Bayesian Networks - I : Definition and probabilistic inference

Philippe LERAY
philippe.leray@univ-nantes.fr

DUKe (Data User Knowledge) Research group
Laboratoire des Sciences du Numérique de Nantes – UMR 6004
Site de l'Ecole Polytechnique de l'université de Nantes

Definition
○○○○○○○
Basic concepts
○○○○○○○○○○○○○○○○○

Inference
○○○○○○○○○

Reminders of basic probabilistic theory

Reminders of basic probabilistic theory

Conditional probability

- let A and M denote two events
- a priori information about A : $\qquad\qquad\qquad\qquad P(A)$

- M happened : $\qquad\qquad\qquad\qquad\qquad P(M) \neq 0$

- if there is a link between A and M, this event will modify our knowledge about A
- a posteriori information : $\qquad\qquad\quad P(A|M) = \frac{P(A,M)}{P(M)}$

Definition
○●○○○○○

Basic concepts
○○○○○○○○○○○○○○○○○

Inference
○○○○○○○○○

Reminders of basic probabilistic theory

Reminders of basic probabilistic theory

Independence

- A and B are independent iff :
  $P(A, B) = P(A) \times P(B)$
  $P(A|B) = P(A)$
  $P(B|A) = P(B)$

Conditional independence

- A and B are independent conditionally to C iff :
  $P(A|B, C) = P(A|C)$

Definition
○○●○○○○

Basic concepts
○○○○○○○○○○○○○○○○○

Inference
○○○○○○○○○

Reminders of basic probabilistic theory

Reminders of basic probabilistic theory

$\{M_i\}$ complete set of mutually exclusive events

Marginalization :
$$P(A) = \sum_i P(A, M_i)$$

Total probability theorem

Event A can result from various causes M_i. What is the probability of A if we know :

- *the prior probabilities $P(Mi)$*
- *the conditional probabilities of A given each Mi*

$$P(A) = \sum_i P(A|M_i)P(M_i)$$

Definition
○○○●○○○

Basic concepts
○○○○○○○○○○○○○○○○○○

Inference
○○○○○○○○○

Reminders of basic probabilistic theory

Reminders of basic probabilistic theory

$\{M_i\}$ complete set of mutually exclusive events

Bayes' theorem

Event A happened. What is the probability that the cause M_i is responsible of this event ?

$$P(M_i|A) = \frac{P(A|M_i) \times P(M_i)}{P(A)}$$

- $P(Mi|A)$: a posteriori probability
- $P(A)$: constant w.r.t. Mi (cf. Total probability theorem)

Chain rule

$$P(A_1 \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\ldots P(A_n|A_1 \ldots A_{n-1})$$

Definition
○○○○●○○
Basic concepts
○○○○○○○○○○○○○○○○○
Inference
○○○○○○○○○

Reminders of basic probabilistic theory

Example

Bayesian network definition

Theoretical principle

- taking into account some extra knowledge (conditional independence between some variables) to simplify the joint probability distribution given by the chain rule.

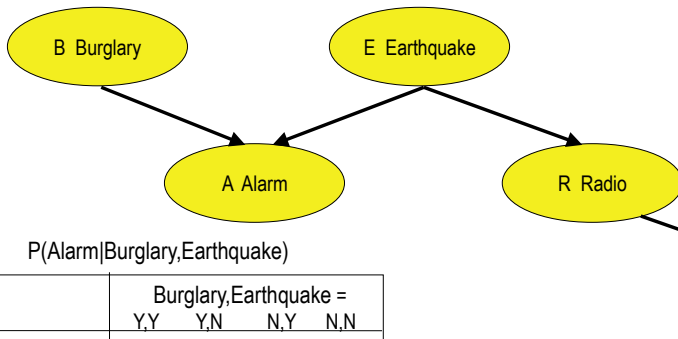### Definition   [Pearl, 1985]

- a Bayesian network (BN) is defined by
  - one qualitative description of (conditional) dependences/independences between variables
    directed acyclic graph (DAG)
  - one quantitative description of these dependences
    conditional probability distributions (CPDs)

Example

one topological order : B, E, A, R, T (not unique)

P(Radio|Earthquake)

	Earthquake =	
	Y	N
Radio=Y	0.99	0.01
Radio=N	0.01	0.99

P(Burglary)=[0.001 0.999]   P(Earthquake)=[0.0001 0.9999]

B Burglary   E Earthquake

P(TV|Radio)

	Radio =	
	Y	N
TV=Y	0.99	0.50
TV=N	0.01	0.50

A Alarm   R Radio

T TV

P(Alarm|Burglary,Earthquake)

	Burglary,Earthquake =			
	Y,Y	Y,N	N,Y	N,N
Alarm=Y	0.75	0.10	0.99	0.10
Alarm=N	0.25	0.90	0.01	0.90

Definition
○○○○○○○
Basic concepts
●○○○○○○○○○○○○○○○○○
Inference
○○○○○○○○○

BN as a dependence model

BN as a dependence model

Dependence is a symmetrical relationship, so why using directed edges ?

