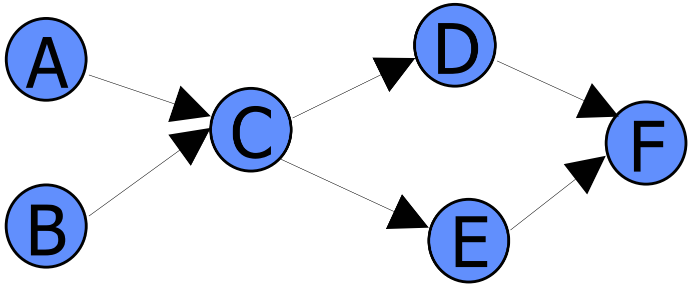
Réseaux Bayésiens

Définition - Inférence

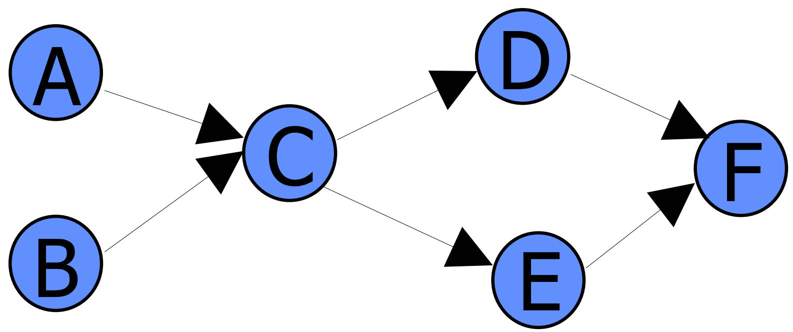
Philippe Leray
Philippe.Leray@insa-rouen.fr

INSA Rouen – Département ASI – Laboratoire PSI (FRE CNRS 2645)



Introduction

- Modèle graphique (MG) = mariage entre la théorie des graphes et la théorie des probabilités
  - graphe d'états
  - probabilités de transition



- De nombreux modèles utilisés en Machine Learning peuvent être vus comme des cas précis de MG
  - Modèles de Markov Cachés (HMM)
  - Réseaux Bayésiens (RB)
  - . . .

Architecture des Systèmes d'Information

Plan

- Définition
- Algorithmes d'inférence
  - Bucket Elimination
  - Message Passing (Pearl)
  - Junction Tree (Jensen)
- Applications et Offre logicielle

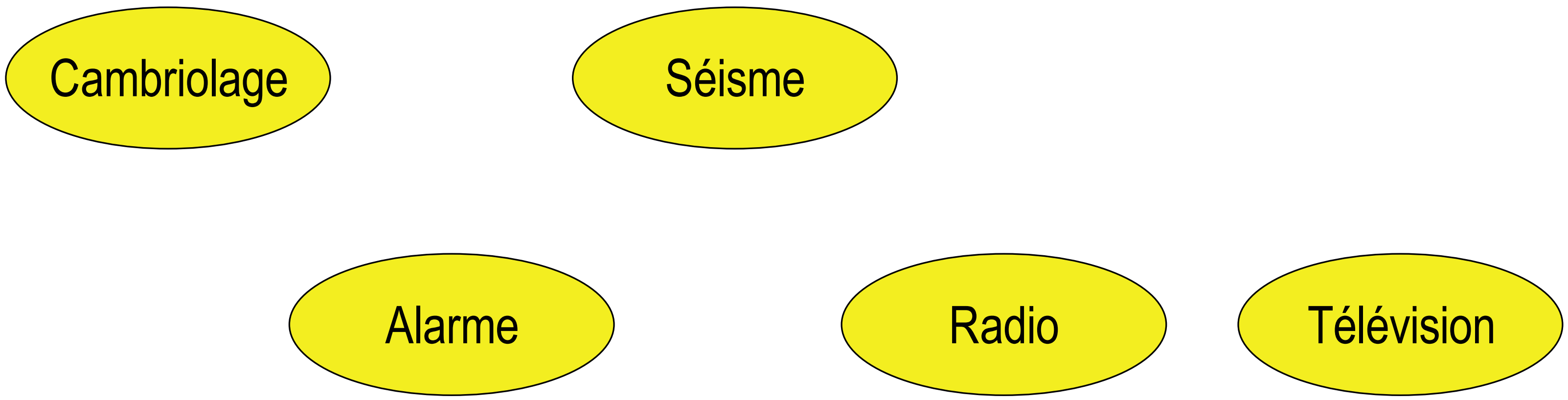
- Apprentissage
  - des paramètres
  - de la structure
- Modèles étendus
  - variables continues : modèles conditionnels gaussiens
  - problèmes temporels : modèles dynamiques
  - théorie de la décision : diagrammes d'influence

Origine

- Inspiration des systèmes experts :
  - règles :
    - si soleil=VRAI et arrosage=FAUX, alors sol=sec
  - comment rajouter des informations probabilistes ?
    - $P(soleil) = 0.3 \qquad P(arrosage) = 0.5 \qquad P(sec) = ?$
  - comment inverser l'inférence d'un SE ?
    - $P(sec) = 0.9 \qquad P(arrosage) = 0.5 \qquad P(soleil)?$
- (Pearl 1988) Raisonnement probabiliste
- De nombreuses appellations :
  - SE bayesien, SE probabiliste, réseau de croyance, réseau causal
  - belief network, bayesian network, probabilistic independence networks

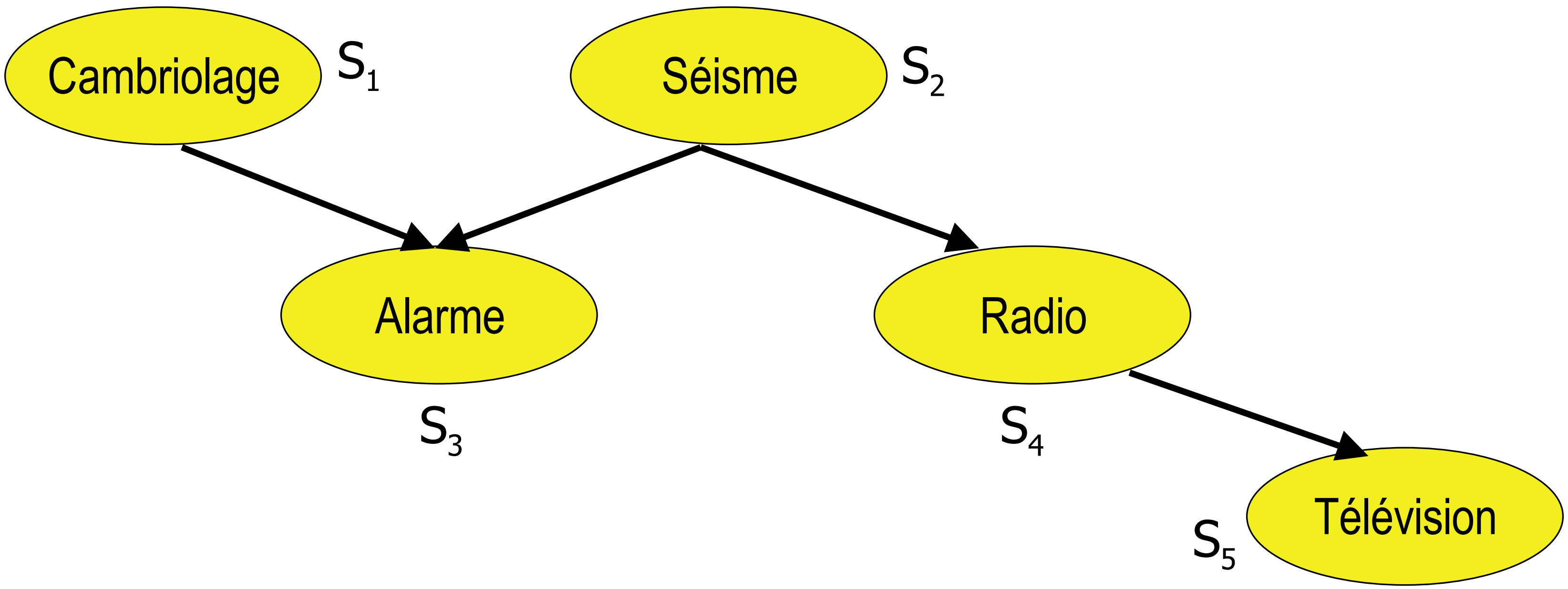
Réseau Bayésien

- Réseau bayésien =
  - description qualitative des dépendances entre des variables (graphe causal)
  - description quantitative de ces dépendances
- Exemple



