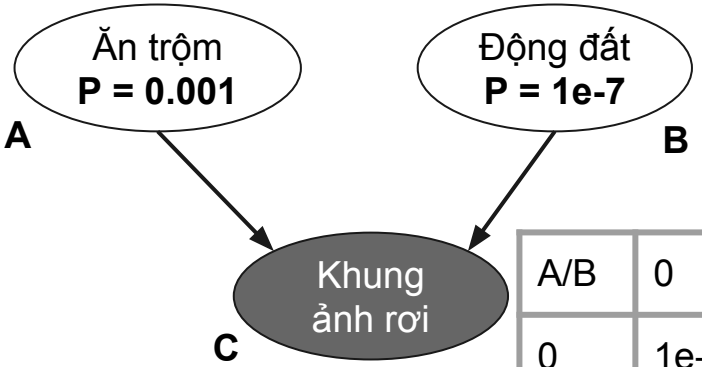


A/B	0	1
0	1e-4	0.8
1	0.4	0.99

Các đặc trưng cần mô hình hóa

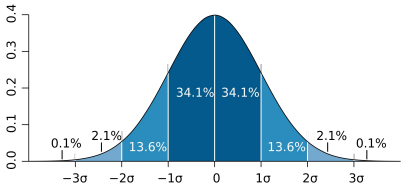
Tính cấu trúc

- Sự vật, hiện tượng có mối liên hệ với nhau
  - Nguyên nhân / kết quả



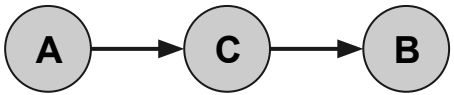
A/B	0	1
0	1e-4	0.8
1	0.4	0.99

Công cụ xác suất



Tính không chắc chắn ⇒ **Phân bố xác suất**

Tính cấu trúc ⇒ **Độc lập xác suất**



$A \perp B \mid C$

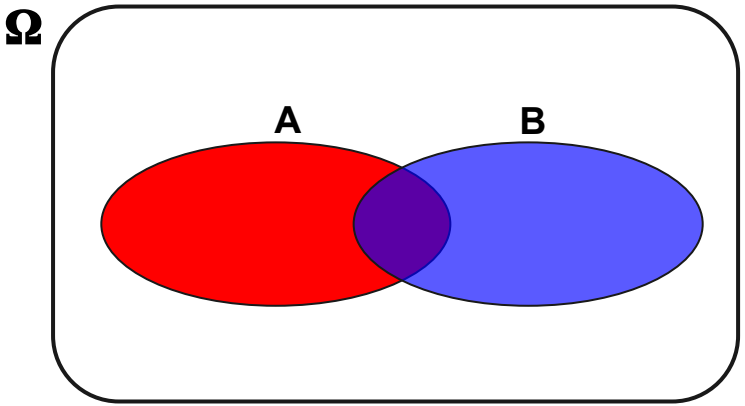
Độc lập xác suất

- Biến cố độc lập

$$A \perp B$$

$$P(A, B) = P(A) \times P(B)$$

$$P(A \mid B) = P(A)$$



$$\frac{A,B}{B} = \frac{A}{\Omega}$$

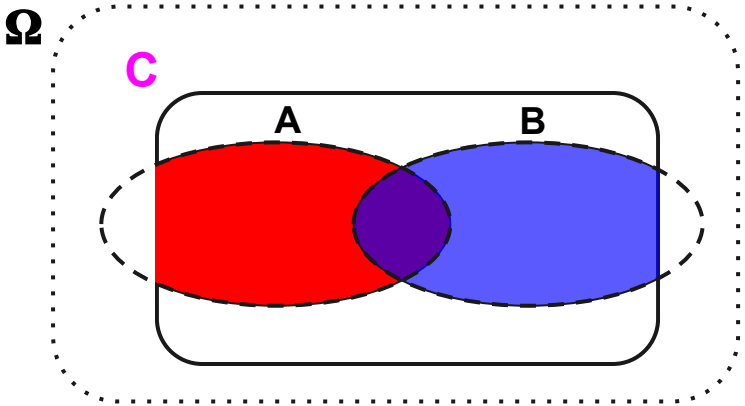
Độc lập xác suất

- Độc lập có điều kiện

$$A \perp B \mid C$$

