COMP9418: Advanced Topics in Statistical Machine Learning

Bayesian Networks 2

Instructor: Gustavo Batista

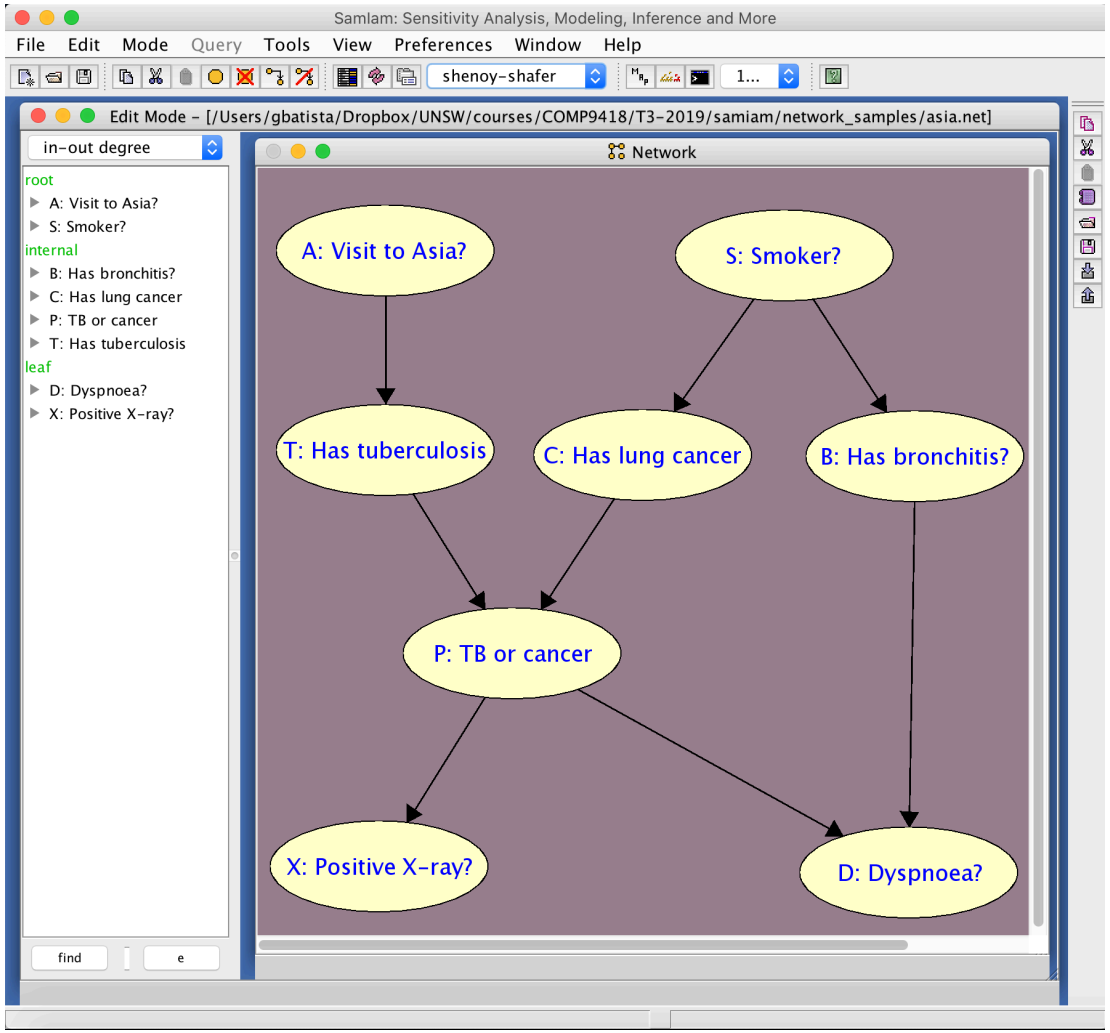
University of New South Wales

Introduction

- This lecture continues the study of Bayesian networks
  - We will see which types of queries can be answered by this probabilistic reasoning framework
  - There are four main types of queries and we will see how to use them in specific situations
- We will discuss several use cases for Bayesian networks
  - We start with a problem description and we will design a model for the problem
  - Also, we will discuss what answers we can get with different types of queries

Probability of Evidence

- Ones of the simplest queries is the probability of some variable instantiation, $\boldsymbol{e}$, $P(\boldsymbol{e})$
  - For instance, the probability a patient has a positive x-ray but no dyspnoea $P(X = true, D = false)$
  - The variables $\boldsymbol{E} = \{X, D\}$ are *evidence variables*
  - $P(\boldsymbol{e})$ is known as *probability of evidence* query
- There are other types of evidence beyond variable instantiation
  - For instance, the probability that a patient has either a positive x-ray or dyspnoea, $P(X = true \vee D = true)$
  - Bayesian networks do not directly support queries with arbitrary pieces of evidence
  - But these probabilities can be computed indirectly



Probability of Evidence

- We can add an auxiliary node E
  - Declare nodes X and D as parents of E
  - Adopt the following CPT for E
  - We can use the augmented network and compute $P(E = yes)$
- This technique is known as *auxiliary-node method*
  - It is practical only when the number of evidence variables is small
  - However, this CPT has only 0 and 1 values, known as *deterministic CPT*
  - We can use some techniques to represent deterministic CPT that do not grow exponentially

X	D	E	$P(E\|X,D)$
x	d	e	1
x	$\bar{d}$	e	1
$\bar{x}$	d	e	1
$\bar{x}$	$\bar{d}$	e	0

Prior and Posterior Marginals

- *Posterior-marginal queries* are the most common ones
  - Let us first discuss the terms posterior and marginal
- Given a joint probability distribution $P(x_1, \ldots, x_n)$
  - The marginal distribution $P(x_1, \ldots, x_m), m \leq n$ is defined as
  - The marginal distribution can be viewed as a projection of the joint distribution on a smaller set of variables

$$P(x_1, \ldots, x_m) = \sum_{x_{m+1}, \ldots, x_n} P(x_1, \ldots, x_n)$$

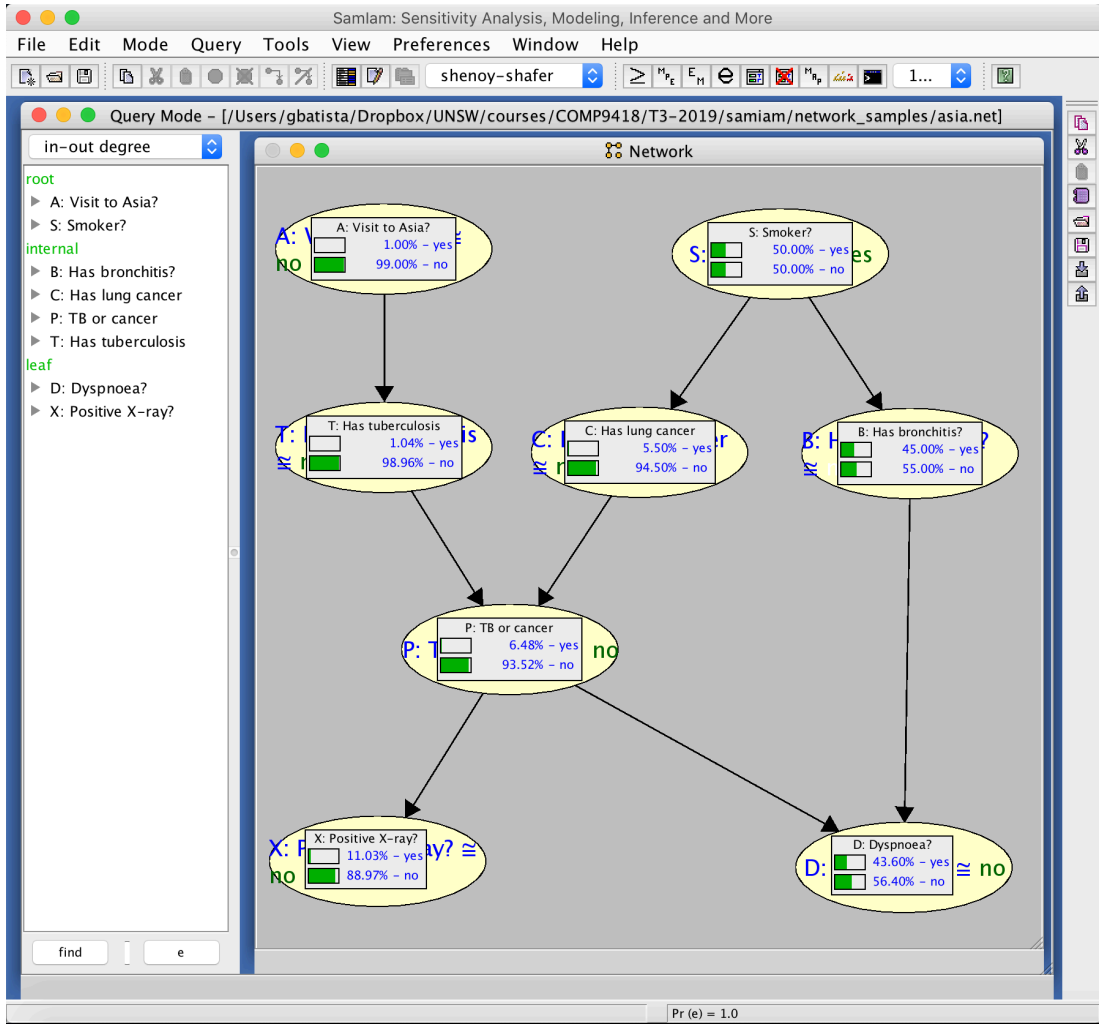
- When the marginal distribution is computed given some evidence $\boldsymbol{e}$
  - It is known as a *posterior marginal*
  - This contrasts with marginal distribution given no evidence, known as *prior marginal*

$$P(x_1, \ldots, x_m|\boldsymbol{e}) = \sum_{x_{m+1}, \ldots, x_n} P(x_1, \ldots, x_n|\boldsymbol{e})$$

Prior and Posterior Marginals

- This screen shows the marginals for each variable
  - For instance, the distribution of variable C, lung cancer is