Exemple

- Dépendance entre les variables (graphe)
  - DAG = Directed Acyclic Graph
  - Numérotation des variables dans l'ordre topologique

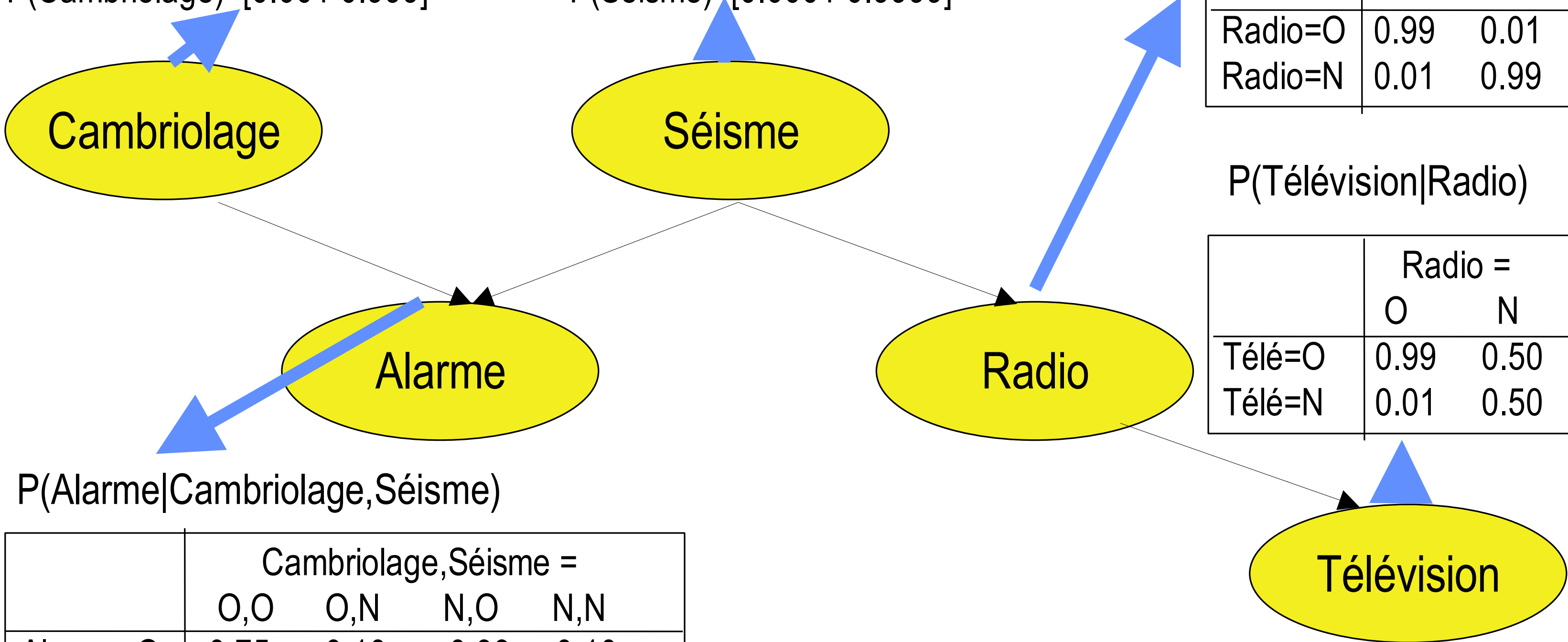


Exemple

- Description quantitative des dépendances (probabilités conditionnelles

P(Cambriolage)=[0.001 0.999]

P(Séisme)=[0.0001 0.9999]



P(Radio|Séisme)

	Séisme = O	N
Radio=O	0.99	0.01
Radio=N	0.01	0.99

P(Télévision|Radio)

	Radio = O	N
Télé=O	0.99	0.50
Télé=N	0.01	0.50

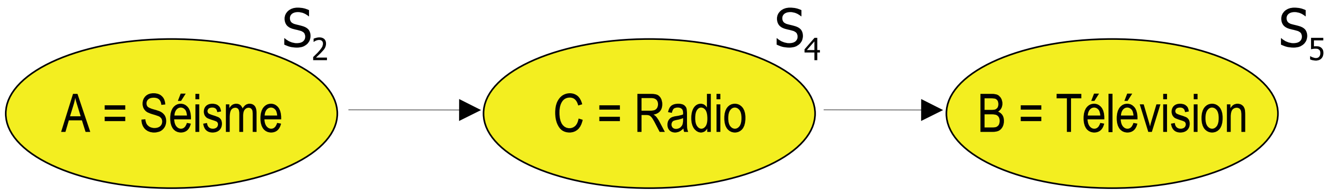
P(Alarme|Cambriolage,Séisme)

	Cambriolage,Séisme = O,O	O,N	N,O	N,N
Alarme=O	0.75	0.10	0.99	0.10
Alarme=N	0.25	0.90	0.01	0.90

Indépendance Conditionnelle

- Indépendance conditionnelle :
  - A et B sont indépendants conditionnellement à C ssi :
    - lorsque l'état de C est connu, toute connaissance sur B n'altère pas A
    - $\boxed{P(A|B,C) = P(A|C)}$
- Les RB vont servir à représenter graphiquement les indépendances conditionnelles
- Exemple sur 3 nœuds
  - 3 types de relations possibles entre A, B et C ...

Indépendance Conditionnelle

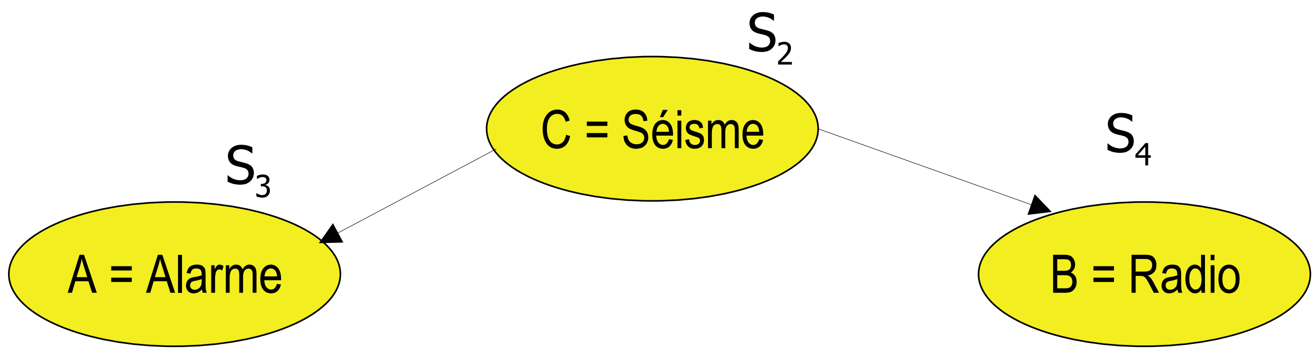


- Connexion série
- A et B sont dépendants

- A et B sont indépendants conditionnellement à C
  - si $P(C)$ est connue,
    A n'intervient pas dans le calcul de $P(B)$

  - $P(S_5|S_4, S_2) = P(S_5|S_4) = P(S_5|parents(S_5))$

Indépendance Conditionnelle

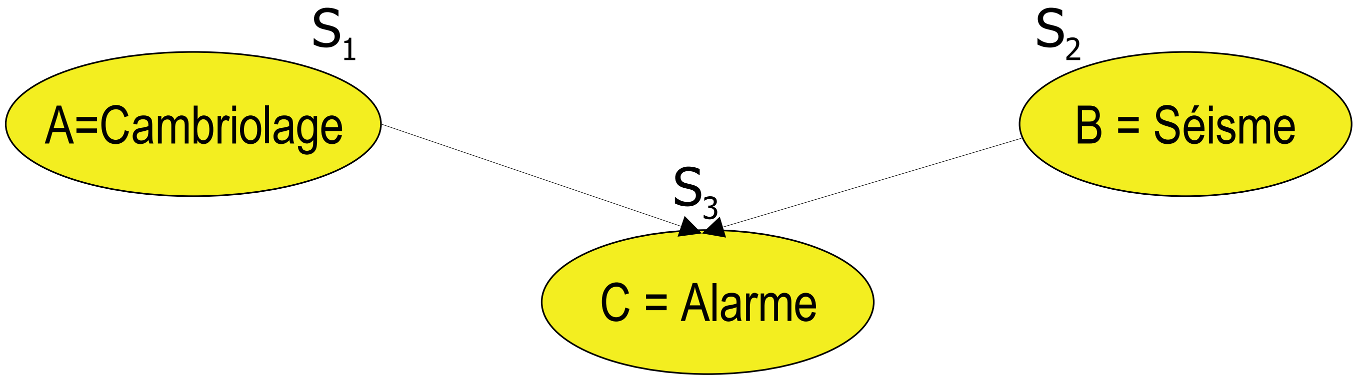


- Connexion divergente
- A et B sont dépendants

- A et B sont indépendants conditionnellement à C
  - si $P(C)$ est connue,
    A n'intervient pas dans le calcul de $P(B)$