$$P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$$

$$P(A \mid B, C) = P(A \mid C)$$



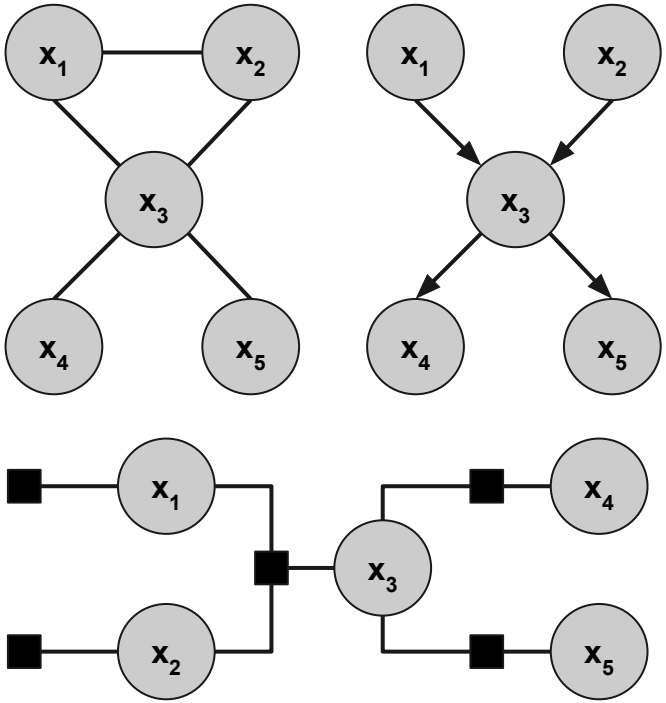
$$\frac{A,B,C}{B,C} = \frac{A,C}{C}$$

Các mô hình đồ thị

- Quy ước trên mô hình đồ thị
- Mạng Bayes (đồ thị có hướng)
- Mạng Markov (đồ thị vô hướng)

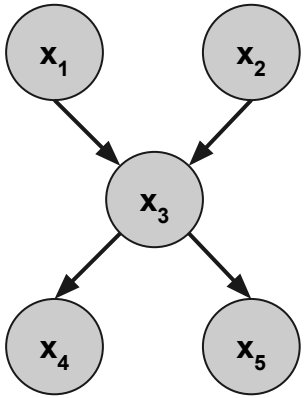
Mô hình đồ thị

- Biến ⇔ đỉnh của đồ thị
- Các mối quan hệ ⇔ cạnh
  - Có hướng hoặc vô hướng
- Quy ước về mối quan hệ / độc lập xác suất giữa các biến
- Dạng hàm của phân bố liên hợp



Mạng Bayes

- **Đồ thị DAG:**
  - Có hướng
  - Không có chu trình
- Mỗi đỉnh chỉ phụ thuộc vào các đỉnh cha mẹ



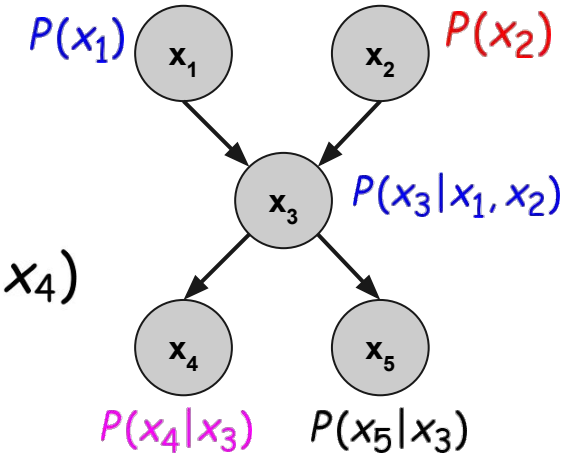
$x_4 \perp x_2 \mid x_3$ $x_1 \perp x_2$ $x_4 \perp x_5 \mid x_3$???