C	$P(C)$
c	5.5%
$\bar{c}$	94.5%



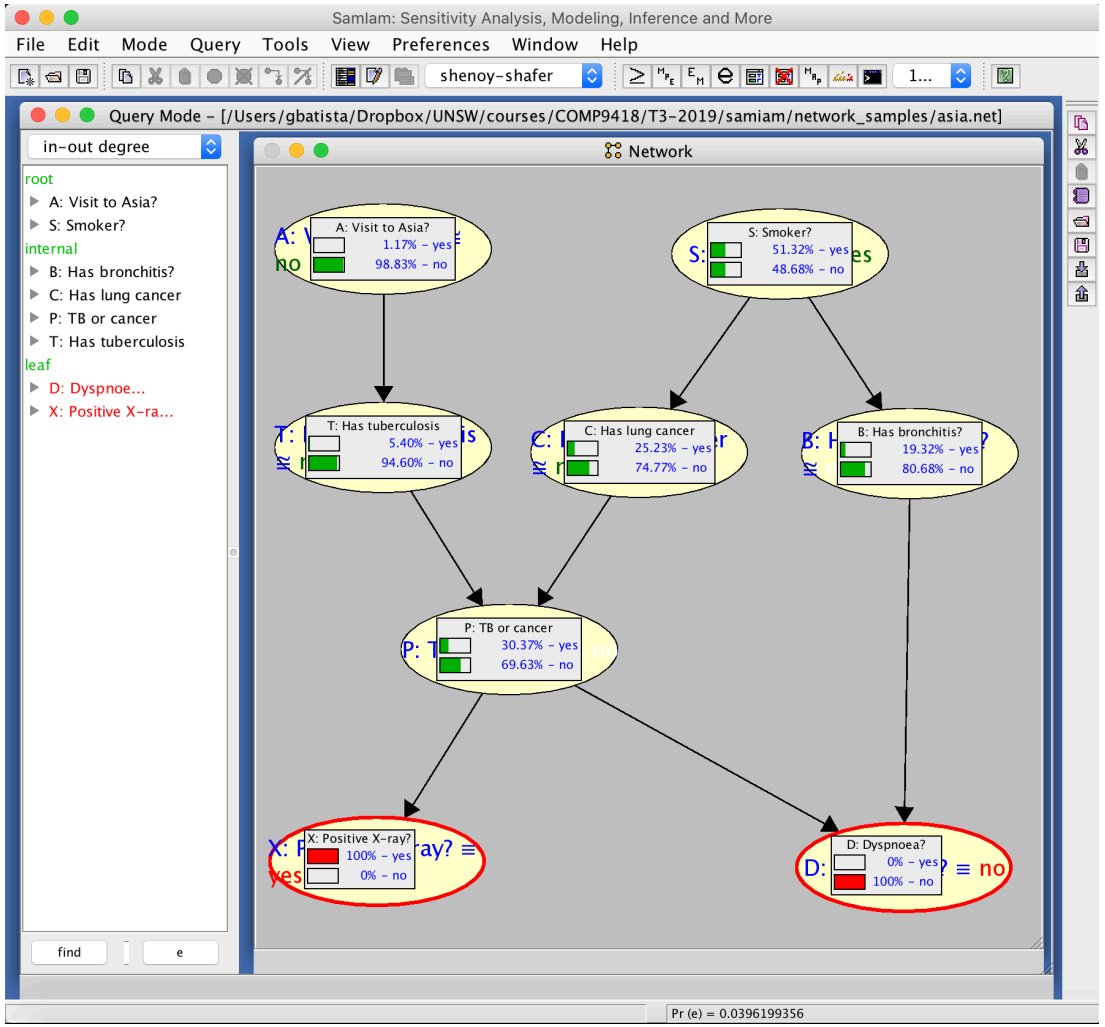
Prior and Posterior Marginals

- This screen shows the marginals for each variable
  - For instance, the distribution of variable C, lung cancer is

C	$P(C)$
c	5.5%
$\bar{c}$	94.5%

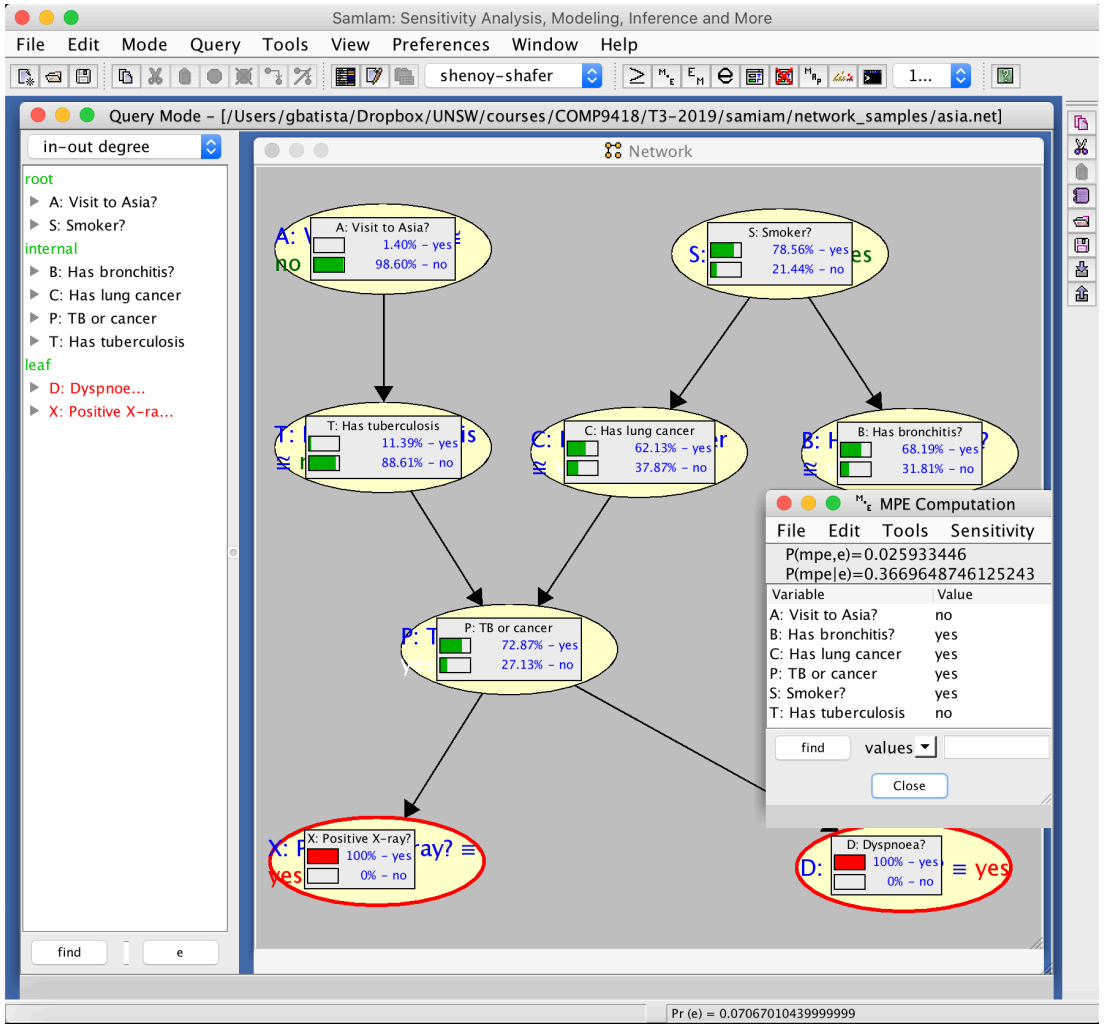
- Suppose the patient has a positive x-ray but no dyspnoea B: $X = true, D = false$.
  - The posterior marginal for variable C is

C	$P(C\|\boldsymbol{e})$
c	25.23%
$\bar{c}$	74.77%



Most Probable Explanation

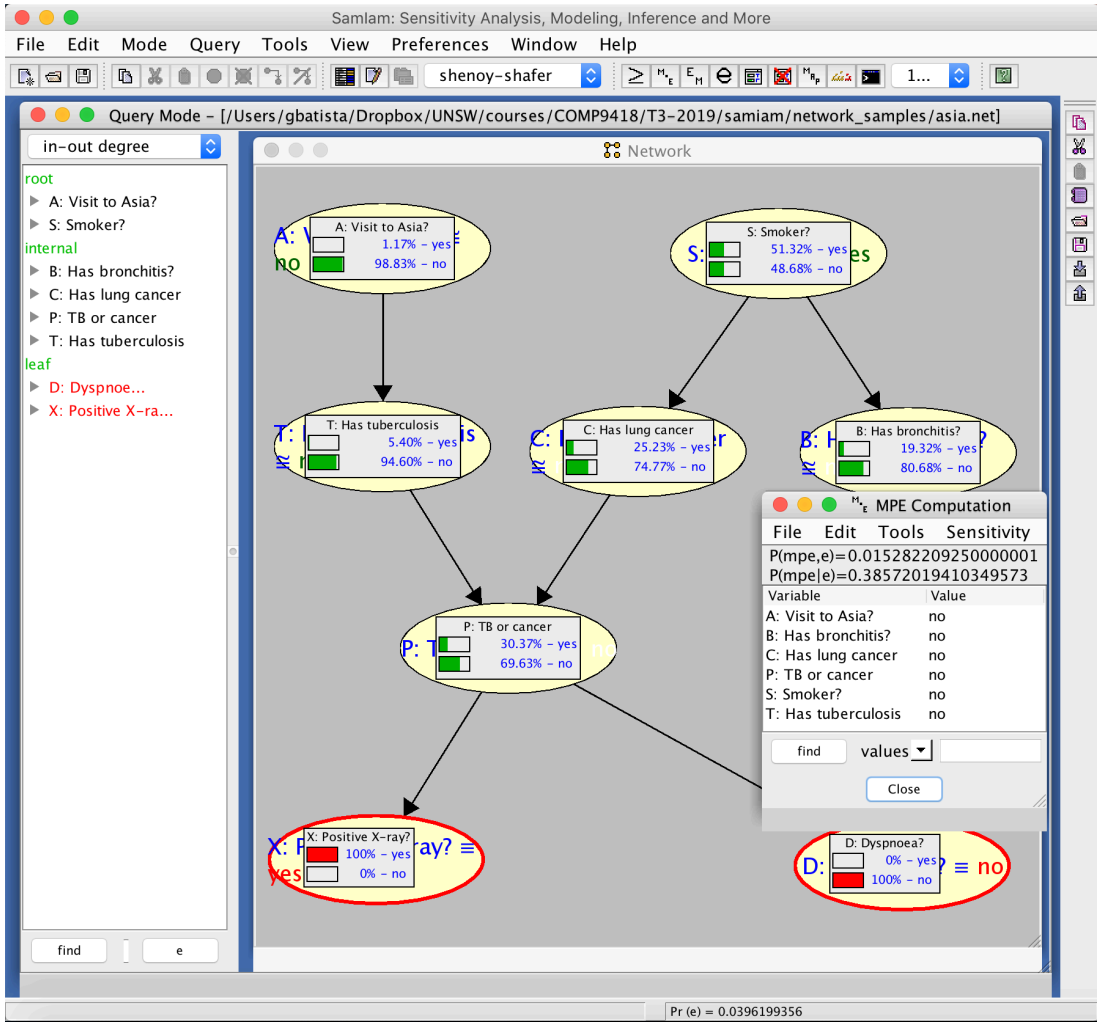
- We now consider the *most probable explanation* (MPE) queries
  - The goal is to identify the most probable instantiation given some evidence
  - Given $X_1, \ldots, X_n$ variables and $\boldsymbol{e}$ is the evidence, the goal is to identify the instantiation $x_1, \ldots, x_n$ for which the probability $P(x_1, \ldots, x_n | \boldsymbol{e})$ is maximal
  - Such instantiation is called the *most probable explanation* given evidence $\boldsymbol{e}$
- For example, the MPE for a patient with positive x-ray and dyspnoea
  - It is a person that made no visit to Asia, is a smoker, and has lung cancer and bronchitis but no tuberculosis



Most Probable Explanation

- MPE cannot be obtained directly from the posterior marginals
  - That is, choosing each value x_i to maximize $P(x_i|\boldsymbol{e})$

- Consider the case in which the patient has a positive x-ray but no dyspnoea
  - We get an explanation the patient is not a smoker with probability of approximately 38.57%
  - However, if we choose for each variable the value with maximal probability, we get an explanation in which the patient is a smoker with probability of approximately 20.03%