  - $P(S_4|S_2, S_3) = P(S_4|S_2) = P(S_4|parents(S_4))$

Indépendance Conditionnelle



- Connexion convergente (A, B, C est une V-structure)
- A et B sont indépendants

- A et B sont dépendants conditionnellement à C
  - si $P(C)$ est connue,
    $P(A)$ intervient pas dans le calcul de $P(B)$

  - $P(S_3|S_1, S_2) = P(S_3|parents(S_3))$

Conséquence

- RB = représentation compacte de la loi jointe $P(S)$
- Théorème de Bayes :

$$P(S) = P(S_1) \times P(S_2|S_1) \times P(S_3|S_1, S_2) \times \cdots \times P(S_n|S_1 \ldots S_{n-1})$$

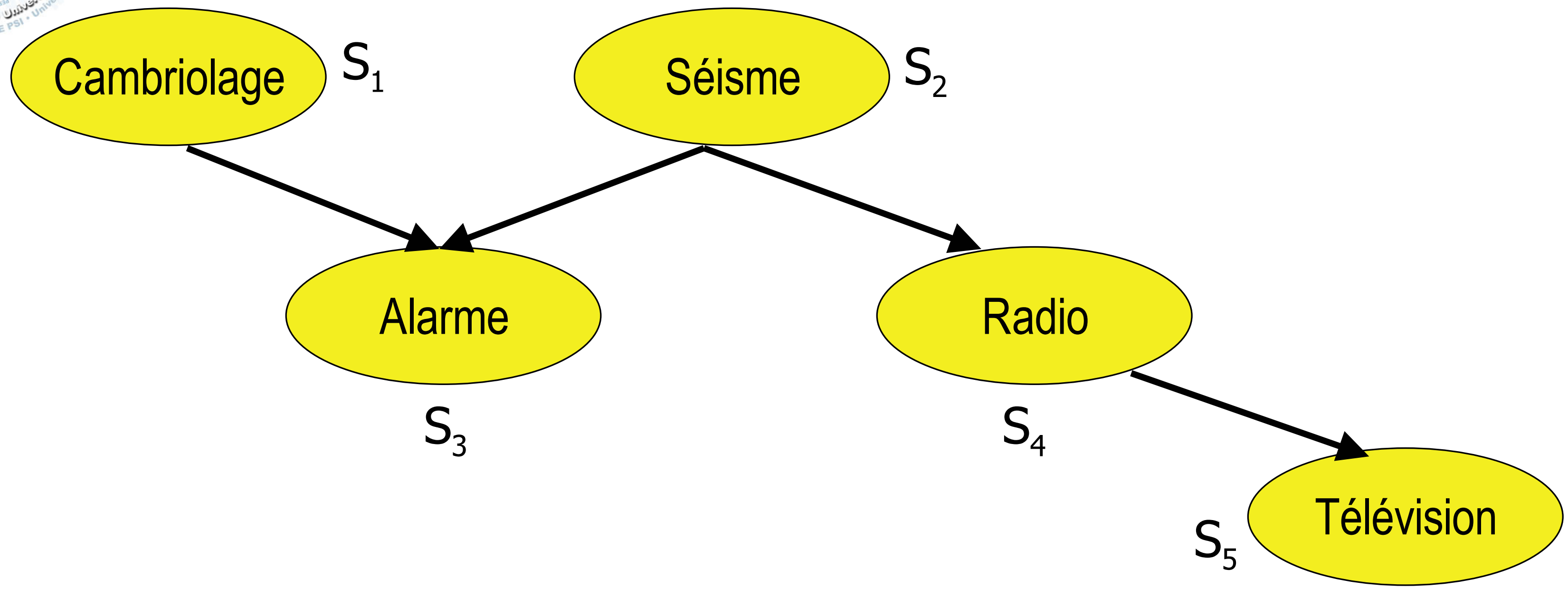
- Mais dans un RB, $P(S_i|S_1 \ldots S_{i-1}) = P(S_i|parents(S_i))$ d'où

$$\boxed{P(S) = \Pi_{i=1}^{n} P(S_i|parents(S_i))}$$

- La loi jointe (globale) se décompose en un produit de lois locales

Architecture des Systèmes d'Information

Exemple



$$P(Cambriolage, Seisme, Alarme, Radio, Tele) =$$

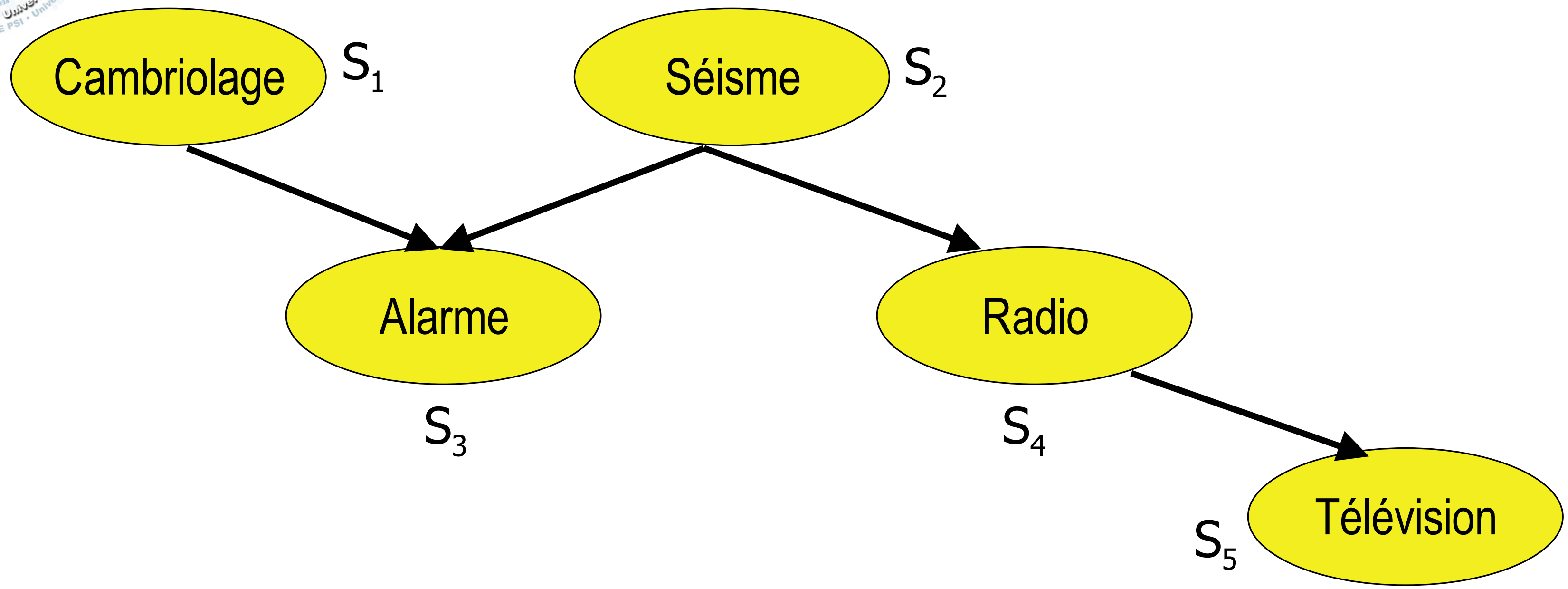
$$P(S_1)P(S_2|S_1)P(S_3|S_1,S_2)P(S_4|S_1,S_2,S_3)P(S_5|S_1,S_2,S_3,S_4)$$

$$P(S_1) \quad P(S_2) \quad P(S_3|S_1,S_2) \quad P(S_4|S_2) \quad P(S_5|S_4)$$

La d-séparation

- Principe
  - déterminer si 2 variables quelconques sont indépendantes conditionnellement à un ensemble de variables instantiées
- Définition
  - Deux variables A et B sont d-séparées si pour tous les chemins entre A et B, il existe une variable intermédiaire V différente de A et B telle que
    - la connexion est série ou divergente et V est instancié
    - la connexion est convergente et ni V ni ses descendants ne sont instanciés
- Si A et B ne sont pas d-séparés, ils sont d-connectés