$x_1, x_2 \perp x_4, x_5 \mid x_3$ $x_1 \perp x_2 \mid x_3$???

Mạng Bayes

- Xác suất liên hợp

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$
$$P(x_4|x_1, x_2, x_3)P(x_5|x_1, x_2, x_3, x_4)$$
$$= P(x_1)P(x_2)P(x_3|x_1, x_2)$$
$$P(x_4|x_3)P(x_5|x_3)$$



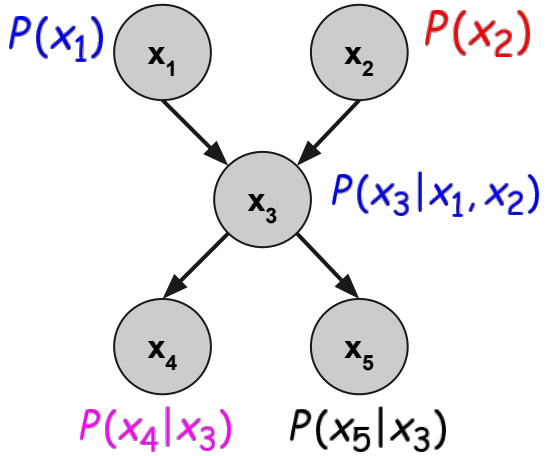
- Số lượng tham số (biến nhị phân)

31 → 1 + 1 + 4 + 2 + 2 = **10**

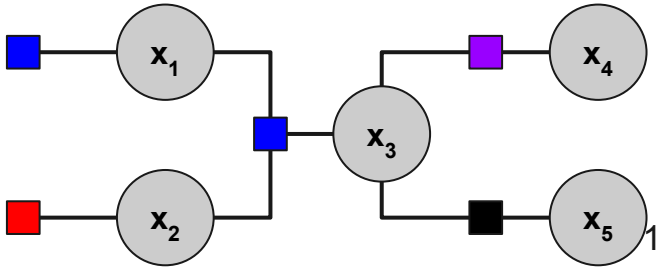
Mạng Bayes

- Xác suất liên hợp = tích các xác suất có điều kiện (nút cha)

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1) \times P(x_2) \times P(x_3|x_1, x_2) \times P(x_4|x_3) \times P(x_5|x_3)$$

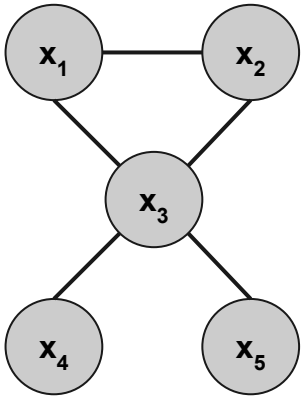


- Đồ thị nhân tử



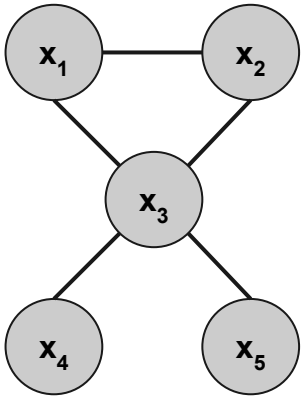
Mạng Markov

- **Đồ thị vô hướng**
  - Cạnh vô hướng
  - Có thể có chu trình



Mạng Markov

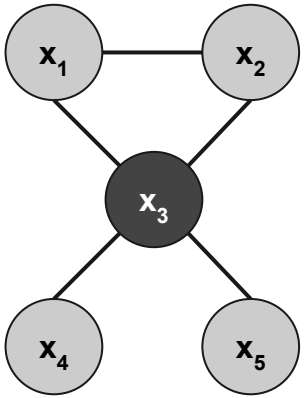
- **Tính Markov cục bộ:** Mỗi đỉnh chỉ phụ thuộc vào các đỉnh kề



$$x_1 \perp x_4, x_5 \mid x_2, x_3$$

Mạng Markov

- **Tính Markov toàn cục:** Hai tập đỉnh độc lập nếu bị chia cắt bởi các đỉnh đã biết