Example with 3 nodes

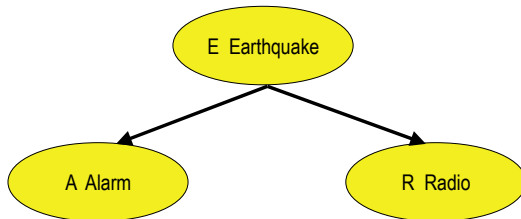- 3 simple structures between A, B and C :
    - $A \rightarrow C \rightarrow B$: serial connexion
    - $A \leftarrow C \rightarrow B$: divergent connexion
    - $A \rightarrow C \leftarrow B$: convergent connexion (V-structure)

Serial connexion

- E and T are dependent

- E and T are independent conditionally to R
  - if R is known, T will not give any new information about E

  - $P(T|E, R) = P(T|R) = P(T|parents(T))$

Definition
○○○○○○○

Basic concepts
○○○●○○○○○○○○○○○○

Inference
○○○○○○○○○

Divergent connexion

Divergent connexion

- A and R are dependent

- A and R are independent conditionally to E
  - if E is known, A will not give any new information about R

  - $P(R|A, E) = P(R|E) = P(R|parents(R))$

Definition
○○○○○○○

Basic concepts
○○○●○○○○○○○○○○○○○

Inference
○○○○○○○○○

V-structure

Convergent connexion – V-structure
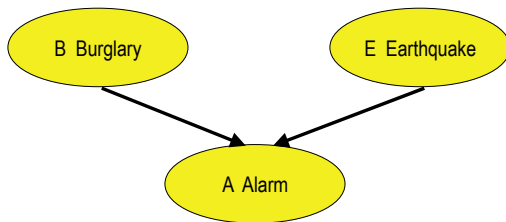
- B and E are independent

- B and E are dependent conditionally to A
    - if A is known, E will give some new information about B

    - $P(A|B, E) = P(A|parents(A))$

Consequence

Chain rule

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

Consequence with a BN

- $P(S_i|S_1 \ldots S_{i-1}) = P(S_i|parents(S_i))$ so

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

- the (global) joint probability distribution is decomposed in a product of (local) conditional distributions

- BN = compact representation of the joint distribution $P(S)$ given some information about dependence relationships between variables

Definition
○○○○○○○

Basic concepts
○○○○○●○○○○○○○○○○○○

Inference
○○○○○○○○○

Example

Example
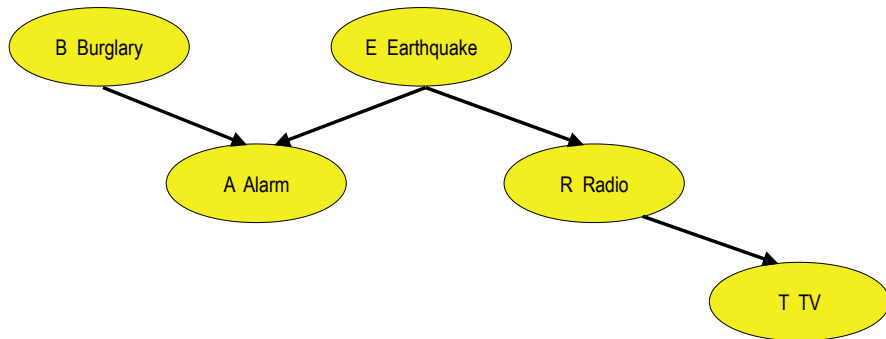
$P(B, E, A, R, T) =$
$P(B) \times P(E|B) \times P(A|B, E) \times P(R|B, E, A) \times P(T|B, E, A, R)$
$P(B) \times \ P(E) \times P(A|B, E) \times \quad P(R, E) \quad \ \times P(T|R)$

Definition
○○○○○○○

Basic concepts
○○○○○○●○○○○○○○○○

Inference
○○○○○○○○○

D-separation

D-separation

Principle

- determine if two variables in the DAG are (conditionally) independent (given some instantiated variables)

Definition

- two variables A and B are d-separated if for all undirected path between A and B, there exist one intermediate variable V different from A and B such as one of the following properties is true :
    - the connexion is serial or divergent in V and V is instantiated
    - the connexion is convergent and neither V nor its descendants are instantiated
- if A and B are not d-separated, they are d-connected