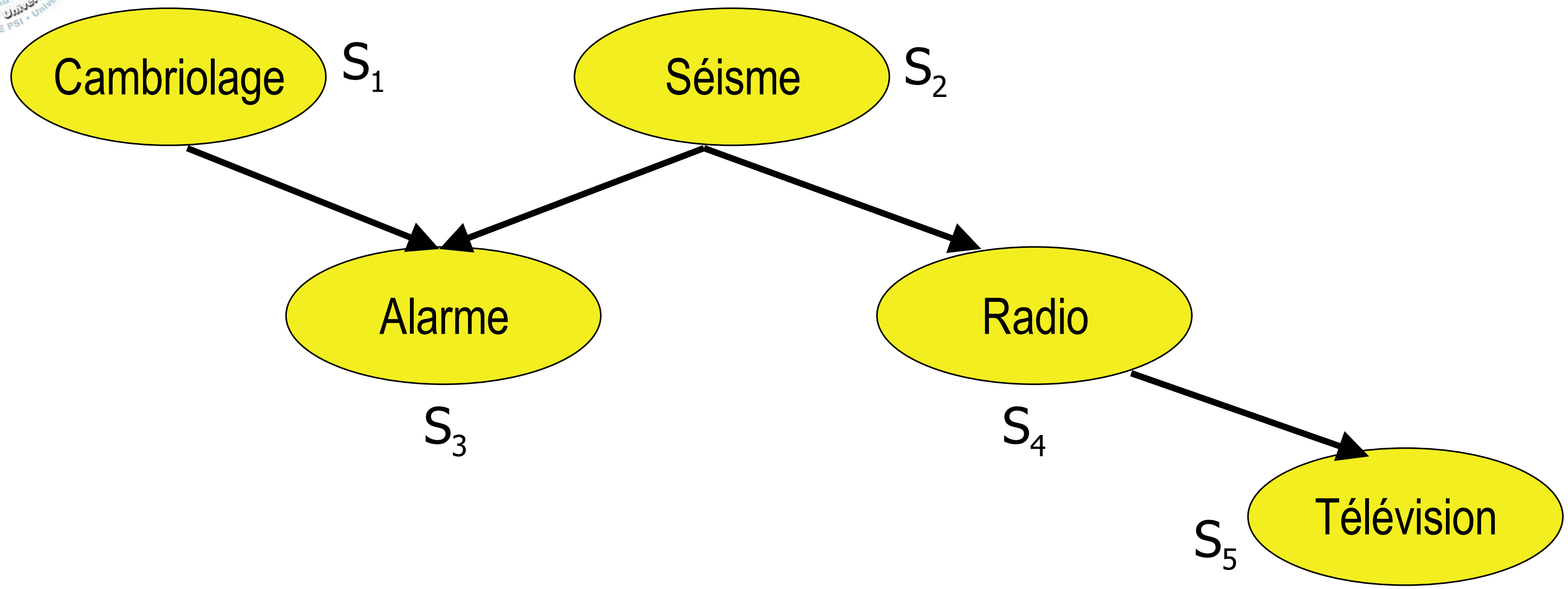
Exemple



- exemple de d-séparation : $S_1 \ldots S_4$?
  - $V = S_3$ sur le chemin entre S_1 et S_4.
  - la connexion est convergente en V
  - V n'est pas instancié

$\rightarrow$ S_1 et S_4 sont d-séparés
(si S_3 était mesuré, S_1 et S_4 seraient d-connectés)

Exemple



- autre exemple de d-séparation : $S_2 \ldots S_5$?
  - $V = S_4$ sur le chemin entre S_2 et S_5.
  - la connexion est série en V
  - V n'est pas instancié

$\longrightarrow$ S_2 et S_5 sont d-connectés
(si S_4 était mesuré, S_2 et S_5 seraient d-séparés)

Inférence

- Inférence = calcul de n'importe quelle $P(S_i|S_j = x)$
  NB : l'observation $\{S_j = x\}$ est appellée l'évidence
- A quoi sert la loi jointe $P(S) = P(S_1, \ldots, S_n)$?
- Rappel : marginalisation

$$P(S_i) = \sum_{s_1, s_2 \ldots s_n} P(S_1 = s_1, S_2 = s_2, \ldots, S_i, \ldots S_n = s_n)$$

- d'où

$$P(S_i|S_j = x) = \frac{P(S_i, S_j = x)}{P(S_j = x)} = \frac{\sum_{\{s_k\}_{k \neq i,j}} P(S_1 = s_1, \ldots, S_i, S_j = x, \ldots S_n = s_n)}{\sum_{\{s_k\}_{k \neq j}} P(S_1 = s_1, \ldots, S_j = x, \ldots S_n = s_n)}$$

- Un RB décompose cette loi jointe, ce qui permet de simplifier les calculs

Quelques algorithmes d'inférence

- Algorithmes exacts
  - Bucket Elimination
  - Message Passing (Pearl 88) pour les arbres
  - Junction Tree (Jensen 90)

Problème = explosion combinatoire de ces méthodes pour des graphes fortement connectés, etc . . .
(inférence = problème NP-complet)

- Algorithmes approchés
  - Echantillonnage : Markov Chain Monte Carlo, ...
  - Méthodes variationnelles

Bucket Elimination

- Principe
  - grâce à la décomposition de la loi jointe, certaines étapes de la marginalisation de $P(S_i, S_j = x)$ se simplifient

- Exemple
  - évidence $E = \{S_4 = O\}$, on cherche $P(S_2|E)$

$$P(S,E) = P(S_1)P(S_2)P(S_3|S_1S_2)P(S_4 = 0|S_2)P(S_5|S_4 = O)$$

$$P(S_2, E) = \sum_{S_1, S_3, S_5} P(S,E)$$

  - et si on choisit l'ordre des variables pour la marginalisation ?

Architecture des Systèmes d'Information

Bucket Elimination

- Commençons par S_5

$$\sum_{S_5} P(S_1, S_2, S_3, S_4 = O, S_5) = P(S_1)P(S_2)P(S_3|S_1, S_2)\ldots$$
$$\ldots \quad P(S_4 = O|S_2) \sum_{S_5} P(S_5|S_4 = O)$$

- Cette dernière somme vaut 1 ! On a éliminé S_5

$$P(S_1, S_2, S_3, S_4 = O) = P(S_1)P(S_2)P(S_3|S_1, S_2)P(S_4 = O|S_2)$$

- Au tour de S_1

Bucket Elimination

$$\sum_{S_1} P(S_1, S_2, S_3, S_4 = O) = P(S_2)P(S_4 = O|S_2)\dots$$
$$\dots \sum_{S_1} P(S_1)P(S_3|S_1, S_2)$$

- Cette dernière somme nous rend une table dépendant de S_2 et S_3 : $T(S_2, S_3)$

$$P(S_2, S_3, S_4 = O) = P(S_2)P(S_4 = O|S_2)T(S_2, S_3)$$

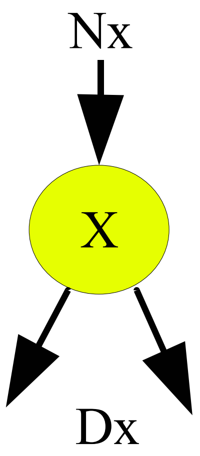
- Idem avec S_3 pour obtenir $P(S_2, S_4 = O)$

Marginalisation = série de produits locaux de matrices et de marginalisations locales

Message Passing (Pearl 1988)

- Chaque nœud envoie des messages à ses voisins
- L'algorithme ne marche que dans le cas des arbres
- (mais est généralisable au cas des poly-arbres)