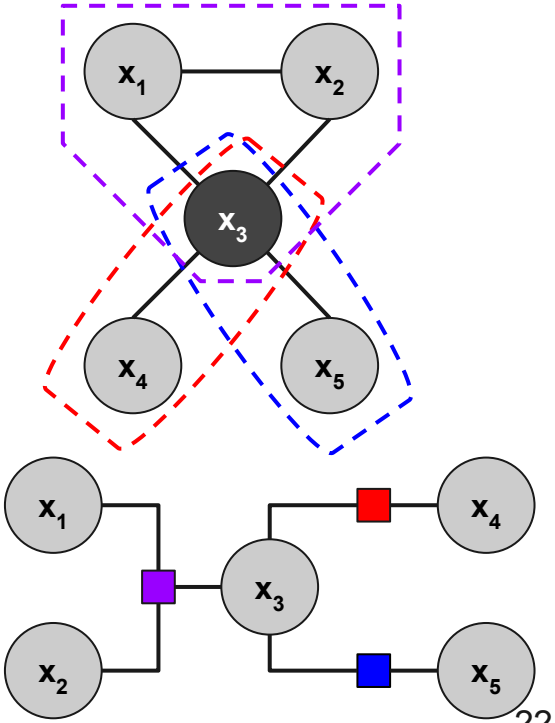
$$x_1, x_2 \perp x_4, x_5 \mid x_3$$

Mạng Markov

Định lý Hammersley–Clifford:

$$P(x_1, \cdots, x_5) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

= tích các hàm trên các đồ thị con đầy đủ (clique)



Suy diễn

- Truy vấn mô hình đồ thị
- Suy diễn chính xác
- Suy diễn xấp xỉ

Truy vấn mô hình đồ thị

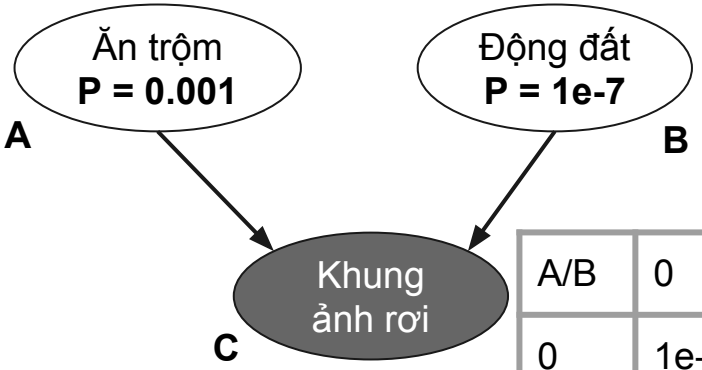
Câu hỏi:

Nếu khung ảnh rơi, khả năng ăn trộm

$$P(A = 1 | C = 1) = ?$$

Nếu khung ảnh rơi và động đất, khả năng ăn trộm

$$P(A = 1 | B = 1, C = 1) = ?$$



A/B	0	1
0	1e-4	0.8
1	0.4	0.99

Truy vấn mô hình đồ thị

$$P(A = 1 | C = 1) = \frac{P(A = 1, C = 1)}{P(C = 1)}$$

$$P(A = 1, C = 1) = P(A = 1, B = 0, C = 1) + P(A = 1, B = 1, C = 1)$$

$$P(C = 1) = P(A = 1, B = 0, C = 1) + P(A = 1, B = 1, C = 1) + P(A = 0, B = 0, C = 1) + P(A = 0, B = 1, C = 1)$$

Nhanh hơn, áp dụng ***tính phân phối*** của phép nhân

$$A \times B + A \times C = A \times (B + C)$$

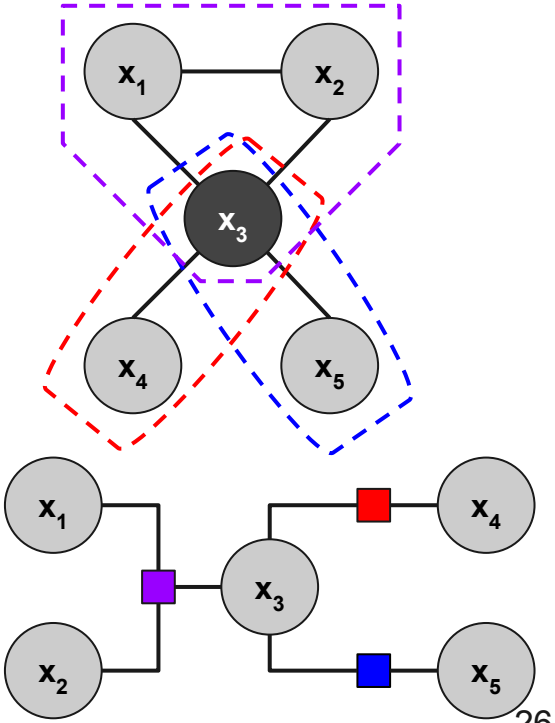
Truy vấn mô hình đồ thị

Cho mô hình đồ thị

$$P(x_1, \cdots, x_5) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

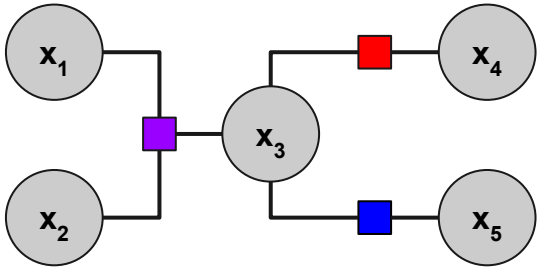
Tính $P(x_A \mid x_B)$

- $\mathbf{x_A}$: tập các nút cần truy vấn
- $\mathbf{x_B}$: tập các nút đã biết giá trị



Đồ thị nhân tử

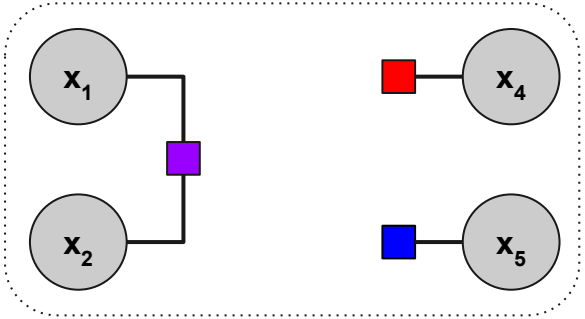
$$P(x_1, \cdots, x_5) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$



Nếu biết $x_3 = 1$

$$P(x_1, x_2, x_4, x_5 | x_3 = 1) = \frac{1}{Z'}\psi_{123}(x_1, x_2, 1)\psi_{34}(1, x_4)\psi_{35}(1, x_5)$$

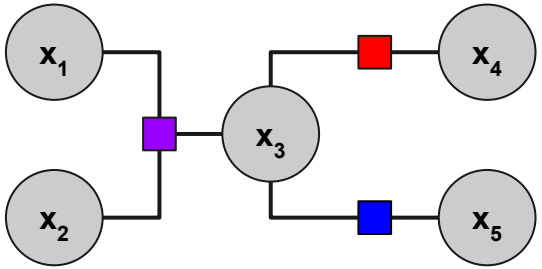
$$= \frac{1}{Z'}\psi'_{12}(x_1, x_2)\psi'_4(x_4)\psi'_5(x_5)$$