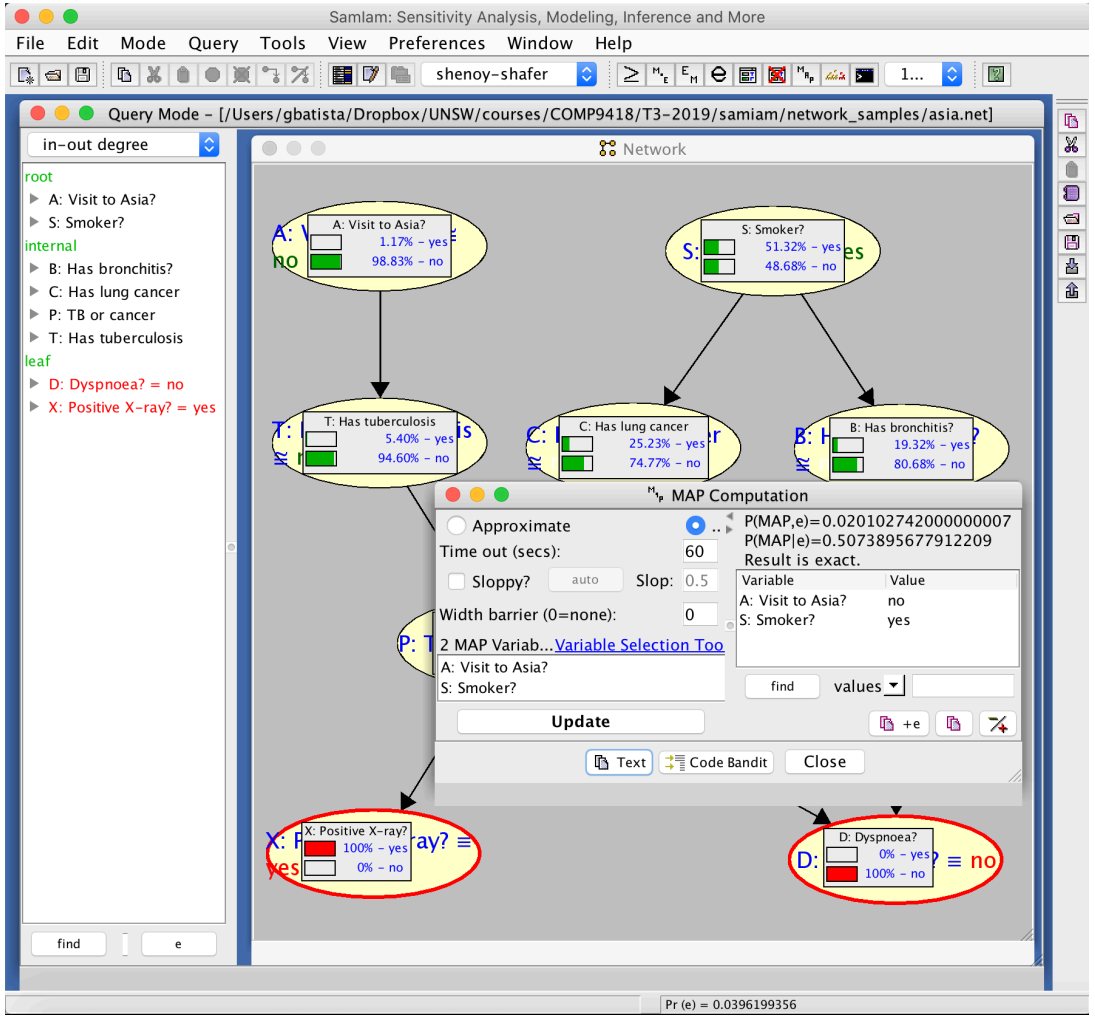


Maximum a Posteriori Hypothesis

- MPE is a special case of a more general class of queries
  - We may want to find the most probable instantiation for a subset of variables
  - Given $\boldsymbol{M} \subseteq \boldsymbol{X}$ and some evidence $\boldsymbol{e}$, our goal is to find the instantiation $\boldsymbol{m}$ of variables $\boldsymbol{M}$ for which $P(\boldsymbol{m}|\boldsymbol{e})$ is maximal
- Such instantiation $\boldsymbol{m}$ is known as *maximum a posteriori hypothesis* (MAP)
  - The variables $\boldsymbol{M}$ are known as MAP variables
  - MPE is a special case of MAP when MAP variables include all network variables
  - Such distinction exists because MPE is much easier to compute

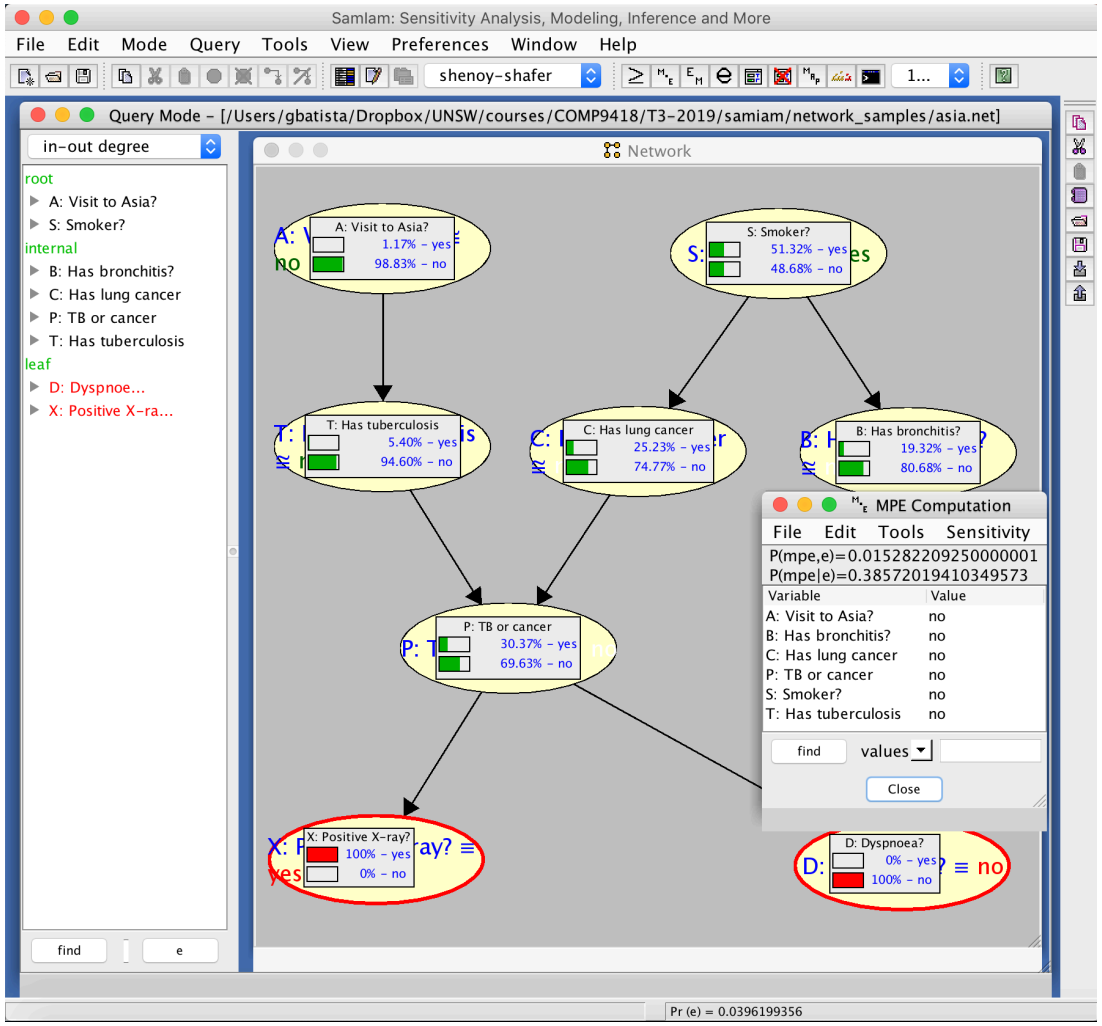
Maximum a Posteriori Hypothesis

- In this example we consider that
  - The patient has a positive x-ray and no dyspnoea
  - MAP variables are $\boldsymbol{M} = \{A, S\}$
  - Which answer is $A = no, S = yes$ with probability of approximately 50.74%
- MPE is frequently used to approximate MAP
  - We say we are projecting the MPE on MAP variables



Maximum a Posteriori Hypothesis

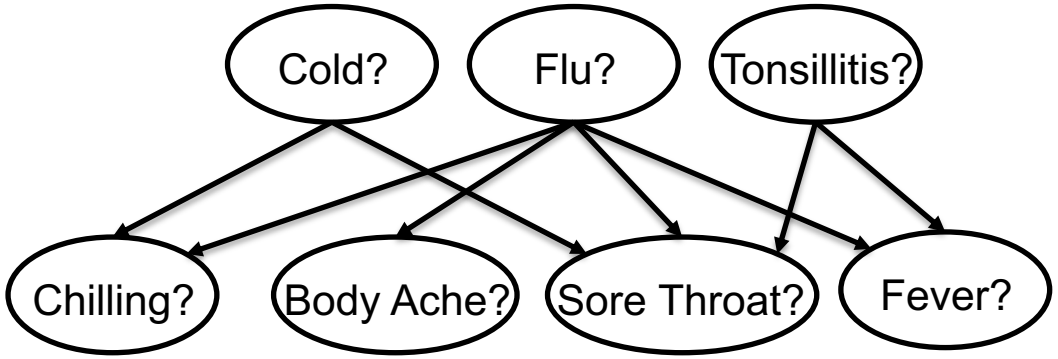
- In this example we consider that
  - The patient has a positive x-ray and no dyspnoea
  - MAP variables are $\boldsymbol{M} = \{A, S\}$
  - Which answer is $A = no, S = yes$ with probability of approximately 50.74%
- MPE is frequently used to approximate MAP
  - We say we are projecting the MPE on MAP variables
  - However, it is just an approximation.
  - This figure shows the MPE answer for $\boldsymbol{M}$ variables is $A = no, S = no$ with probability approximately 48.09%



Diagnosis Model from Expert I

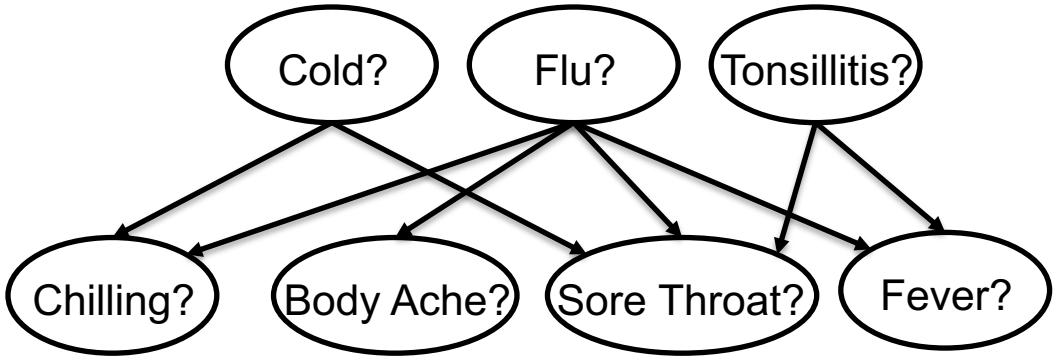
- Consider the following medical information
- Which variables
  - Disease: flu, cold and tonsillitis
  - Symptoms: chilling, body ache and pain, sore throat, and fever
- Values
  - True or false
- Structure
  - From the problem statement

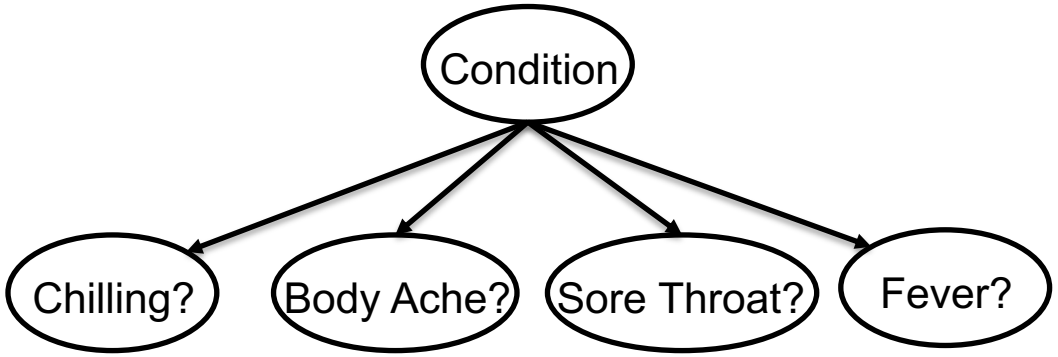
The flu is an acute disease characterised by fever, body aches and pains, and can be associated with chilling and a sore throat. The cold is a bodily disorder popularly associated with chilling and can cause a sore throat. Tonsillitis is inflammation of the tonsils that leads to a sore throat and can be associated with fever