- E = ensemble de variables instanciées.
  $E = N_x \cup D_x$



- 2 types de messages λ et π serviront à calculer
  - $\lambda(X) \propto P(D_x|X)$
  - $\pi(X) \propto P(X|N_x)$

- et ensuite on peut montrer que

$$\boxed{P(X|E = e) \propto \lambda(X)\pi(X)}$$

Architecture des Systèmes d'Information

Message Passing

- Les messages λ
  - Pour chaque enfant Y de X,

$$\lambda_Y(X = x) = \sum_y P(Y = y|X = x)\lambda(Y = y)$$

- Comment calculer λ en chaque nœud ?
  - Si X instancié, $\lambda(X) = [001\dots0]$
    (la position du 1 correspond à la valeur donnée à X
  - sinon
    - si X est une feuille, $\lambda(X) = [1\dots1]$
    - sinon

$$\lambda(X = x) = \Pi_{Y \in Enf(X)} \lambda_Y(X = x)$$

Architecture des Systèmes d'Information

Message Passing

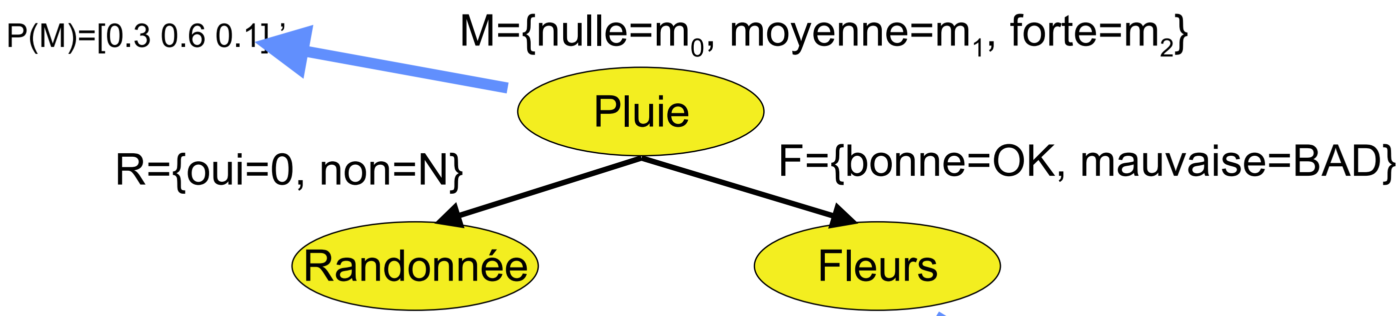
- Les messages π
  - Pour Z l'unique parent de X,

$$\pi_X(Z=z) = \pi(Z=z)\Pi_{U\in Enf(Z)\setminus\{X\}}\lambda_U(Z=z)$$

- Comment calculer π en chaque nœud ?
  - Si X instancié, $\lambda(X) = [001\ldots 0]$
    (la position du 1 correspond à la valeur donnée à X
  - sinon
    - si X est la racine, $\pi(X) = P(X)$
    - sinon

$$\pi(X=x) = \sum_z P(X=x|Z=z)\pi_X(Z=z)$$

Exemple



P(R | M)

	Pluie =		
	m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

	Pluie =		
	m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $\boxed{E = \emptyset}$

$$P(F) = ?$$

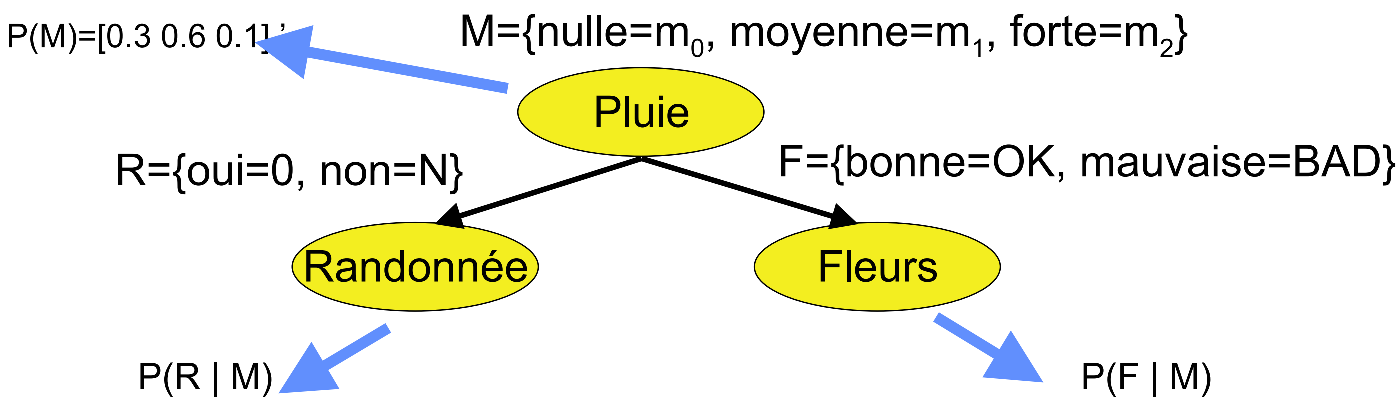
$$P(R) = ?$$

$$P(F) = \sum_m P(F|M=m)P(M=m)$$

$$\begin{bmatrix} 0.20 & 0.75 & 0.90 \\ 0.80 & 0.25 & 0.10 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.6 \\ 0.1 \end{bmatrix} = \boxed{\begin{matrix} 0.6 \\ 0.4 \end{matrix}}$$

$$\begin{bmatrix} 0.85 & 0.50 & 0.05 \\ 0.15 & 0.50 & 0.95 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.6 \\ 0.1 \end{bmatrix} = \boxed{\begin{matrix} 0.56 \\ 0.44 \end{matrix}}$$

Exemple



P(R | M)

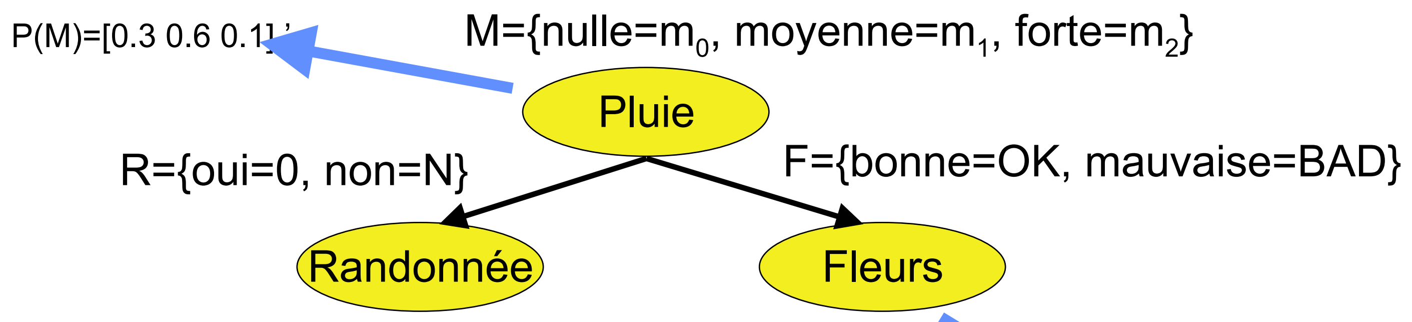
	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $\boxed{E = \{M = m_2\}}$
  - $\lambda(M) = \pi(M) = [0 \;\; 0 \;\; 1]$ (nœud instancié)
  - $P(M|E) \propto \lambda(M)\pi(M) = [0 \;\; 0 \;\; 1]$ (logique)
- messages envoyés aux enfants :
  - $\pi_R(M) = \pi(M)\lambda_F(M) = [0 \;\; 0 \;\; 1]$
  - $\pi_F(M) = \pi(M)\lambda_R(M) = [0 \;\; 0 \;\; 1]$

Exemple



$P(R \mid M)$

	Pluie =		
	m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

$P(F \mid M)$

	Pluie =		
	m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $\boxed{E = \{M = m_2\}}$ suite...
  - en R :
    - $\pi(R) = P(R|M)\pi_R(M) = [0.05 \;\; 0.95]$
    - $\lambda(R) = [1 \;\; 1]$ (feuille)
    - $P(R|E) \propto \lambda(R)\pi(R) = [0.05 \;\; 0.95]$

$$\begin{bmatrix} 0.85 & 0.50 & 0.05 \\ 0.15 & 0.50 & 0.95 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0.95 \end{bmatrix}$$