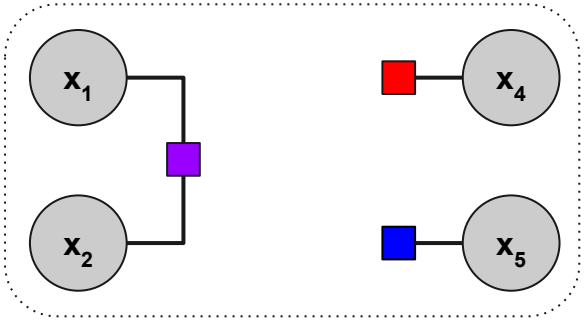
→ Đồ thị nhân tử mới

Đồ thị nhân tử

Tính $P(x_A \mid x_B)$ ở đồ thị cũ



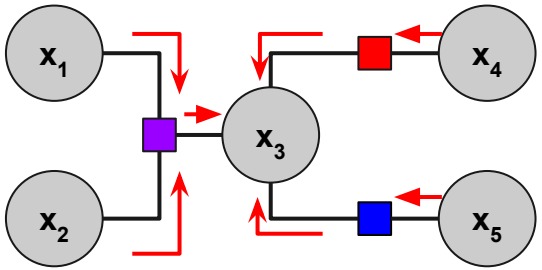
⇔ Tính $P(x_A)$ ở đồ thị mới



Thuật toán Tổng - tích

Trên đồ thị nhân tử dạng cây

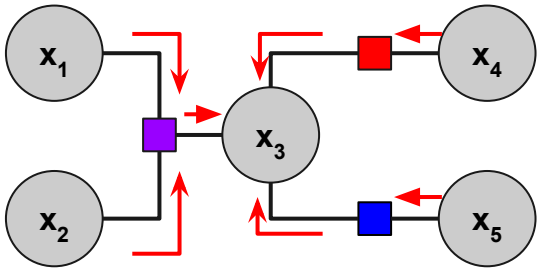
Tính $P(x_A)$



- **Chính xác**
- **Hiệu quả**
- **Ý tưởng**: *tính phân phối → gom các nhân tử thành **tin nhắn** phát trên các cạnh đồ thị*

Thuật toán Max - tổng

Trên đồ thị nhân tử dạng cây



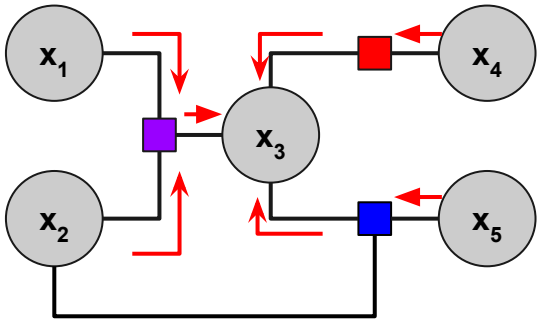
Tính $max_{x_A} P(x_A)$

- **Lấy logarit:** *Tích các nhân tử* → *Tổng*
- Tính phân phối của phép cộng:

$$\max(a + b, a + c) = a + \max(b, c)$$

Suy diễn xấp xỉ

Trên đồ thị tổng quát



- Tổng - tích, max - tổng
  có thể lặp vô hạn
- Suy diễn biến phân: EM, mean-field, …
- Suy diễn Monte-Carlo: Gibbs, Metropolis
- Heuristic: lan truyền tin nhắn như thường