Diagnosis Model from Expert I

- Another modelling has one variable "Condition" with values normal, cold, flu, tonsillitis
  - This network structure is known as *naïve Bayes*
  - The naïve Bayes has the structure $C \rightarrow \mathrm{A}_1, \ldots, C \rightarrow A_m$, where C is the *class* and $A_1, \ldots, \mathrm{A}_m$ are the *attributes*
- The naïve Bayes structure
  - Has a *single-fault* assumption
  - Attributes are independent given a condition
  - All attributes are connected to the condition node





Diagnosis Model from Expert I

- CPTs for the conditions
  - Must provide the belief in developing the condition by a person we have no knowledge of any symptom
- CPTs for the symptoms
  - Must provide the belief in this symptom under all possible combinations of possible conditions
- The probabilities for the CPTs can come
  - From medical statistics or subjective beliefs
  - Estimating from medical records of previous patients
- The diagnosis problem
  - Symptoms represent known evidence
  - Compute the most probable combination of conditions given the evidence
  - MAP or MPE query

Case?	Cold?	Flu?	Tonsillitis?	Chilling?	Bodyache?	Sorethroat?	Fever?
1	true	false	?	true	false	false	false
2	false	true	false	true	true	false	true
3	?	?	true	false	?	true	false
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

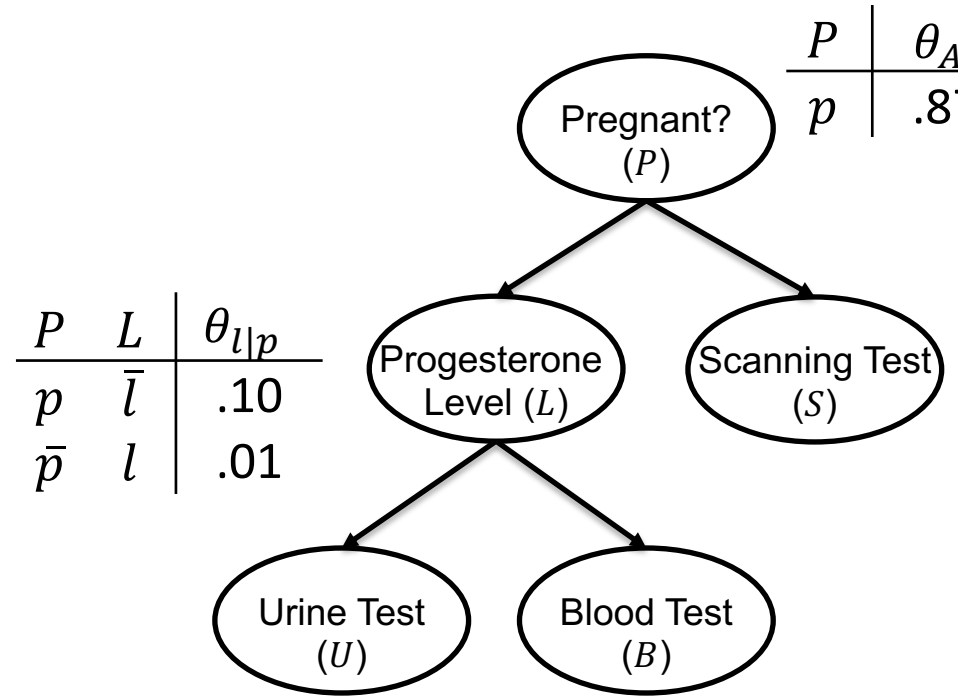
Diagnosis Model from Expert II

- Consider the problem of computing the probability of pregnancy given some tests
- Variables
  - Query variable to represent pregnancy (P)
  - Three evidence variables to represent test results: scanning (S), blood (B) and urine (U)
  - One intermediary variable to represent progesterone level (L)
- Values
  - Binary, depending on the variable
- Structure
  - Causal from the problem statement

A few weeks after inseminating a cow, we have three possible tests to confirm pregnancy. The first is a scanning test that has a false positive of 1% and a false negative of 10%. The second is a blood test that detects progesterone with a false positive of 10% and a false negative of 30%. The third test is a urine test that also detects progesterone with a false positive of 10% and a false negative of 20%. The probability of a detectable progesterone level is 90% given pregnancy and 1% given no pregnancy. The probability that insemination will impregnate a cow is 87%

Diagnosis Model from Expert II

- Some independencies
  - Blood and urine tests are independent, given the progesterone level
  - The scanning test is independent of the blood and urine tests, given the status of pregnancy
  - Urine and blood tests are not independent given the status of pregnancy
- Query
  - Suppose we inseminate a cow and after a few weeks the three tests come out negative
    $\boldsymbol{e}: S = false, B = false, U = false$
  - $P(P|\boldsymbol{e}) \approx 10.21\%$
  - It is relatively high given all three tests came out negative



P	θ_A
p	.87

P	L	$\theta_{l\|p}$
p	$\bar{l}$	.10
$\bar{p}$	l	.01

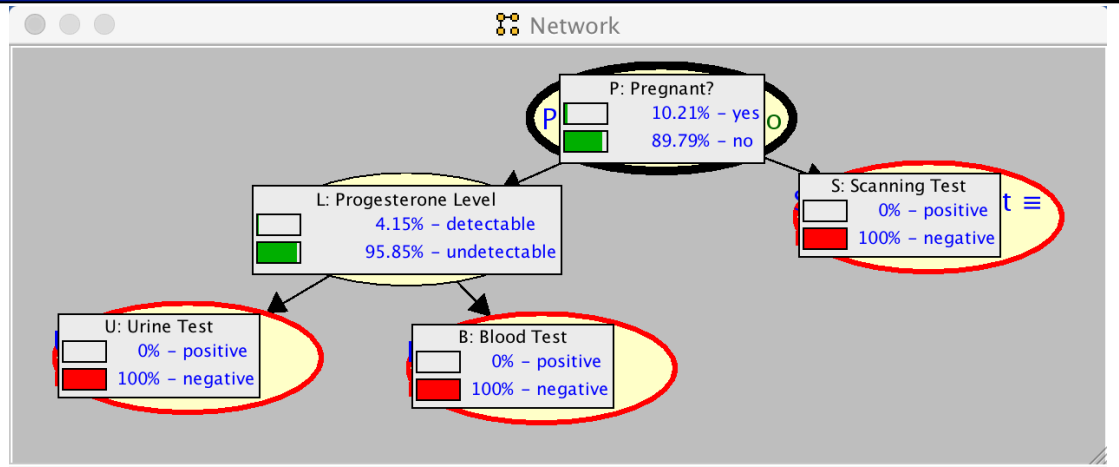
L	U	$\theta_{u\|l}$
l	$\bar{u}$	.20
$\bar{l}$	u	.10

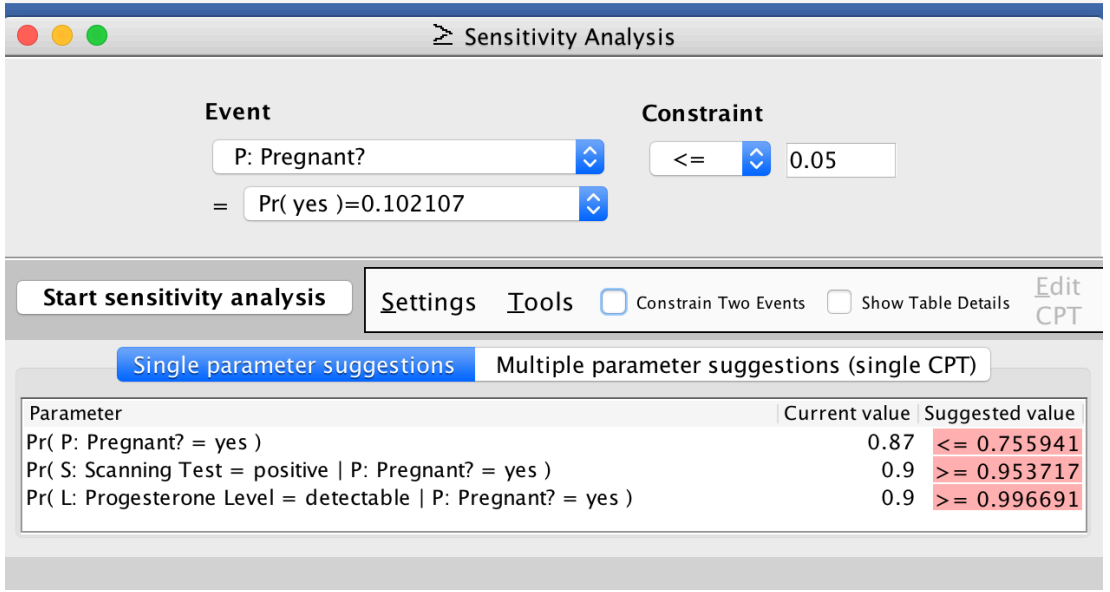
L	B	$\theta_{b\|l}$
l	$\bar{b}$	.30
$\bar{l}$	b	.10

P	S	$\theta_{s\|p}$
p	$\bar{s}$	.10
$\bar{p}$	s	.01

Sensitivity Analysis

- Suppose the farmer is not happy
  - The three negative tests need to drop the probability of pregnancy to less than 5%
  - We will need to replace the tests, but we need to know the new false positive and negative rates
- Sensitivity analysis
  - Which network parameters do we have to change, and by how much, to ensure that the probability of pregnancy would be no more than 5% given three negative tests?
- Solution is to change the scanning test by one with 4.63% false negative rate
  - Urine and blood test cannot help
  - The uncertainty of the progesterone level is such that even perfect urine and blood tests cannot achieve the desired confidence level

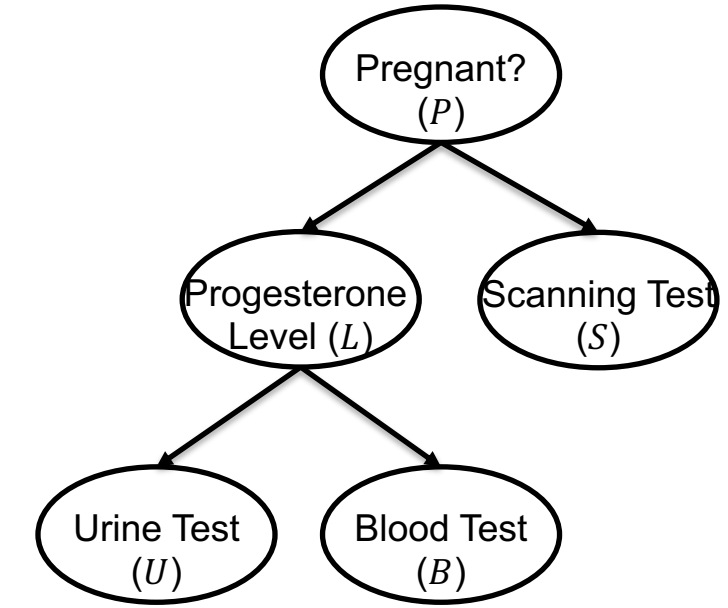


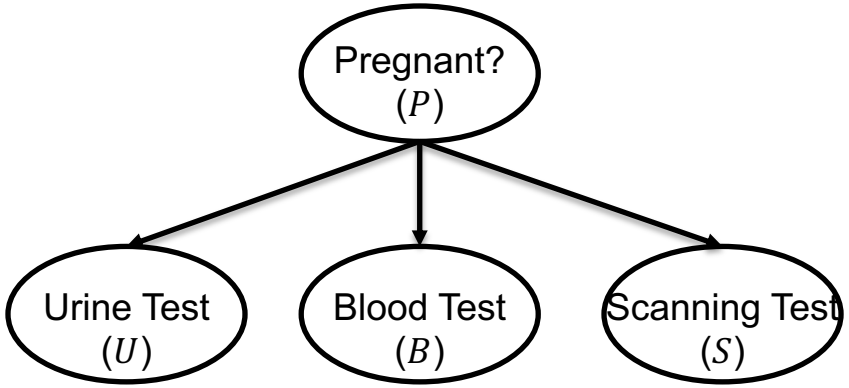


Network Granularity

- The progesterone level is neither a query or an evidence variable
  - Why do we need to include it in the network?
- Intermediate variables are a modelling convenience
  - It helps modelling urine and blood tests with pregnancy
  - However, the network allow us to compute the following

