COMP9414: Artificial Intelligence

Lecture 5a: Uncertainty

Wayne Wobcke

e-mail:w.wobcke@unsw.edu.au

Reasoning with Uncertainty

- An agent can not always ascertain the truth of all propositions, so may not only have "flat out" beliefs (P or $\neg P$)
- Some environments themselves generate uncertainty for the agent, due to unpredictability or nondeterminism, so propositions inadequately model those environments
- Rational decisions for an agent require tradeoffs between the importance of goals and the likelihood of achieving them, and the cost of acting and not achieving them

This Lecture

- Uncertainty
- Probability Theory
- Conditional Probability and Bayes' Rule
- Bayesian Networks
  - Semantics of Bayesian Networks
  - Inference in Bayesian Networks

Problems with Logical Approach

- Consider trying to formalize a medical diagnosis system

$$\forall p(Symptom(p, AbdominalPain) \rightarrow Disease(p, Appendicitis))$$

- This rule is not correct since patients with abdominal pain may be suffering from other diseases

$$\forall p(Symptom(p, AbdominalPain) \rightarrow$$

$$Disease(p, Appendicitis) \vee Disease(p, Ulcer) \vee Disease(p, Indig) \cdots)$$

- How about a causal rule?

$$\forall p(Disease(p, Ulcer) \rightarrow Symptom(p, AbdominalPain))$$

Sources of Uncertainty

- Difficulties arise with the logical approach due to

  incompleteness: agent may not have complete theory for domain

  ignorance: agent may not have enough information about domain

  noise: information agent does have may be unreliable

  nondeterminism: environment itself may be stochastic

  unpredictability: environment may be inherently unpredictable

- Probability gives a way of summarizing this uncertainty
  - e.g. agent believes that there is a probability of 0.75 that patient suffers from appendicitis if they have abdominal pains

What Do the Numbers Mean?

Statistical/Frequentist View

Long-range frequency of a set of "events" e.g. probability of the event of "heads" appearing on the toss of a coin = long-range frequency of heads that appear on coin toss

Objective View

Probabilities are real aspects of the world – objective

Personal/Subjective/Bayesian View

Measure of belief in proposition based on agent's knowledge, e.g. probability of heads is a degree of belief that coin will land heads based on beliefs about the coin or could be just a guess; different agents may assign a different probability – subjective

Sample Space and Events

- Flip a coin three times
- The possible outcomes are

  TTT TTH THT THH

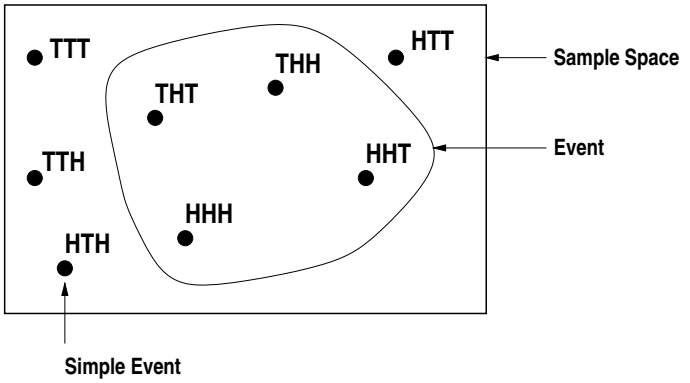
  HTT HTH HHT HHH

- Set of all possible outcomes

  $S = \{\texttt{TTT}, \texttt{TTH}, \texttt{THT}, \texttt{THH}, \texttt{HTT}, \texttt{HTH}, \texttt{HHT}, \texttt{HHH}\}$

- Any subset of the sample space is known as an event
- Any singleton subset of the sample space is known as a simple event

Sample Space and Events



Prior Probability

- $P(A)$ is the prior or unconditional probability that an event A occurs
- For example, $P(Appendicitis) = 0.3$
- In the absence of any other information, agent believes there is a probability of 0.3 (30%) that the patient suffers from appendicitis
- To account for the effect of new information on probabilities, the agent must reason with conditional probabilities to update probabilities

Random Variables

- Propositions are random variables that can take on several values
  $P(Weather = Sunny) = 0.8$
  $P(Weather = Rain) = 0.1$
  $P(Weather = Cloudy) = 0.09$
  $P(Weather = Snow) = 0.01$
- Every random variable X has a domain of possible values $\langle x_1, x_2, \cdots, x_n \rangle$
- Probabilities of all possible values $\mathbf{P}(Weather) = \langle 0.8, 0.1, 0.09, 0.01 \rangle$ is a probability distribution
- $\mathbf{P}(Weather, Appendicitis)$ is a combination of random variables represented by cross product (can also use logical connectives $P(A \wedge B)$ to represent compound events)