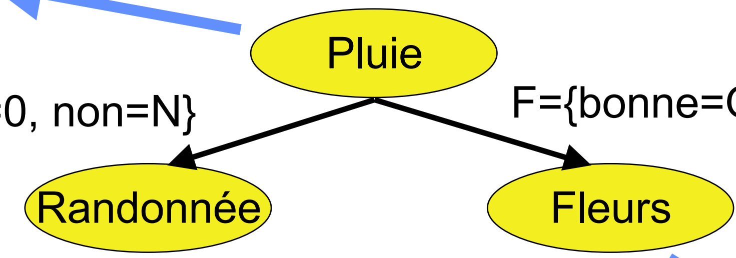
Pluie=Forte $\Rightarrow$ Randonnée=Non

Architecture des Systèmes d'Information

Exemple

P(M)=[0.3 0.6 0.1]'

M={nulle=m_0, moyenne=m_1, forte=m_2}



R={oui=0, non=N}

F={bonne=OK, mauvaise=BAD}

P(R | M)

	Pluie =		
	m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

	Pluie =		
	m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

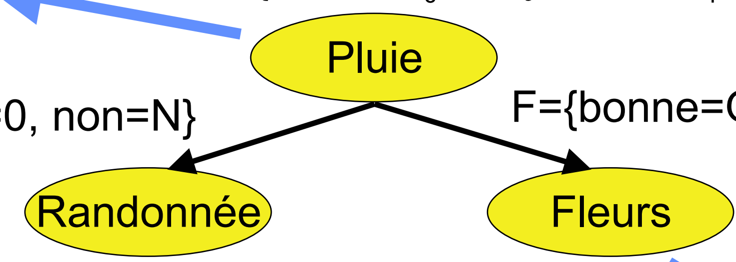
- $\boxed{E = \{M = m_2\}}$ fin
  - en F :
    - $\pi(F) = P(F|M)\pi_F(M) = [0.9 \;\; 0.1]$
    - $\lambda(F) = [1 \;\; 1]$ (feuille)
    - $P(F|E) \propto \lambda(F)\pi(F) = [0.9 \;\; 0.1]$

$\boxed{\text{Pluie=Forte} \Rightarrow \text{Fleurs arrosées}}$

Exemple

P(M)=[0.3 0.6 0.1]'

M={nulle=m_0, moyenne=m_1, forte=m_2}



P(R | M)

	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

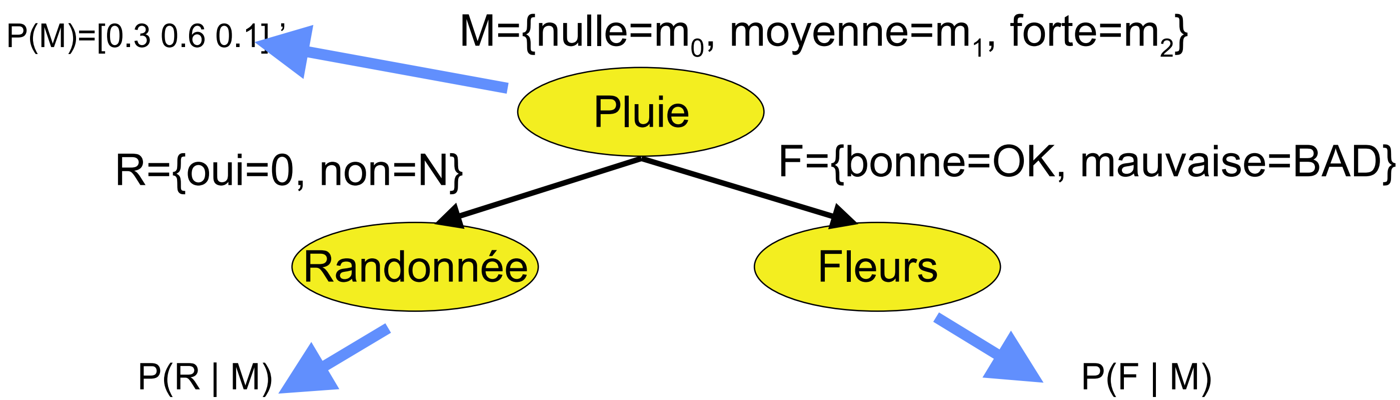
- $\boxed{E = \{F = OK(0.2) | BAD(0.8)\}}$ (soft evidence)

En F :

- $\lambda(F) = \pi(F) = [0.2 \ \ 0.8]$
- $P(F|E) \propto \lambda(F)\pi(F) = [0.2 \ \ 0.8]$
- message envoyé au parent :
  - $\lambda_F(M) = \lambda(F)P(F|M) = [0.68 \ \ 0.35 \ \ 0.26]$

$$\begin{array}{c|ccc} & 0.20 & 0.75 & 0.90 \\ & 0.80 & 0.25 & 0.10 \\ \hline 0.2 \ \ 0.8 & 0.68 & 0.35 & 0.26 \end{array}$$

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

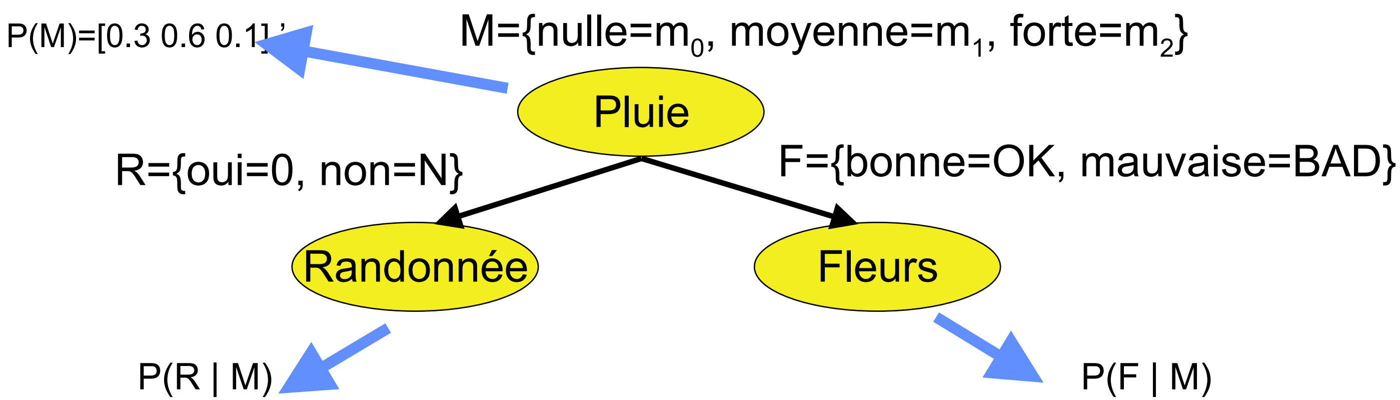
	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $\boxed{E = \{F = OK(0.2)|BAD(0.8)\}}$ (soft evidence)

En R :

- $\lambda(R) = [1 \;\; 1]$ (feuille)
- $\pi(R) = ?$
- message envoyé au parent :
  - $\lambda_R(M) = \lambda(R)P(R|M) = [1 \;\; 1 \;\; 1]$

Exemple



	Pluie = m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

	Pluie = m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

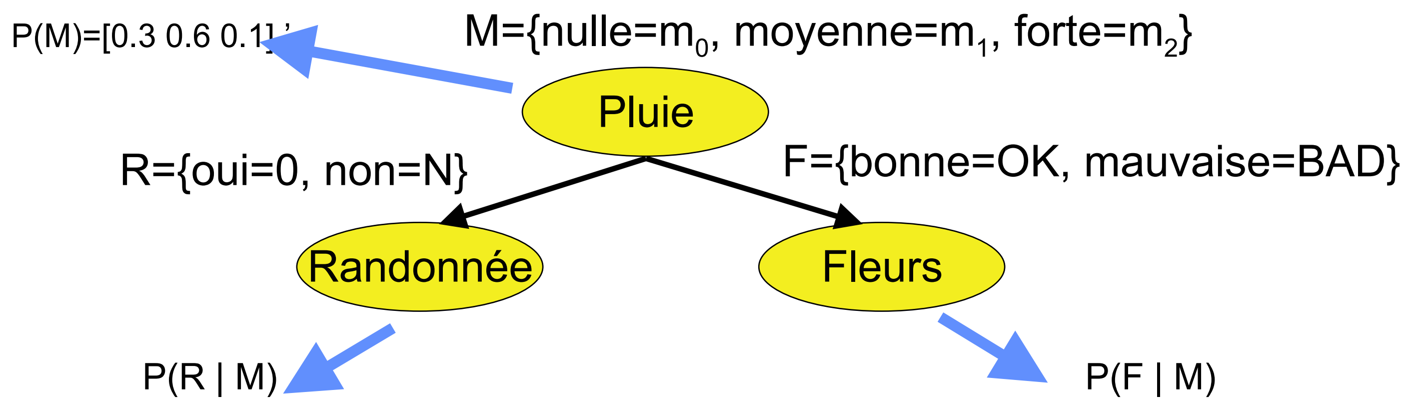
- $\boxed{E = \{F = OK(0.2)|BAD(0.8)\}}$ (soft evidence)