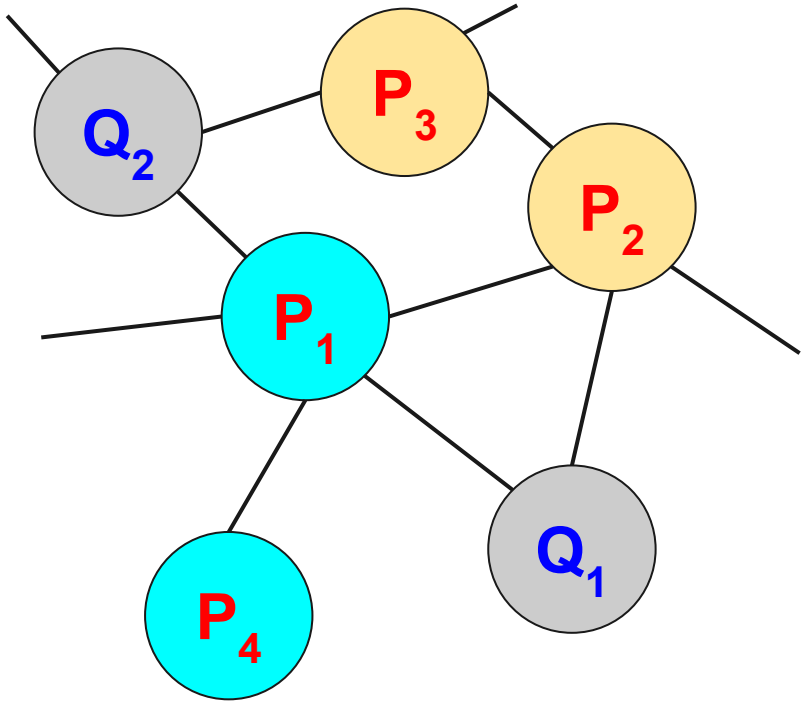
Ứng dụng

- Tin sinh học
- Truyền tin (mã hoá, giải mã)
- Xử lý ảnh / âm thanh / video, xử lý ngôn ngữ tự nhiên

Đoán chức năng protein

Mạng tương tác protein (PPI)

- Biết P_1, … P_n có F hay không.
- Hỏi Q_1, … Q_m có F không ?

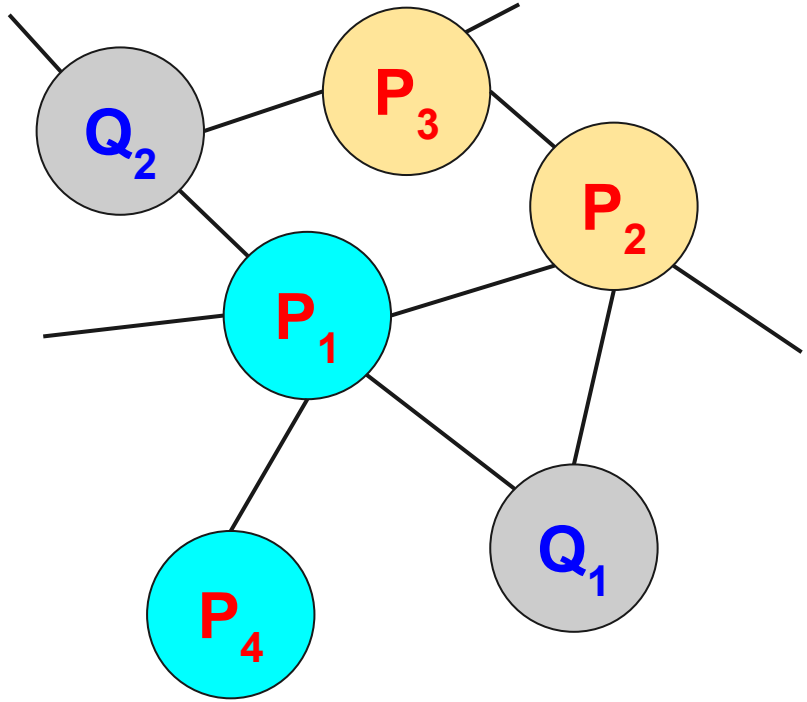


Đoán chức năng protein

Deng (2003)