Axioms of Probability

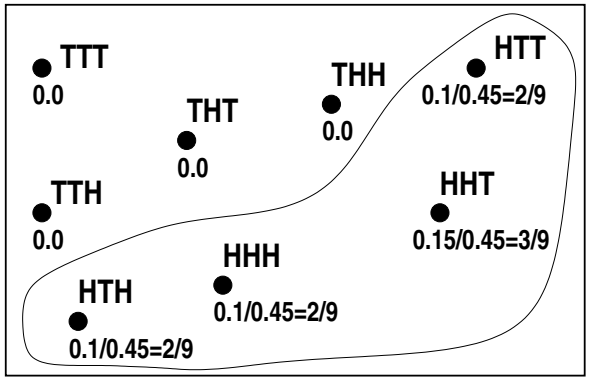
1. $0 \leq P(A) \leq 1$
   - All probabilities are between 0 and 1
2. $P(True) = 1 \qquad P(False) = 0$
   - Valid propositions have probability 1
   - Unsatisfiable propositions have probability 0
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

- Can determine probabilities of all other propositions
- For example, $P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$
  $P(True) = P(A) + P(\neg A) - P(False)$
  $1 = P(A) + P(\neg A) - 0$
  Therefore $P(\neg A) = 1 - P(A)$

Conditional Probability

- Need to update probabilities based on new information
- Use conditional or posterior probability
  $P(A|B)$ is the probability of A given that all we know is B
  - e.g. $P(Appendicitis|AbdominalPain) = 0.75$
- Definition: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$ provided $P(B) > 0$
- Product Rule: $P(A \wedge B) = P(A|B).P(B)$
- $\mathbf{P}(X|Y) = P(X = x_i|Y = y_j)$ for all i, j
  $\mathbf{P}(X,Y) = \mathbf{P}(X|Y).\mathbf{P}(Y)$ – a set of equations

Normalization



- Conditional probability distribution given that first coin is `H`

Joint Probability Distribution

- Complete specification of probabilities to all events in domain
- Suppose random variables $X_1, X_2, \cdots, X_n$
- An atomic (simple) event is an assignment of values to all variables
- Joint probability distribution $\mathbf{P}(X_1, X_2, \cdots, X_n)$ assigns probabilities to all possible atomic events
- Simple medical domain with two Boolean random variables

	AbdominalPain	¬*AbdominalPain*
Appendicitis	0.04	0.06
¬*Appendicitis*	0.01	0.89

Joint Probability Distribution

- Simple events are mutually exclusive and jointly exhaustive
- Probability of complex event is sum of probabilities of compatible simple events

  $P(Appendicitis) = 0.04 + 0.06 = 0.10$

  $P(Appendicitis \vee AbdominalPain) = 0.04 + 0.06 + 0.01 = 0.11$

  $P(Appendicitis | AbdominalPain)$

  $= \frac{P(Appendicitis \wedge AbdominalPain)}{P(AbdominalPain)} = \frac{0.04}{0.04+0.01} = 0.8$
- Problem: With many variables, the number of probabilities is vast

Joint Probability Distribution

Assume there is some underlying joint probability distribution over three random variables *toothache*, *cavity* and *catch*, which we can write in the form of a table

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Note that the sum of the entries in the table is 1.0

For any proposition, sum the atomic events where it is true

Inference by Enumeration Example

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition, sum the atomic events where it is true

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by Enumeration Example

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition, sum the atomic events where it is true

$P(cavity \vee toothache)$
$= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Conditional Probability by Enumeration

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

$$P(\neg cavity|toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Bayes' Rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- AI systems abandon joint probabilities and work directly with conditional probabilities using Bayes' Rule
- Deriving Bayes' Rule:
  $P(A \wedge B) = P(A|B)P(B)$ (Definition)
  $P(B \wedge A) = P(B|A)P(A)$ (Definition)
  So $P(A|B)P(B) = P(B|A)P(A)$ since $P(A \wedge B) = P(B \wedge A)$
  Hence $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ if $P(A) \neq 0$
- Note: If $P(A) = 0$, $P(B|A)$ is undefined

Applying Bayes' Rule

- Example (Russell & Norvig, 1995)
- Doctor knows that
  - meningitis causes a stiff neck 50% of the time
  - chance of patient having meningitis is $\frac{1}{50000}$
  - chance of patient having a stiff neck is $\frac{1}{20}$
- $P(StiffNeck|Meningitis) = 0.5$
  $P(Meningitis) = \frac{1}{50000}$
  $P(StiffNeck) = \frac{1}{20}$
  $P(Meningitis|StiffNeck) = \frac{P(StiffNeck|Meningitis).P(Meningitis)}{P(StiffNeck)}$
  $= 0.5\frac{1}{50000}\frac{1}{\frac{1}{20}} = 0.0002$