En M :

- $\lambda(M) = \lambda_R(M)\lambda_F(M) = [0.68 \;\; 0.35 \;\; 0.26]$
- $\pi(M) = P(M) = [0.3 \;\; 0.6 \;\; 0.1]$ (racine)
- $P(M|E) \propto \lambda(M)\pi(M) = \boxed{[0.463 \;\; 0.477 \;\; 0.060]}$
- message : $\pi_R(M) = \pi(M)\lambda_F(M) = [0.204 \; 0.216 \; 0.026]$

Architecture des Systèmes d'Information

Exemple



P(R | M)

	Pluie =		
	m_0	m_1	m_2
R=O	0.85	0.50	0.05
R=N	0.15	0.50	0.95

P(F | M)

	Pluie =		
	m_0	m_1	m_2
F=OK	0.20	0.75	0.90
F=BAD	0.80	0.25	0.10

- $\boxed{E = \{F = OK(0.2)|BAD(0.8)\}}$ (soft evidence)

Retour en R :

- $\lambda(R) = [1 \;\; 1]$ (feuille)
- $\pi(R) = P(R|M)\pi_R(M) = [0.283 \;\; 0.229]$
- $P(R|E) \propto \lambda(R)\pi(R) = \boxed{[0.553 \;\; 0.447]}$

Fleurs plutôt en mauvais état $\Rightarrow$ Randonnée = plutôt oui

Junction Tree (Jensen 1990)

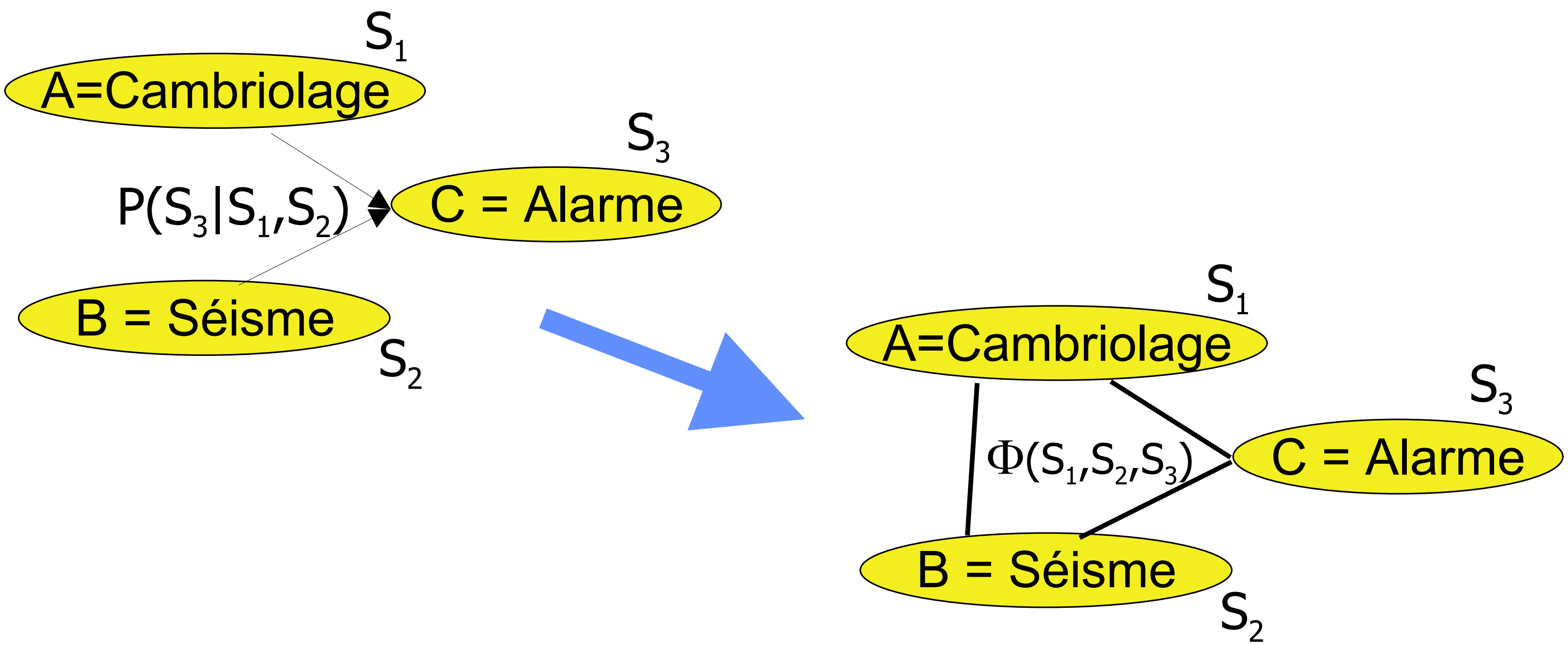
- Message Passing ne s'applique bien qu'aux arbres
- Besoin d'un algorithme plus général

- Principe
  - Transformer le graphe en un arbre (non orienté)...
  - Arbre = arbre de jonction des cliques maximales du graphe moralisé et triangulé

- Moralisation = marier les parents et "désorienter" le graphe
- Triangulation = éviter les cycles dans le graphe non orienté.

Junction Tree

- Moralisation : marier les parents de chaque nœud



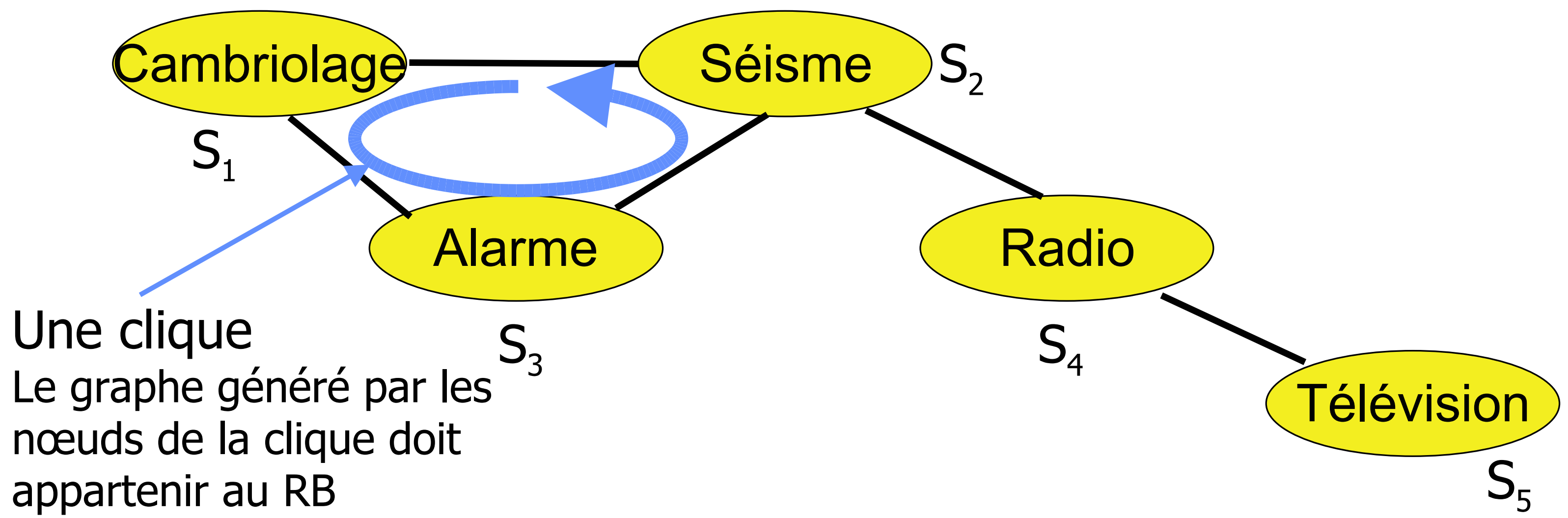
Junction Tree

- Triangulation : tout cycle de longueur au moins 4 doit contenir une corde (arête reliant deux sommets non consécutifs sur le cycle)
- (= aucun sous-graphe cyclique de longueur > 3).