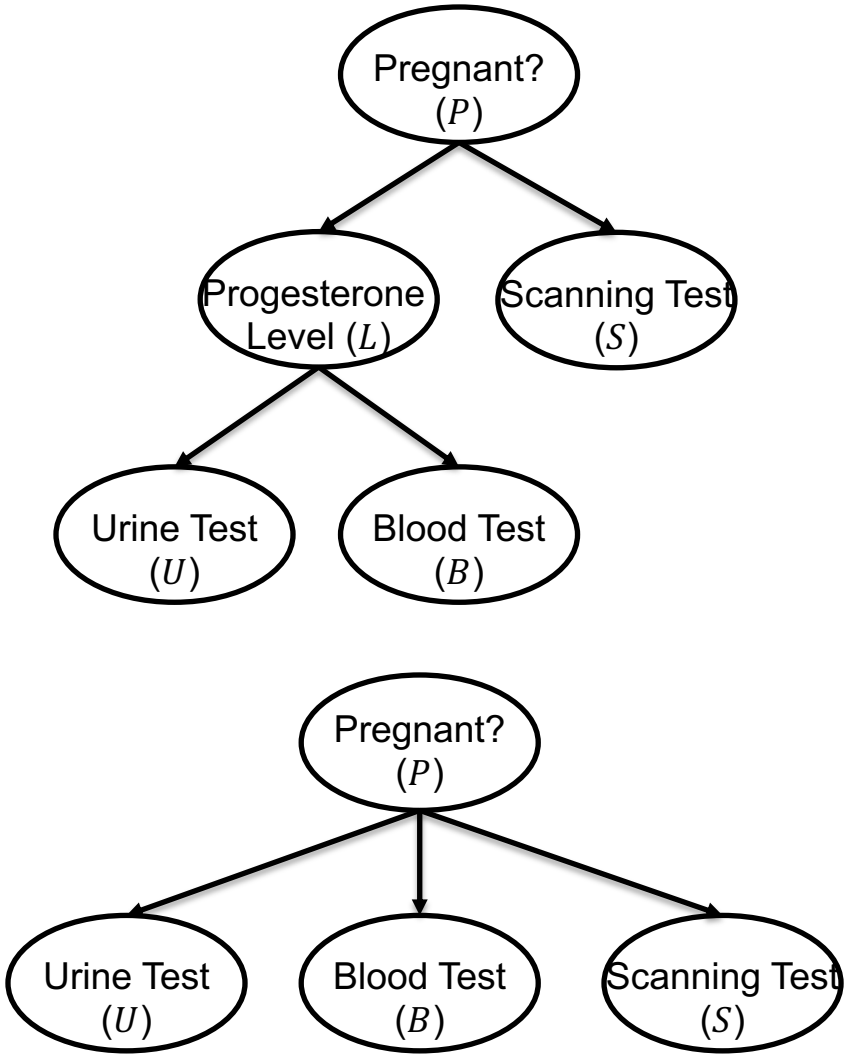
$$P(B = false|P = true) = 36\%$$
$$P(B = true|P = false) = 10.6\%$$
$$P(U = false|P = true) = 27\%$$
$$P(U = true|P = true) = 10.7\%$$





Network Granularity

- Is this simpler network equivalent to the original one?
  - Simpler: negative blood and urine tests will count more in ruling out pregnancy (45.09% vs 52.96%)
  - We cannot remove intermediate variables without undesirable effects in certain cases
- In general, an intermediate variable can be bypassed without affecting *model accuracy* if
  - $P(\boldsymbol{q}, \boldsymbol{e}) = P'(\boldsymbol{q}, \boldsymbol{e})$
  - For all instantiations $\boldsymbol{q}$ of the query variables $\boldsymbol{Q}$ and $\boldsymbol{e}$ of the evidence variables $\boldsymbol{E}$
  - P is induced by the original Bayesian network and P' by the new network

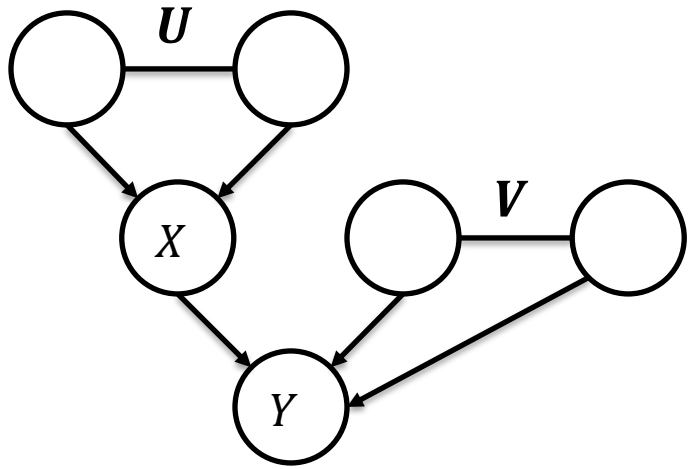


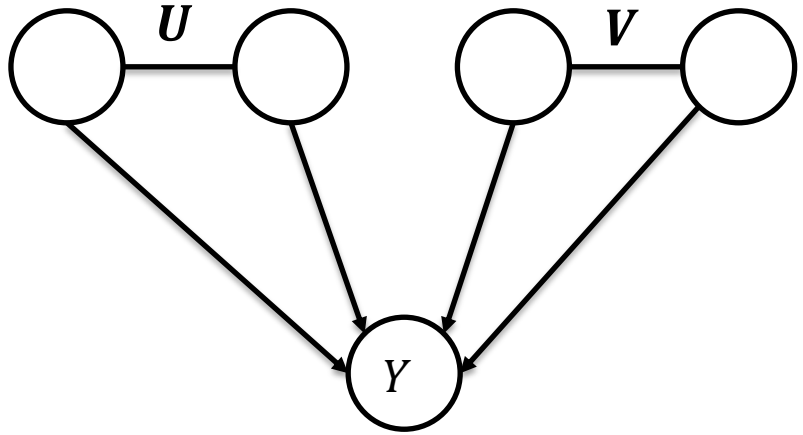
Network Granularity

- Suppose that X is not a query or an evidence variable
  - X can be bypassed if it has a single child Y
  - In this case, the CPT for variable Y is

$$\theta'_{y|\boldsymbol{uv}} = \sum_{x} \theta_{y|x\boldsymbol{v}}\theta_{x|\boldsymbol{u}}$$

  - $\boldsymbol{U}$ are the parents of X and $\boldsymbol{V}$ the parents of Y other than X
- In most case, we do not bypass intermediate variables
  - It tends to create larger CPTs even it does not affect model accuracy

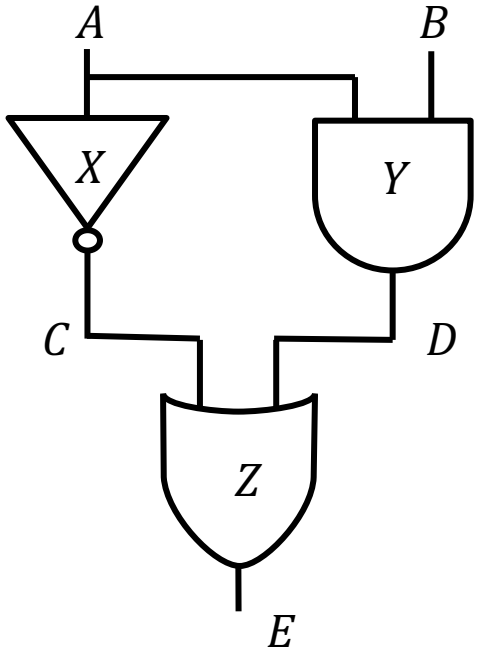




Diagnosis Model from Design

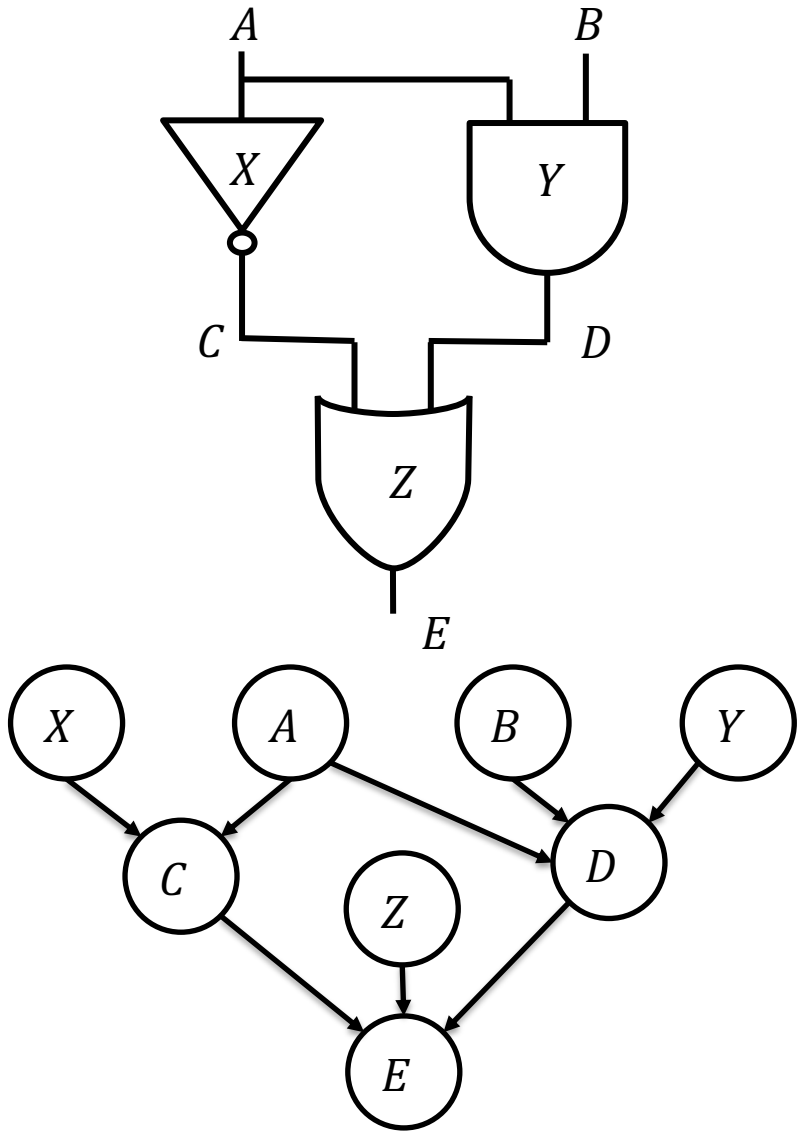
- This is another diagnosis problem
  - The model will be general to the point it can be generated automatically for similar instances
- Evidence variables
  - Primary inputs and outputs of the circuit: A, B and E
- Query variables
  - One for each component: X, Y, and Z
- Intermediate variables
  - Internal wires: C and D
  - For larger circuits, it would be unfeasible to model the circuit without representing the internal states