Normalization

- Avoiding assessment of symptoms
  $P(Meningitis|StiffNeck) = \frac{P(StiffNeck|Meningitis).P(Meningitis)}{P(StiffNeck)}$
  $P(\neg Meningitis|StiffNeck) = \frac{P(StiffNeck|\neg Meningitis).P(\neg Meningitis)}{P(StiffNeck)}$
- $P(StiffNeck) = P(StiffNeck|Meningitis).P(Meningitis)$
  $+ P(StiffNeck|\neg Meningitis).P(\neg Meningitis)$
- So $P(Meningitis|StiffNeck) =$
  $\frac{P(StiffNeck|Meningitis).P(Meningitis)}{P(StiffNeck|Meningitis).P(Meningitis)+P(StiffNeck|\neg Meningitis).P(\neg Meningitis)}$
  Similarly for $P(\neg Meningitis|StiffNeck)$
- Therefore, from both $P(StiffNeck|\cdots)$ can derive $P(StiffNeck)$ and the denominator is a normalization factor

Using Bayes' Rule

- Suppose there are two conditional probabilities for appendicitis
  $P(Appendicitis|AbdominalPain) = 0.8$
  $P(Appendicitis|Nausea) = 0.1$
- $P(Appendicitis|AbdominalPain \wedge Nausea)$
  $= \frac{P(AbdominalPain \wedge Nausea|Appendicitis).P(Appendicitis)}{P(AbdominalPain \wedge Nausea)}$
- Need to know $P(AbdominalPain \wedge Nausea|Appendicitis)$
- With many symptoms that is a daunting task ...

Conditional Independence

- Appendicitis is direct cause of both abdominal pain and nausea
- If we know a patient is suffering from appendicitis, the probability of nausea should not depend on the presence of abdominal pain; likewise probability of abdominal pain should not depend on nausea
- Nausea and abdominal pain are conditionally independent given appendicitis
- An event X is independent of event Y, conditional on background knowledge K, if knowing Y does not affect the conditional probability of X given K:
  $$P(X|K) = P(X|Y,K)$$

Bayesian Networks

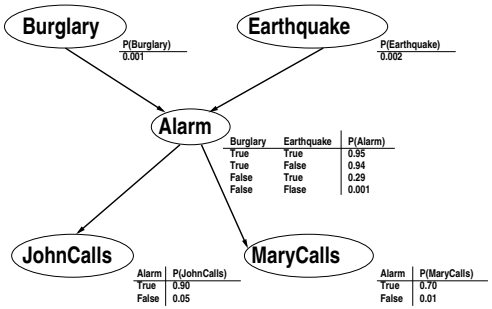
- A Bayesian network (also Bayesian Belief Network, probabilistic network, causal network, knowledge map) is a directed acyclic graph (DAG) where
  - Each node corresponds to a random variable
  - Directed links connect pairs of nodes – a directed link from node X to node Y means that X has a direct influence on Y
  - Each node has a conditional probability table quantifying effect of parents on node
- Independence assumption of Bayesian networks
  - Each random variable is (conditionally) independent of its nondescendants given its parents

Bayesian Networks

- Example (Pearl, 1988)
- You have a new burglar alarm at home that is quite reliable at detecting burglars but may also respond at times to an earthquake. You also have two neighbours, John and Mary, who promise to call you at work when they hear the alarm. John always calls when he hears the alarm but sometimes confuses the telephone ringing with the alarm and calls then, also Mary likes loud music and sometimes misses the alarm. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Bayesian Networks

- Example (Pearl, 1988)



- Probabilities summarize potentially infinite set of possible circumstances

Conditional Probability Table

- Row contains conditional probability of each node value for a conditioning case (possible combination of values for parent node)

		$\mathbf{P}(Alarm \mid Burglary \wedge Earthquake)$	
Burglary	*Earthquake*	True	False
True	True	0.950	0.050
True	False	0.940	0.060
False	True	0.290	0.710
False	False	0.001	0.999

Semantics of Bayesian Networks

- Bayesian network provides a complete description of the domain
- Joint probability distribution can be determined from the network
  - $P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$
- For example, $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J|A).P(M|A).P(A|\neg B \wedge \neg E).P(\neg B).P(\neg E) = 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628$
- Bayesian network is a complete and non-redundant representation of domain (and can be far more compact than joint probability distribution)