- Triangulation optimale pour des graphes non-dirigés = NP-difficile (comment choisir les meilleures cordes ?)

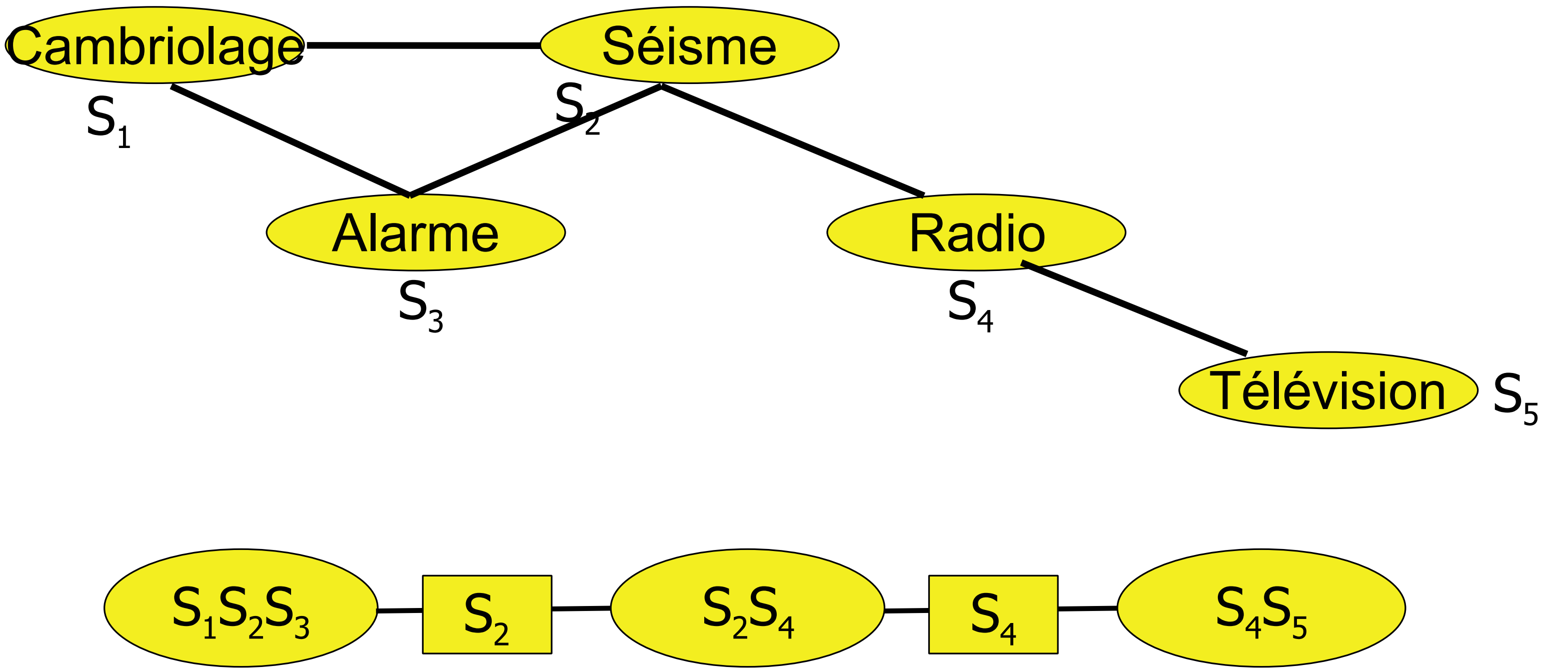
Junction Tree

- Clique = sous-graphe du RB dont les nœuds sont complétement connectés
- Clique maximale = l'ajout d'un autre nœud à cette clique ne donne pas une clique



Junction Tree

- Théorème : Si le graphe est moralisé et triangulé, alors les cliques peuvent être organisées en un arbre de jonction



$$P(S) = \Phi(S1, S2, S3)\Phi(S2, S4)\Phi(S4, S5)$$

- L'inférence se fait au niveau des Φ

Applications

- Diagnostic et raisonnement dans des systèmes complexes

- Marketing/Finance (modélisation de risques) :
  - ATT : détection de fraudes (mauvais payeurs) pour les factures de téléphone
  - Altaprofit : optimation de portefeuilles (contrats d'assurance vie)

- Informatique :
  - Microsoft : printer troubleshooting, assistant Office
  - MODIST : évaluation de la qualité pour des développements logiciels

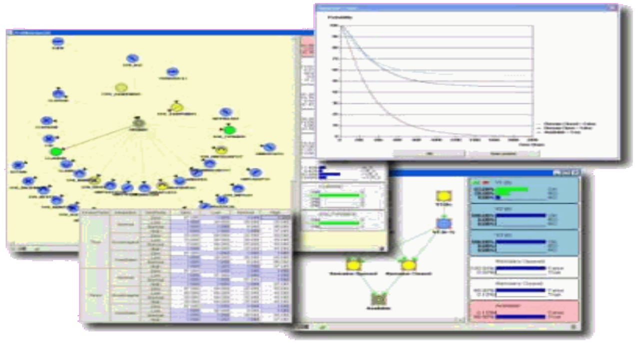
Applications

- Médecine :
  - Aide au diagnostic de problèmes cardio-vasculaires
  - Surveillance transfusionnelle, ...

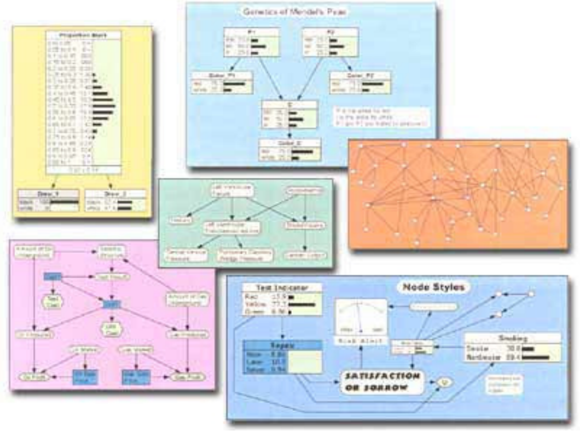
- Industrie :
  - NASA : aide au diagnostic de pannes en temps réel pour les systèmes de propulsion de la navette spatiale
  - Lockheed Martin : système de contrôle d'un véhicule sous-marin autonome
  - Ricoh : aide au télédiagnostic
  - EDF : modélisation de groupes electrogènes

Offre logicielle

- Toolbox
  - Bayes Net Toolbox (BNT) pour Matlab
  - gR, GRAPPA, ... pour $\mathcal{R}$
  - BNJ, JavaBayes, ... pour Java
- Logiciels non commerciaux
  - Microsoft Belief Network [US]
  - Genie 2/Smile [US]
- Logiciels commerciaux
  - Bayesia [FR]
  - ProBT (inférence probabiliste) [FR]
  - Hugin [DK]
  - Netica [CA]



A Complete Software Package to Solve Problems Using Bayesian Belief Networks and Influence Diagrams



Références

- **Les Réseaux Bayésiens** - P. Naïm, P.H. Wuillemin, Ph. Leray, O. Pourret, A. Becker (Eyrolles)
- **Probabilistic reasoning in Intelligent Systems : Networks of plausible inference** - J. Pearl (Morgan Kaufman)
- **An introduction to Bayesian Networks** - F. Jensen (Springer Verlag)
- **Probabilistic Networks and Expert Systems** - R.G. Cowell & al. (Springer Verlag)
- **Learning Bayesian Networks** - R. Neapolitan (Prenctice Hall)
- **Learning in Graphical Models** - Jordan M.I. ed. (Kluwer)

Architecture des Systèmes d'Information