Consider this digital circuit. Given some values for the circuit primarily inputs and outputs, our goal is to decide whether the circuit is behaving normally. If not, decide the most likely health states of its components



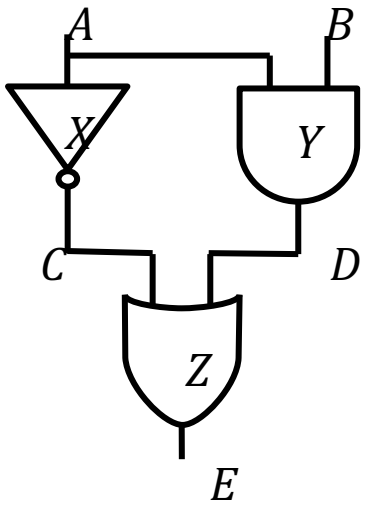
Diagnosis Model from Design

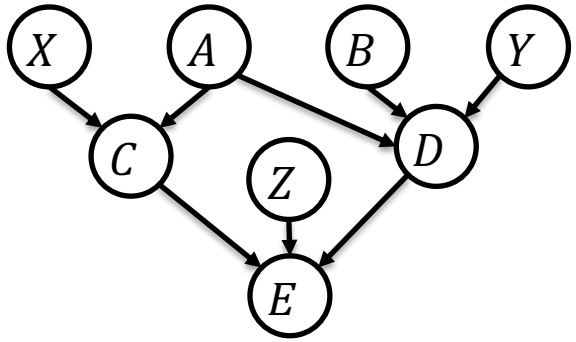
- This structure is general
  - It applied to any system composed by function blocks
  - Outputs determined by inputs and health state
  - Do not contain feedback loops
- Values
  - High or low for wire variables
  - Ok or faulty for health variables
  - Stuck-at-zero or stuck-at-one may be more specific



Diagnosis Model from Design

- CPTs
  - Health variables defined the probability of being faulty
  - Component variables are deterministic
  - CPTs for primary inputs





X	θ_x
ok	.99
faulty	.01

X	θ_x
ok	.99
stuckat0	.005
stuckat1	.005

A	X	C	$\theta_{c\|a,x}$
high	ok	high	0
low	ok	high	1
high	faulty	high	.5
low	faulty	high	.5

A	X	C	$\theta_{c\|a,x}$
high	ok	high	0
low	ok	high	1
high	stuckat0	high	0
low	stuckat0	high	0
high	stuckat1	high	1
low	stuckat1	high	1

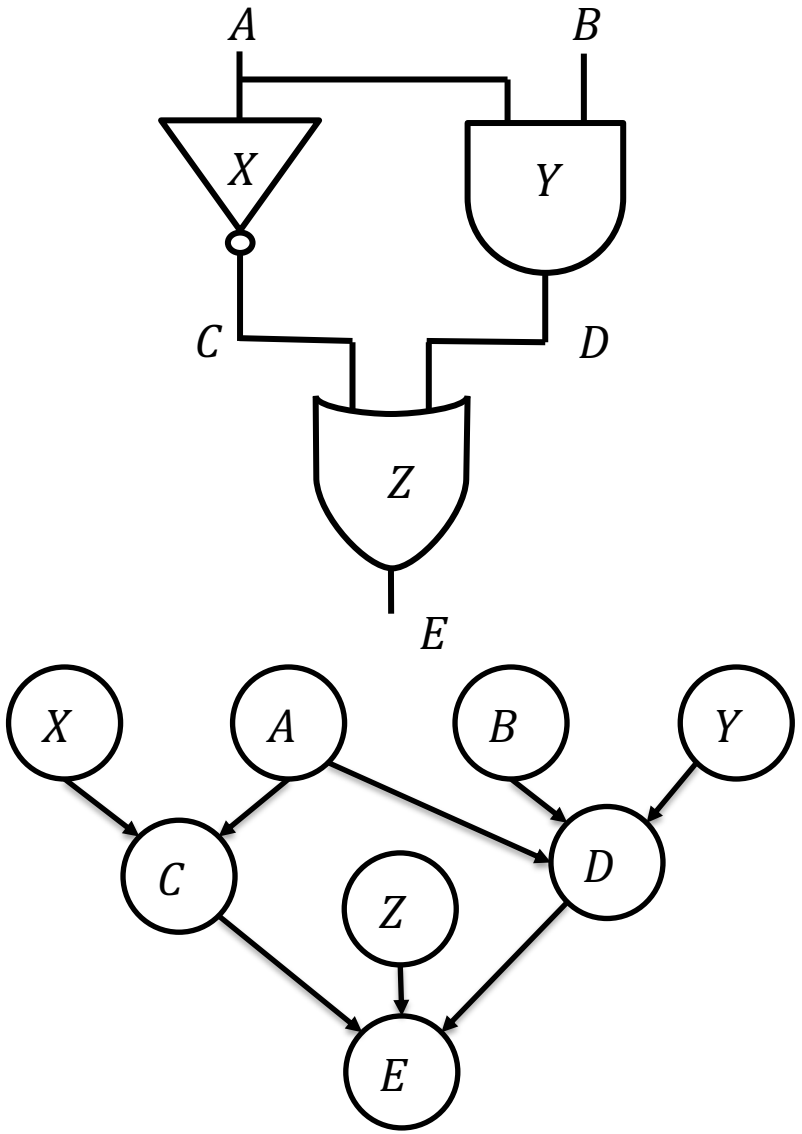
A	θ_a
high	.5
low	.5

Diagnosis Model from Design

- Suppose we have the following test vector
  - $\boldsymbol{e}$: $A = high, B = high, E = low$
  - We want to compute MAP query over health variables

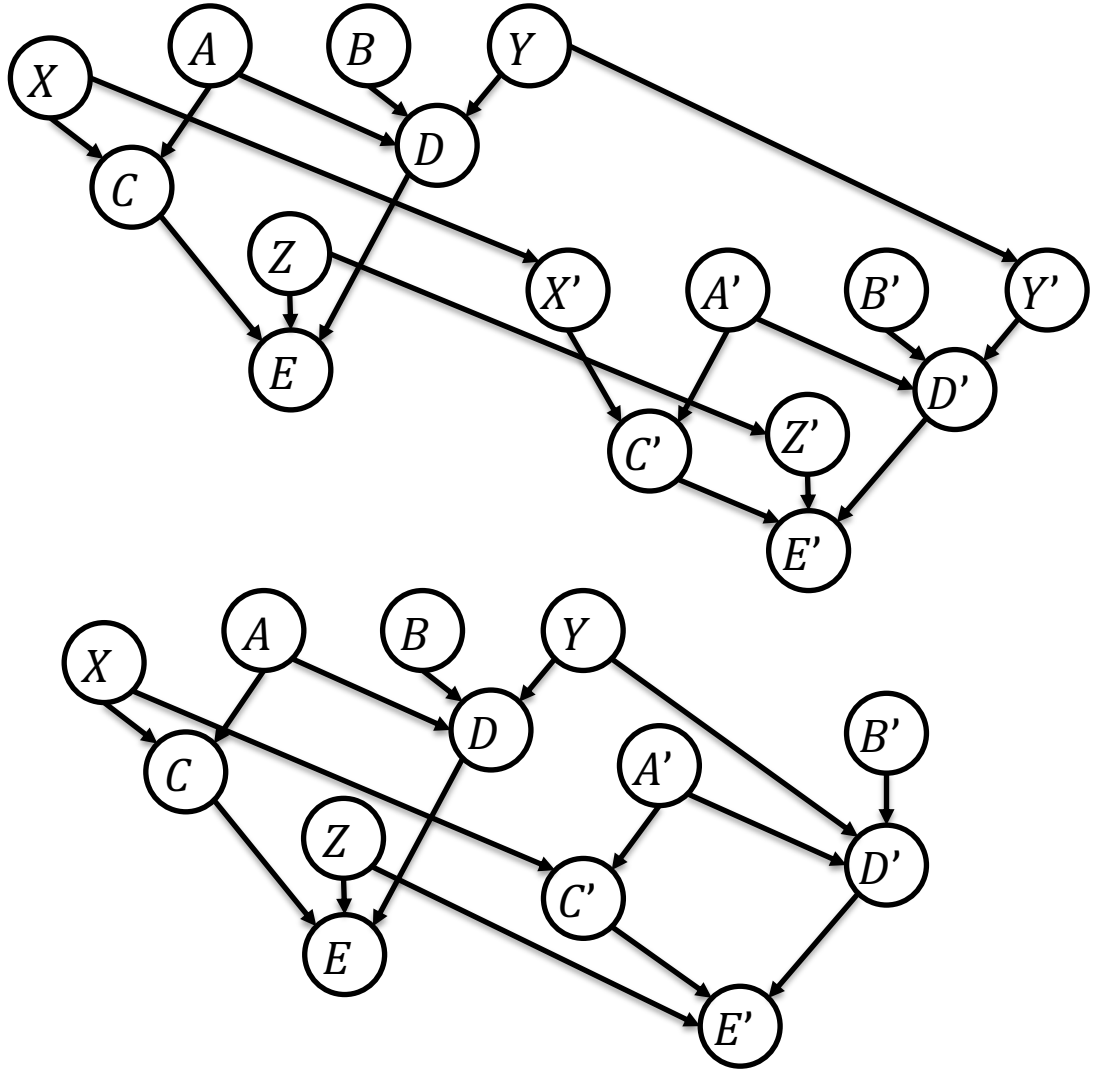
X	Y	Z	MAP
ok	stuckat0	ok	49.4%
ok	ok	stuckat0	49.4%

X	Y	Z	MAP
ok	faulty	ok	49.4%
ok	ok	faulty	49.4%



Diagnosis Model from Design

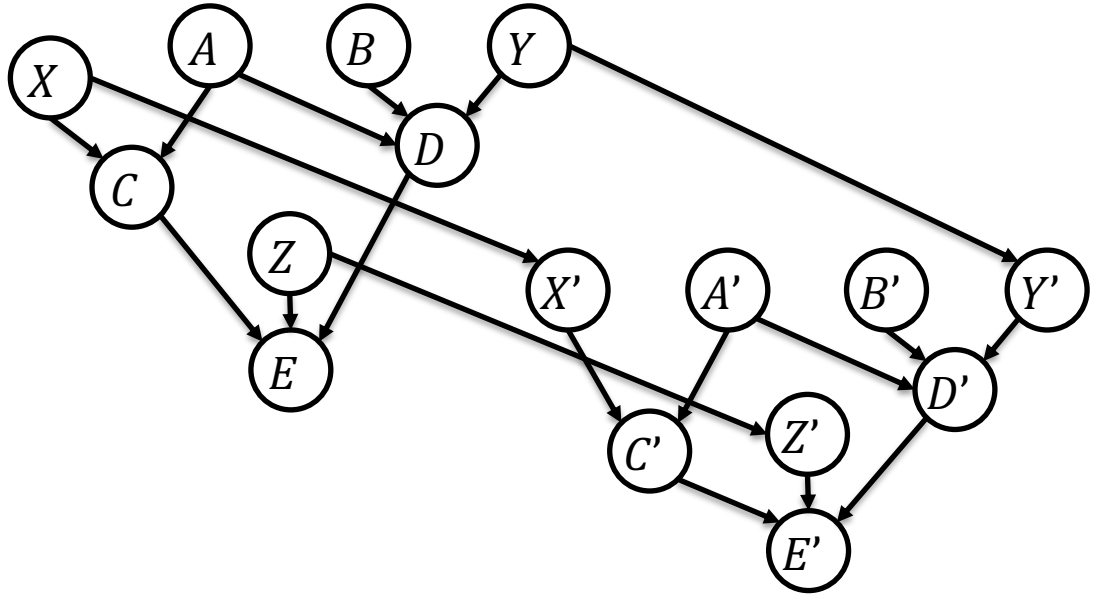
- Suppose we have two test vectors instead of one
  - For instance, we want to solve the ambiguity of the MAE result
  - We apply two low inputs and get an abnormal low as output
- We have evidence variables
  - A', B' and E' for new input vector
  - C' and D' for internal wires
  - X', Y' and Z' are necessary if we want to model intermittent faults