Definition
○○○○○○○
Basic concepts
○○○○○○○●○○○○○○○○○
Inference
○○○○○○○○○

Bayes ball

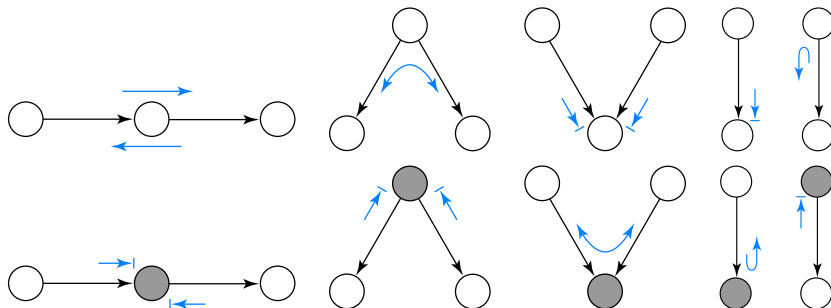## Bayes ball   [Shachter, 1998]

Principle

- are *A* and *B* independent given a set of variables *Z* ?
- shade all nodes instantiated (*Z*)
- if the Bayes ball cannot reach *B* from *A*, they are conditionally independent
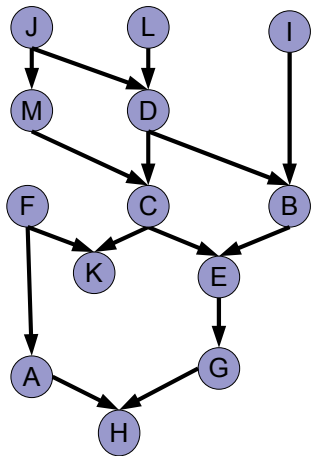- balls can travel opposite to edge directions

Definition
○○○○○○○

Basic concepts
○○○○○○○○●○○○○○○○○

Inference
○○○○○○○○○

Bayes ball

Bayes ball   (Shachter 1998)

The 10 rules for Bayes ball playing :

$\rightarrow\!|$ = ball is blocked

image from M. Paskin course on Graphical Models

Definition
0000000

Basic concepts
000000000●00000000

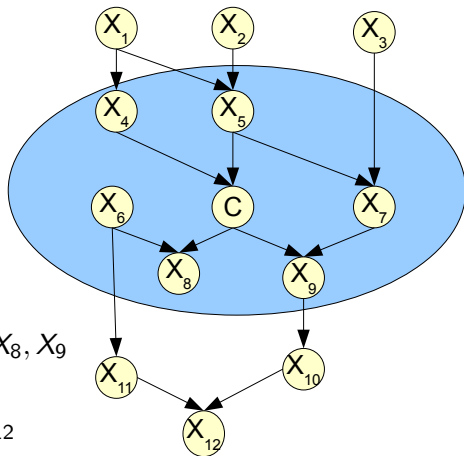Inference
000000000

Example

Example

Markov blanket

Definition

- the Markov blanket MB of node C contains all the variables that shield the node from the rest R of the network, i.e. C is independent of every subset of R conditionally to MB

Graphical interpretation

- MB = set of children, parents and spouses

Markov blanket

- $MB = X_4, X_5, X_6, X_7, X_8, X_9$
- C is independent of $X_1, X_2, X_3, X_{10}, X_{11}, X_{12}$ conditionally to MB
- $P(C|X_1...X_{12}) = P(C|MB)$

Definition
○○○○○○○

Markov equivalence

Basic concepts
○○○○○○○○○○○○●○○○○○
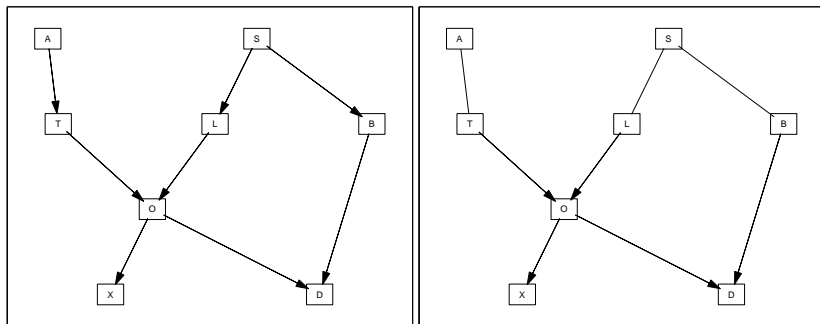
Inference
○○○○○○○○○

Markov equivalence

Definition

B_1 and B_2 are Markov equivalent iff both describe exactly the same conditional (in)dependence statements.