$$\max_{Q} \mathbb{P}(Q|P)$$

$$\mathbb{P}(Q,P) = \exp\{\alpha N_1 + \beta N_{10} + \gamma N_{11} + N_{00}\}$$



Đoán chức năng protein

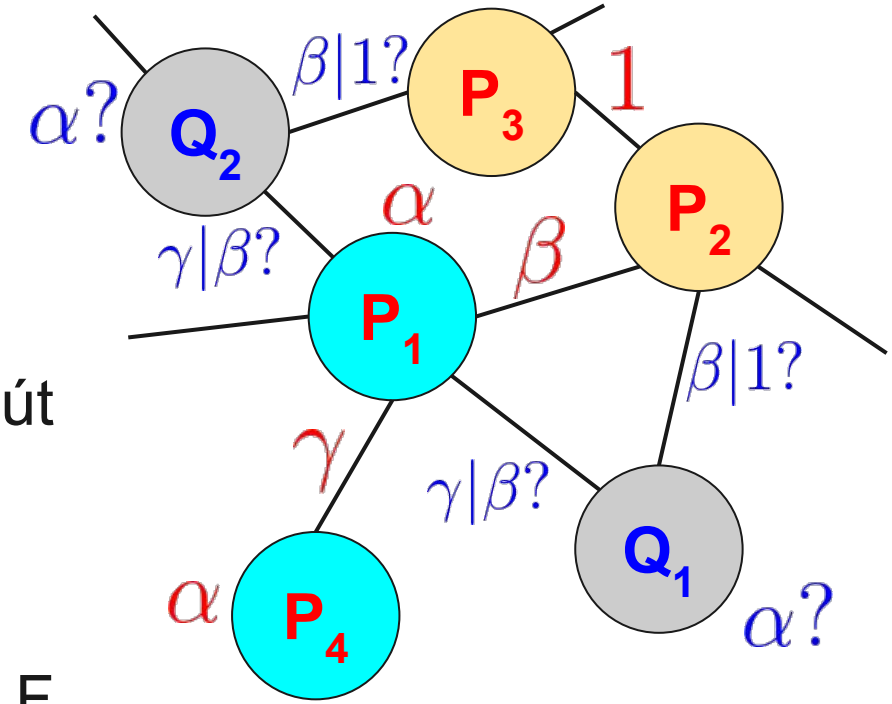
$$\mathbb{P}(Q, P) = \exp\{\alpha N_1 + \beta N_{10} + \gamma N_{11} + N_{00}\}$$

N_1: số nút có F

N_{10}: số cạnh nối nút có F và nút không có F

N_{11}: số cạnh nối nút có F

N_{00}: số cạnh nối nút không có F



Mã LDPC (Gallager)

H: ma trận kiểm tra chẵn lẻ

G: ma trận sinh mã

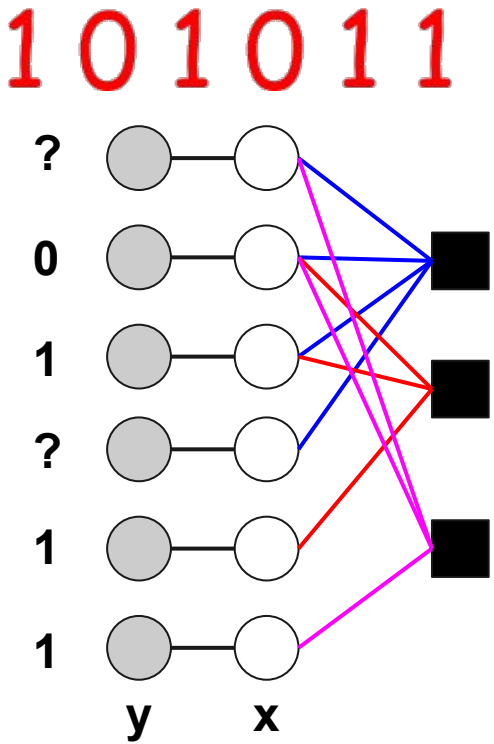
x: dữ liệu cần truyền

xG: dữ liệu + bit kiểm tra

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Mã LDPC



$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\psi_{1234}(x_1, x_2, x_3, x_4)$$

$$\psi_{235}(x_2, x_3, x_5)$$

$$\psi_{126}(x_1, x_2, x_6)$$

Mã LDPC

- Truyền video: DVB-S2 / DVB-T2 / DVB-C2
- Chuẩn WiMAX: IEEE 802.16e
- Mạng không dây: IEEE 802.11n

Ngoài ra: **mã Turbo** - ***liên lạc liên hành tinh***

(cũng dùng thuật toán tổng - tích)