Diagnosis Model from Design

- Suppose we have the following test vectors
  - $\boldsymbol{e}$: $A = high, B = high, E = low$
  - $\boldsymbol{e}$': $A = low, B = low, E = low$

X	Y	Z	MAP
ok	ok	faulty	97.53%



- For intermittent faults, we need as additional CPT (*persistency model*)

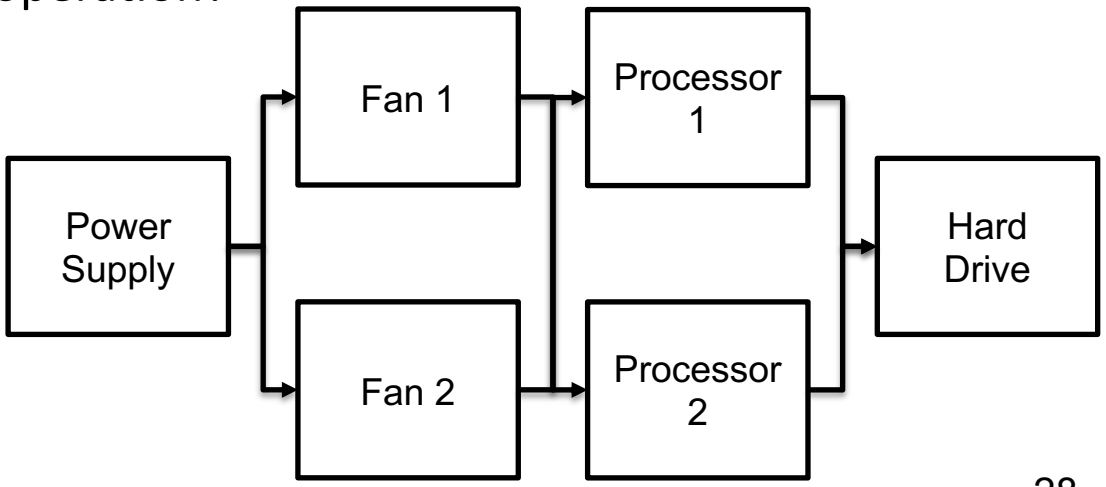
X	X'	$\theta_{x'\|x}$
ok	ok	.99
ok	faulty	.01
faulty	ok	.001
faulty	faulty	.999

- Dynamic Bayesian network (DBN)
  - Include multiple copies of the same variable
  - Different copies represent different states of the variable over time

Reliability Model from Design

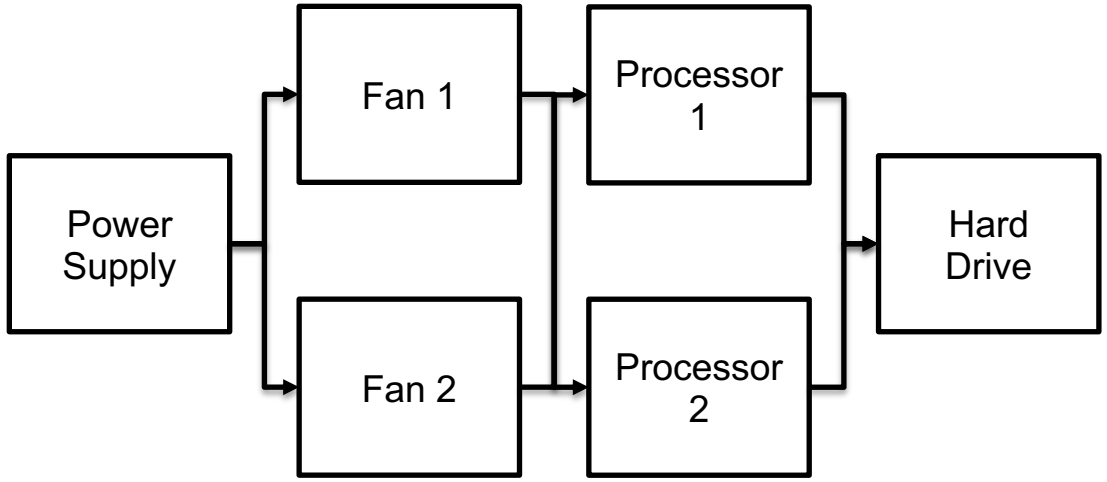
- To address this problem we need an RBD interpretation
  - Each node represents a block
  - Block B represents a subsystem that includes the component B and the subsystems feeding into B
  - For B to be available, at least one of the subsystems feeding into B must be available
  - Our representation has a single leaf node

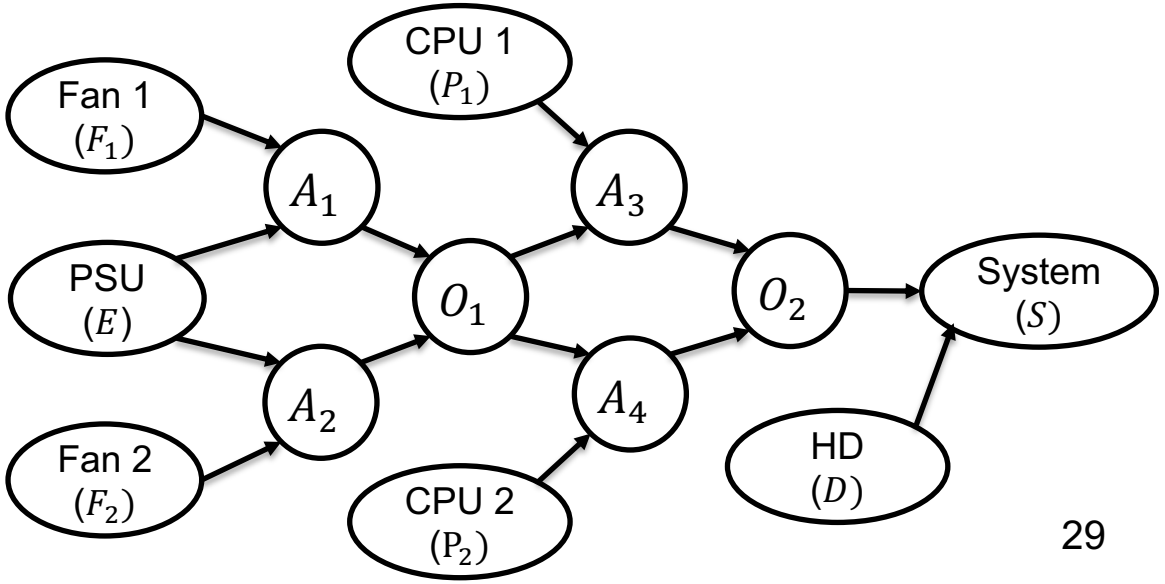
This figure depicts a *reliability block diagram* (RBD) of a computer system, indicating conditions under which the system is guaranteed to be functioning normally (available). At 1,000 days since initial operation, the reliability of different components are as follows: power supply is 99%, fan is 90%, processor is 96%, and the hard drive is 98%. What is the overall system reliability at 1,000 days since operation?



Reliability Model from Design

- Variables
  - Availability of each component: E, F_1, F_2, P_1, P_2, and D
  - Availability of the whole system: S
  - Intermediary variables for AND's and OR's
- CPTs
  - Root variables correspond to system components
  - Intermediate variables $A_1, \ldots, A_4$ and S are AND gates
  - Intermediate variables $O_1, \ldots, O_2$ are OR gates





Reliability Model from Design

- Queries
  - Marginal for variable S provides the system reliability ($\approx 95.9\%$)
  - If we want to increase the system reliability to 96.5% by replacing one component
    - Increase reliability of the hard drive to $\approx 98.6\%$
    - Increase reliability of the power supply to $\approx 99.6\%$
    - Increase reliability of either fan to $\approx 96.2\%$
  - We found the system functioning abnormally at day 1,000. MAP provides the most likely explanation

E	F_1	F_2	P_1	P_2	D	$P(.\|S = \text{un\_avail})$
avail	avail	avail	avail	avail	un_avail	36%

  - The next most likely explanations are

E	F_1	F_2	P_1	P_2	D	$P(.\|S = \text{un\_avail})$
avail	un_avail	un_avail	avail	avail	avail	21.8%
un_avail	avail	avail	avail	avail	avail	17.8%

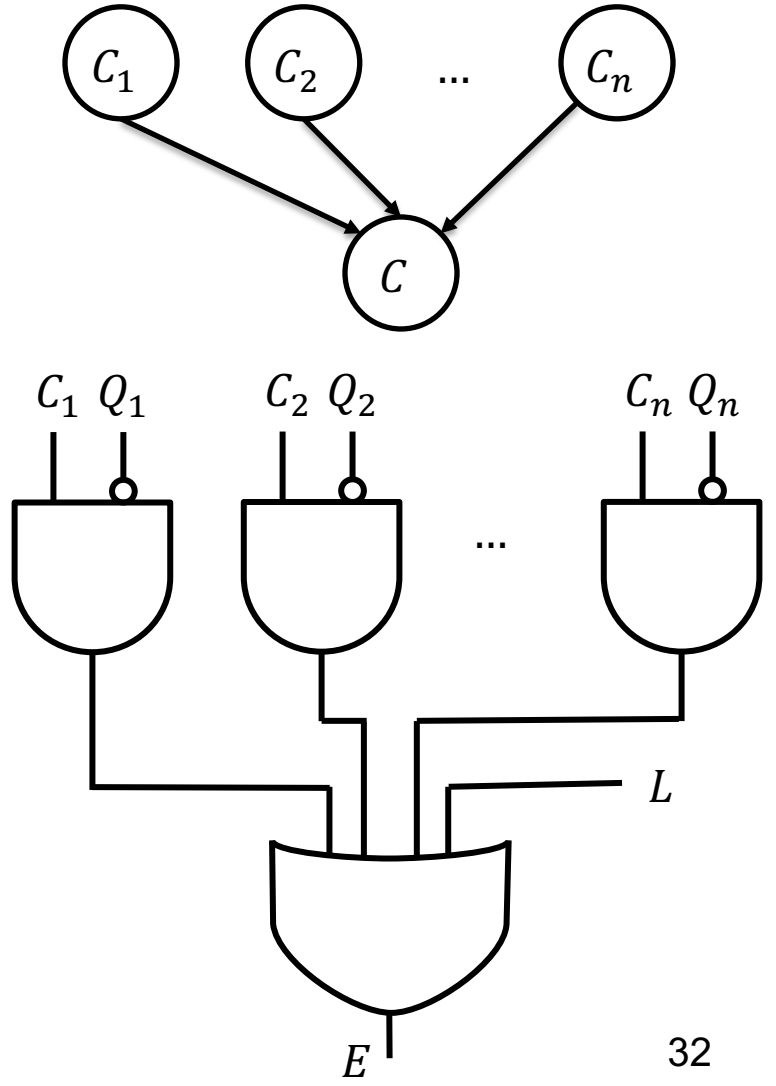
Dealing with Large CPTs

- One major issue is dealing with large CPTs
  - If a binary variable has n parents
  - We have 2^n independent parameters
- This situation causes modelling and computational problems
  - Modelling problems tend to appear first
  - 1,024 entries may be a small number for storage/inference
  - But eliciting 1,024 probabilities to quantify the relationship between, say, headache and ten medical conditions is difficult

Number of parents (n)	Number of parameters (2^n)
2	4
3	8
6	64
10	1,024
20	1,048,576
30	1,073,741,824