Graphical properties

- B_1 and B_2 have the same skeleton, V-structures and inferred edges.
- all the equivalent graphs (= equivalence class) can be summarized by one partially directed DAG named CPDAG or Essential Graph

Definition
○○○○○○○

Basic concepts
○○○○○○○○○○○○○○●○○○○

Inference
○○○○○○○○○

Example

Markov equivalence

Definition
○○○○○○○

Basic concepts
○○○○○○○○○○○○○○●○○○

Inference
○○○○○○○○○

Faithfulness

Faithfulness

Definition

a Bayesian network structure G and an associated probability distribution P are faithful to one another if and only if every conditional independence relationship valid in P can be read in G

Very simple counterexample

- $G = X_1 \longrightarrow X_2$
- $P(X_2|X_1 = 0) = P(X_2|X_1 = 1) = [0.8 \quad 0.2]$
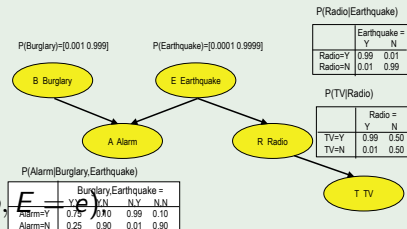- X_1 and X_2 are dependent in G but independent in P

BN as a generative model

Principle

- BN = compact representation of the joint distribution $P(S)$
- we can use classical sampling methods to generate data from this distribution

Example : forward sampling

- if $rand1 < P(B = Y)$, $B = Y$, else N
- if $rand2 < P(E = Y)$, $E = Y$, else N
- if $rand3 < P(A = Y | B = b, E = e)$, $A = Y$, else N

P(Burglary)=[0.001 0.999]   P(Earthquake)=[0.0001 0.9999]

P(Radio|Earthquake)

	Earthquake =	
	Y	N
Radio=Y	0.99	0.01
Radio=N	0.01	0.99

P(TV|Radio)

	Radio =	
	Y	N
TV=Y	0.99	0.50
TV=N	0.01	0.50

P(Alarm|Burglary,Earthquake)

	Burglary,Earthquake =			
	Y,Y	Y,N	N,Y	N,N
Alarm=Y	0.75	0.90	0.99	0.10
Alarm=N	0.25	0.90	0.01	0.90

Definition
○○○○○○○

Basic concepts
○○○○○○○○○○○○○○○●○

Inference
○○○○○○○○○

BN as a generative model

Example - Minesweeper

Principle

- bombs are places in a grid
- each square (i, j) independently has a bomb ($B_{i,j} = true$) with probability b
- what you can observe for a given square is a reading $N_{i,j}$ of the number of bombs in adjacent squares (not including the square itself)

Definition
○○○○○○○

Basic concepts
○○○○○○○○○○○○○○○○●

Inference
○○○○○○○○○

BN as a generative model

Example - Minesweeper

Bayesian network ?

- draw a Bayesian network for a one-dimensional 4x1 Minesweeper grid, showing all eight variables ($B_1 \ldots B_4$ and $N_1 \ldots N_4$). Show the minimal set of arcs needed to correctly model the domain above

- fully specify the CPTs for each variable, assuming that there is no noise in the readings (i.e. that the number of adjacent bombs (or bomb) is reported exactly, deterministically). Your answers may use the bomb rate b if needed

- what are the posterior probabilities of bombs B_i in each of the four squares, given no information? If we observe $N_2 = 1$, what are the posterior probabilities of bombs in each square?

- check your model with Genie/Smile

What is probabilistic inference ?

Inference

- computation of any $P(S_i|\{S_j = x\})$ (NP-hard)
- evidence ε = set of observable variables $\{S_j = x\}$

Exact inference algorithms

- Message Passing (Pearl 1988) for trees or poly-trees
- Junction Tree (Jensen 1990)
- Shafer-Shenoy (1990)

Problem = combinatorial explosion for strongly connected graphs.

Approximate inference algorithms

- sampling

What is probabilistic inference ?

Inference

Exact inference algorithms

- Message Passing (Pearl 1988) for trees or poly-trees
- Junction Tree (Jensen 1990)
- Shafer-Shenoy (1990)

Problem = combinatorial explosion for strongly connected graphs.

Approximate inference algorithms

- sampling
- variational methods

What is probabilistic inference ?

Inference

Exact inference algorithms

- Message Passing (Pearl 1988) for trees or poly-trees
- Junction Tree (Jensen 1990)
- Shafer-Shenoy (1990)

Problem = combinatorial explosion for strongly connected graphs.