Semantics of Bayesian Networks

- Factorization of joint probability distribution
- Chain Rule: Use conditional probabilities to decompose conjunctions $P(X_1 \wedge X_2 \wedge \cdots \wedge X_n) = P(X_1).P(X_2|X_1).P(X_3|X_1 \wedge X_2). \cdots .P(X_n|X_1 \wedge X_2 \wedge \cdots \wedge X_{n-1})$
- Now, order the variables $X_1, X_2, \cdots, X_n$ in a Bayesian network so that a variable comes after its parents – let π_{X_i} be the tuple of parents of variable X_i (this is a complex random variable)

  Using the chain rule, $P(X_1 \wedge X_2 \wedge \cdots \wedge X_n) = P(X_1).P(X_2|X_1).P(X_3|X_1 \wedge X_2). \cdots .P(X_n|X_1 \wedge X_2 \wedge \cdots \wedge X_{n-1})$

Semantics of Bayesian Networks

- Each $P(X_i|X_1 \wedge X_2 \wedge \cdots \wedge X_{i-1})$ has the property that it is not conditioned on a descendant of X_i (given the ordering of variables in the Bayesian network)
- Therefore, by conditional independence
  - $P(X_i|X_1 \wedge X_2 \wedge \cdots \wedge X_{i-1}) = P(X_i|\pi_{X_i})$
- Rewriting gives the chain rule
  - $P(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P(X_i|\pi_{X_i})$

Calculation using Bayesian Networks

- **Fact 1:** Consider random variable X with parents $Y_1, Y_2, \cdots, Y_n$

  $$P(X|Y_1 \wedge \cdots \wedge Y_n \wedge Z) = P(X|Y_1 \wedge \cdots \wedge Y_n)$$

  if Z doesn't involve a descendant of X (including X itself)
- **Fact 2:** If $Y_1, \cdots, Y_n$ are pairwise disjoint and exhaust all possibilities

  $$P(X) = \Sigma P(X \wedge Y_i) = \Sigma P(X|Y_i).P(Y_i)$$

  $$P(X|Z) = \Sigma P(X \wedge Y_i|Z)$$

  - e.g. $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\Sigma P(J \wedge B \wedge e \wedge a \wedge m)}{\Sigma P(j \wedge B \wedge e \wedge a \wedge m)}$ where j ranges over $J, \neg J$, e over $E, \neg E$, a over $A, \neg A$ and m over $M, \neg M$

Calculation using Bayesian Networks

- $P(J \wedge B \wedge E \wedge A \wedge M) = P(J|A).P(B).P(E).P(A|B \wedge E).P(M|A) = 0.90 \times 0.001 \times 0.002 \times 0.95 \times 0.70 = 0.000001197$
- $P(J \wedge B \wedge \neg E \wedge A \wedge M) = 0.0005910156$
- $P(J \wedge B \wedge E \wedge \neg A \wedge M) = 5 \times 10^{-11}$
- $P(J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 2.99 \times 10^{-8}$
- $P(J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000513$
- $P(J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 0.000253292$
- $P(J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 4.95 \times 10^{-9}$
- $P(J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 2.96406 \times 10^{-6}$

Calculation using Bayesian Networks

- $P(\neg J \wedge B \wedge E \wedge A \wedge M) = 0.000000133$
- $P(\neg J \wedge B \wedge \neg E \wedge A \wedge M) = 6.56684 \times 10^{-5}$
- $P(\neg J \wedge B \wedge E \wedge \neg A \wedge M) = 9.5 \times 10^{-10}$
- $P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge M) = 5.6886 \times 10^{-7}$
- $P(\neg J \wedge B \wedge E \wedge A \wedge \neg M) = 0.000000057$
- $P(\neg J \wedge B \wedge \neg E \wedge A \wedge \neg M) = 2.81436 \times 10^{-5}$
- $P(\neg J \wedge B \wedge E \wedge \neg A \wedge \neg M) = 9.405 \times 10^{-8}$
- $P(\neg J \wedge B \wedge \neg E \wedge \neg A \wedge \neg M) = 5.63171 \times 10^{-5}$

Calculation using Bayesian Networks

- Therefore $P(J|B) = \frac{P(J \wedge B)}{P(B)} = \frac{\Sigma P(J \wedge B \wedge e \wedge a \wedge m)}{\Sigma P(j \wedge B \wedge e \wedge a \wedge m)} = \frac{0.00849017}{0.001}$
- $P(J|B) = 0.849017$
- Can often simplify calculation without using full joint probabilities – but not always