Approximate inference algorithms

- sampling
- variational methods

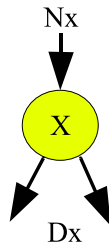## Message Passing   (Pearl 1988)

Principle

- designed for tree structures (generalized to poly-trees)
- every node send messages to its parent and children

- $\varepsilon =$ set of instantiated/observed variables.
  $\varepsilon = N_x \cup D_x$ instantiated (non) descendants of X

- we can demonstrate that $P(X|\varepsilon = e) \propto \lambda(X)\pi(X)$
  with $\lambda(X) \propto P(D_x|X)$ and $\pi(X) \propto P(X|N_x)$

- 2 types of messages $\overrightarrow{\lambda}$ and $\overrightarrow{\pi}$ will help to compute these λ
  and π values for every X

Definition   Basic concepts   **Inference**
○○○○○○○   ○○○○○○○○○○○○○○○○○   ○○●○○○○○○
Lambda messages

Message Passing : $\overrightarrow{\lambda}$ messages

$\lambda(X) \propto P(D_x|X)$ information from descendants

$\lambda(X)$ **initialization**

- if X is an unobserved leaf : $\lambda(X) = [1 \dots 1]$ (no information)
- if X is an observed node : $\lambda(X) = [001 \dots 0]$ (exact info.)
  (1 at i-th position corresponds to the observed value $X = i$)

$\overrightarrow{\lambda}$ **propagation and aggregation**

- for every child Y of X,

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

- aggregation : $\lambda(X = x) = \Pi_{Y \in Child(X)}\overrightarrow{\lambda_Y}(X = x)$

Message Passing : $\overrightarrow{\pi}$ messages

$\pi(X) \propto P(X|N_x)$ information from non descendants

$\pi(X)$ initialization

- if X is the unobserved root : $\pi(X) = P(X)$ (a priori info.)
- if X is an observed node : $\pi(X) = [001 \ldots 0]$ (exact info.)

$\overrightarrow{\pi}$ propagation and aggregation

- for Z, unique parent of X,

$$\overrightarrow{\pi_X}(Z = z) = \pi(Z = z) \prod_{U \in Child(Z) \setminus \{X\}} \overrightarrow{\lambda_U}(Z = z)$$

- aggregation : $\pi(X = x) = \sum_z P(X = x|Z = z)\overrightarrow{\pi_X}(Z = z)$
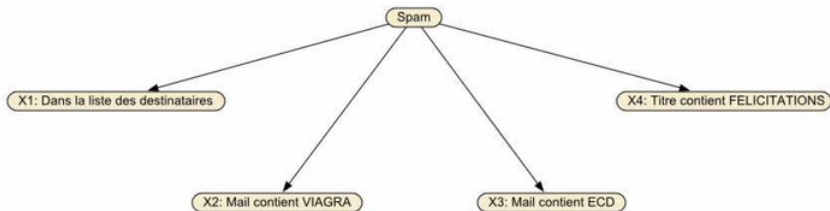
Message Passing complexity

Worst case

- each node send 2 messages : time complexity is linear in the number of nodes
- work done in a node is proportional to the size of its CPD : linear for trees, but need to be bounded for poly-trees

Example

Bayesian anti-spam

Variables

- *Spam* : this mail is a spam = yes / no,
- X_1 : addressee of the mail ? = only one / among N / not,
- X_2 : this mail contains VIAGRA = yes / no,
- X_3 : this mail contains POLYTECH = yes / no,
- X_4 : this mail has a title containing CONGRATULATIONS = yes / no.

Bayesian anti-spam

Parameters

some statistics are computed with the help of one dataset
composed with 100 spams and 400 legitimate mails :

- you are rarely the only addressee of a spam (2%). It's more
  frequent not being one of the addressees (78%). These
  percentages are inverted when the mail is not a spam.

- 1% of spam mails contains POLYTECH, and 10% contain
  VIAGRA. 30% of spam mails can be detected with their title
  containing CONGRATULATIONS.

- 30% of legitimate mails contain POLYTECH, 0.1% contain
  VIAGRA, and 1% have a title containing
  CONGRATULATIONS.

Bayesian anti-spam

Message Passing

- what is the probability of spam if you are the only one addressee ? if you are one of the addressees ? if you do not appear on the addressee list ?

- what is the probability of spam if you are one of the addressees and if the text contains VIAGRA?

- same question if you add the fact that POLYTECH doesn't appear in the mail and the title contains CONGRATULATIONS.

- check your results with Genie/Smile