Inference in Bayesian Networks

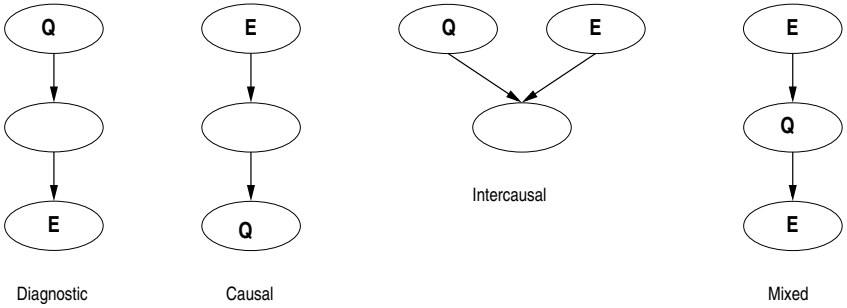
Diagnostic Inference: From effects to causes
$P(Burglary|JohnCalls) = 0.016$

Causal Inference: From causes to effects
$P(JohnCalls|Burglary) = 0.85$; $P(MaryCalls|Burglary) = 0.67$

Intercausal Inference: Explaining away
$P(Burglary|Alarm) = 0.3736$ but adding evidence, $P(Burglary|Alarm \wedge Earthquake) = 0.003$; despite the fact that burglaries and earthquakes are independent, the presence of one makes the other **much** less likely

Mixed Inference: Combinations of the patterns above
Diagnostic + Causal: $P(Alarm|JohnCalls \wedge \neg Earthquake)$
Intercausal + Diagnostic: $P(Burglary|JohnCalls \wedge \neg Earthquake)$

Inference in Bayesian Networks



- Q = query; E = evidence

Example – Causal Inference

- $P(JohnCalls|Burglary)$
- $P(J|B) = P(J|A \wedge B).P(A|B) + P(J|\neg A \wedge B).P(\neg A|B)$
  $= P(J|A).P(A|B) + P(J|\neg A).P(\neg A|B)$
  $= P(J|A).P(A|B) + P(J|\neg A).(1 - P(A|B))$
- Now $P(A|B) = P(A|B \wedge E).P(E|B) + P(A|B \wedge \neg E).P(\neg E|B)$
  $= P(A|B \wedge E).P(E) + P(A|B \wedge \neg E).P(\neg E)$
  $= 0.95 \times 0.002 + 0.94 \times 0.998 = 0.94002$
- Therefore $P(J|B) = 0.90 \times 0.94002 + 0.05 \times 0.05998 = 0.849017$
- **Fact 3:** $P(X|Z) = P(X|Y \wedge Z).P(Y|Z) + P(X|\neg Y \wedge Z).P(\neg Y|Z)$, since $X \wedge Z \equiv (X \wedge Y \wedge Z) \vee (X \wedge \neg Y \wedge Z)$ (conditional version of Fact 2)

Example – Diagnostic Inference

- $P(Earthquake|Alarm)$
- $P(E|A) = \frac{P(A|E).P(E)}{P(A)}$
  $= \frac{P(A|B \wedge E).P(B).P(E) + P(A|\neg B \wedge E).P(\neg B).P(E)}{P(A)}$
  $= \frac{0.95 \times 0.001 \times 0.002 + 0.29 \times 0.999 \times 0.002}{P(A)} = \frac{5.8132 \times 10^{-4}}{P(A)}$
- Now $P(A) = P(A|B \wedge E).P(B).P(E) + P(A|\neg B \wedge E).P(\neg B).P(E) +$
  $P(A|B \wedge \neg E).P(B).P(\neg E) + P(A|\neg B \wedge \neg E).P(\neg B).P(\neg E)$
  And $P(A|B \wedge \neg E).P(B).P(\neg E) + P(A|\neg B \wedge \neg E).P(\neg B).P(\neg E)$
  $= 0.94 \times 0.001 \times 0.998 + 0.001 \times 0.999 \times 0.998 = 0.001935122$
  So $P(A) = 5.8132 \times 10^{-4} + 0.001935122 = 0.002516442$
- Therefore $P(E|A) = \frac{5.8132 \times 10^{-4}}{0.002516442} = 0.2310087$
- **Fact 4:** $P(X \wedge Y) = P(X).P(Y)$ if X, Y are independent

Conclusion

- Due to noise or uncertainty it is useful to reason with probabilities
- Calculating with joint probability distribution difficult due to the large number of values
- Use of Bayes' Rule and independence assumptions simplifies reasoning
- Bayesian networks allow compact representation of probabilities and efficient reasoning with probabilities
- Elegant recursive algorithms can be given to automate the process of inference in Bayesian networks