Micro Models: Noisy-or

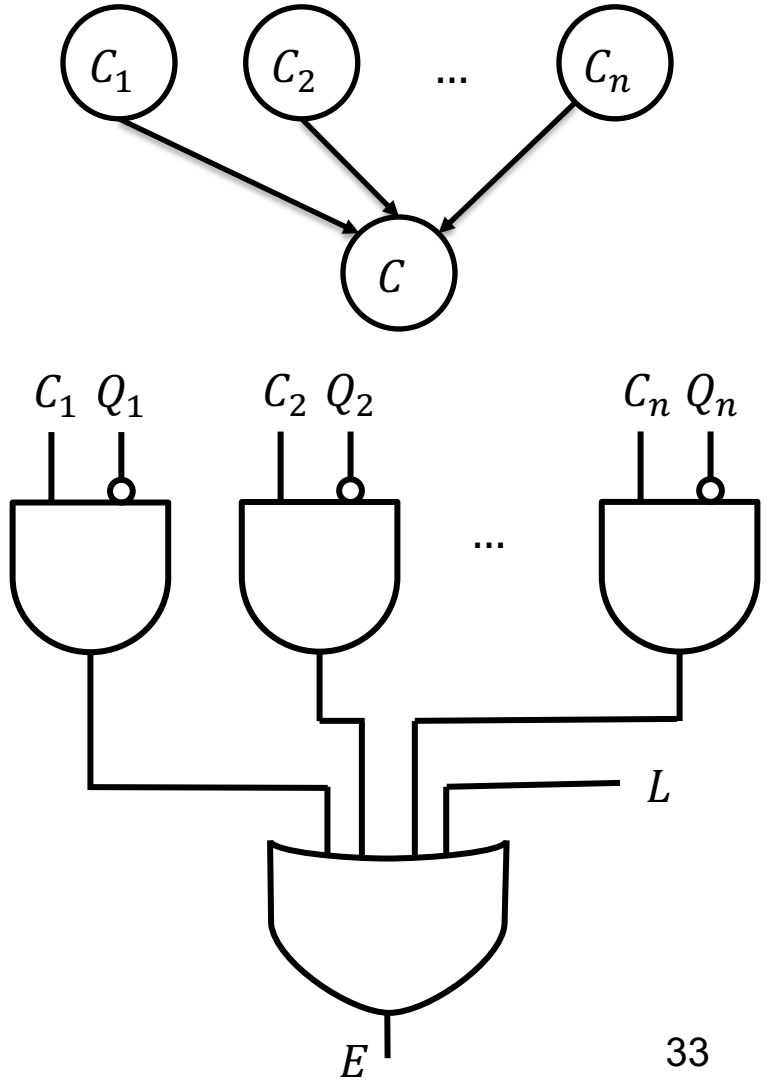
- The first approach is to develop a *micro model*
  - It specifies the relationship between the parents and their common child
  - But with a number of parameters that is smaller then 2^n
- A well-known micro model is the *noisy-or*
  - In a causal interpretation, each cause C_i is capable of establishing the effect E on its own
  - Except under some unusual circumstances summarised by the suppressor variable Q_i
  - Moreover, the leak variable L is meant to represent all other causes of E



Micro Models: Noisy-or

- The noisy-or model can be specified with $n + 1$ parameters
  - $\theta_{q_i} = P(Q_i = \text{active})$: probability that the suppressor of cause C_i is active
  - $\theta_l = P(L = \text{active})$: the probability that the leak variable is active
- Let I_α be the indices of causes that are active in α
  - For instance, if α: $C_1 = \text{active}, C_2 = \text{active}, C_3 = \text{passive}, C_4 = \text{passive}, C_5 = \text{active}$
  - Then $I_\alpha = \{1,2,5\}$

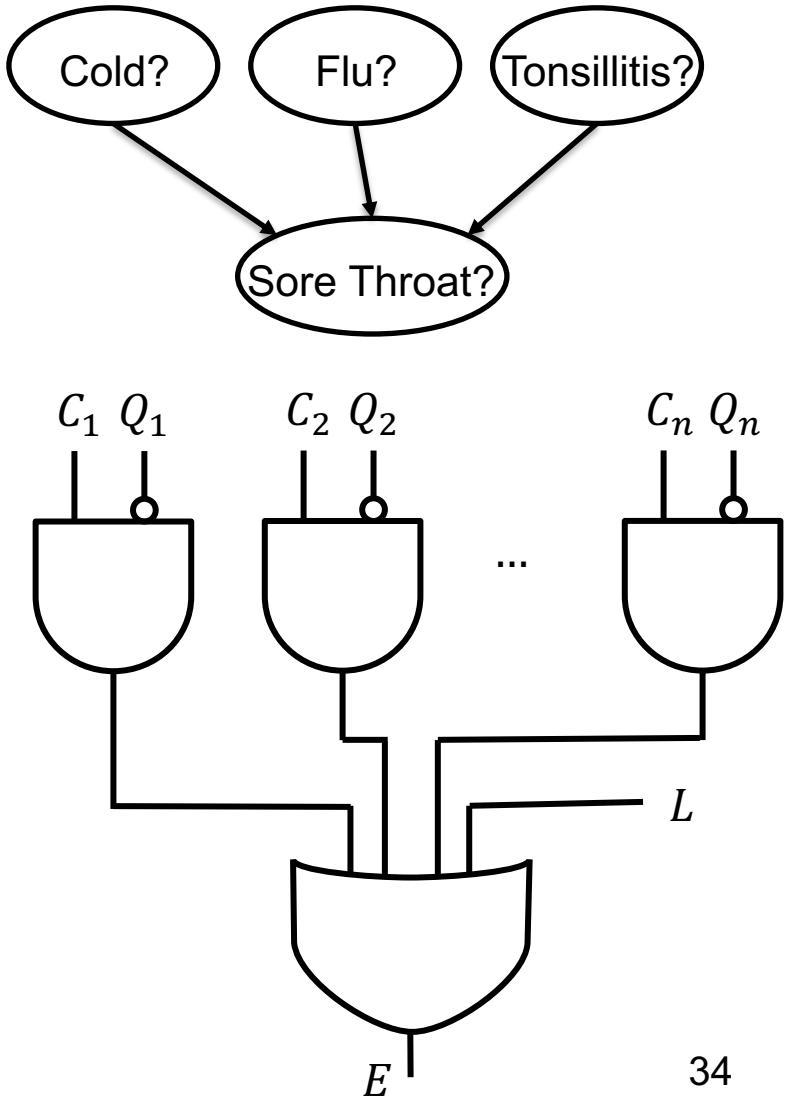
$$P(E = \text{passive}|\alpha) = (1 - \theta_l) \prod_{i \in I\_\alpha} \theta_{q_i}$$



Micro Models: Noisy-or

- Revisiting the medical diagnosis problem
  - Sore throat (S) has three different causes Cold (C), flu (F) and tonsillitis (T)
  - Suppressor probability for C is .15, F is .01, and T is .05
  - Leak probability is .02

C	F	T	S	$\theta_{s\|c,f,t}$	
c	f	t	s	.9999265	$1-(1-.02)(.15)(.01)(.05)$
c	f	$\bar{t}$	s	.99853	$1-(1-.02)(.15)(.01)$
c	$\bar{f}$	t	s	.99265	$1-(1-.02)(.15)(.05)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\bar{c}$	$\bar{f}$	$\bar{t}$	s	.02	$1-(1-.02)$



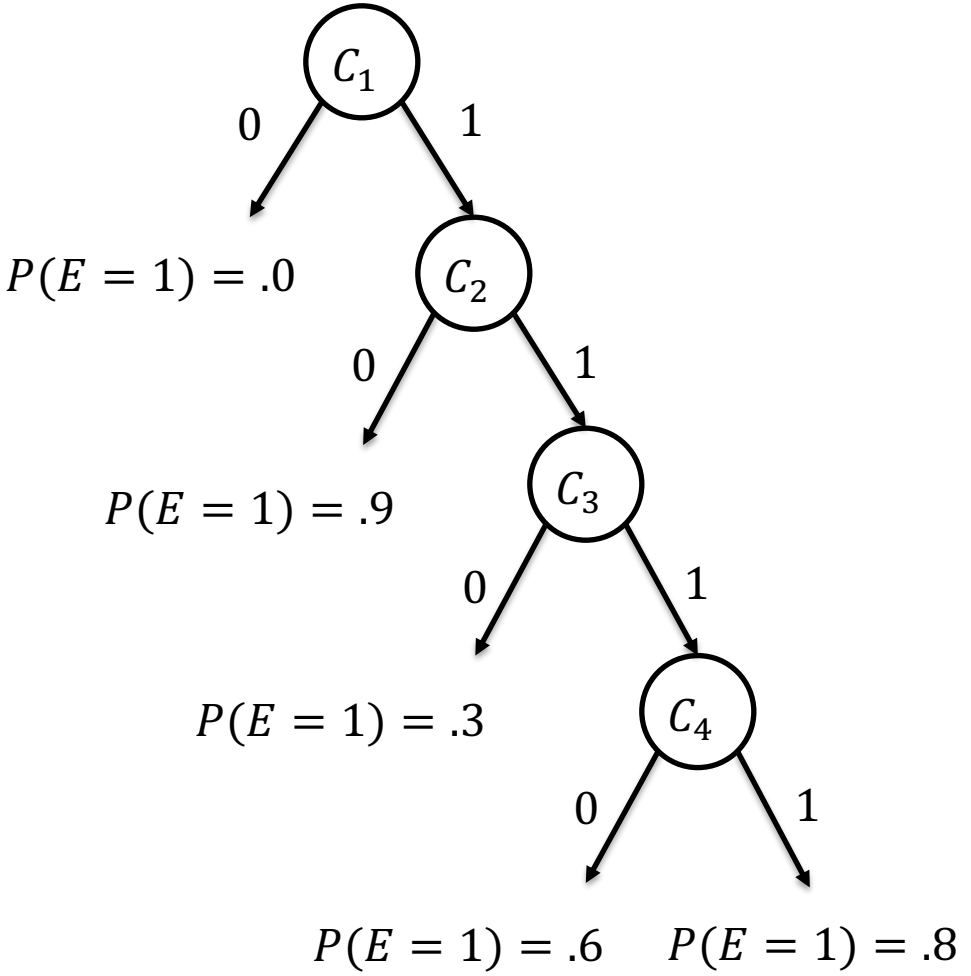
Other Representations

- Many times we have some local structure, but it does not fit any existing micro model such as noisy-or
  - For instance, this CPT has a considerable amount of structure
  - But it does not correspond to the assumptions of a noisy-or model
- For irregular structure, there are several nontabular representation
  - Not necessarily exponential in the number of parents

C_1	C_2	C_3	C_4	$P(E=1)$
1	1	1	1	.0
1	1	1	0	.0
1	1	0	1	.0
1	1	0	0	.0
1	0	1	1	.0
1	0	1	0	.0
1	0	0	1	.0
1	0	0	0	.0
0	1	1	1	.9
0	1	1	0	.9
0	1	0	1	.9
0	1	0	0	.9
0	0	1	1	.3
0	0	1	0	.3
0	0	0	1	.6
0	0	0	0	.8

Decision Trees and Graphs

- Decision tree is a popular representation
  - We start at the root and branch downward depending on the value of the variable
  - It can have a size that is linear in the number of parents if there is enough structure
  - But it may also be exponential if such structure lacks in the CPT



If-then Rules

- A CPT for variable E can be represented using a set of if-then rules

$$\text{If } \alpha_i \text{ then } P(e) = p_i$$

α_i is a propositional sentence with the parent variables of E

If $C_1 = 1$	then $P(E = 1) = 0$
If $C_1 = 0 \wedge C_2 = 1$	then $P(E = 1) = .9$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 1$	then $P(E = 1) = .3$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 0 \wedge C_4 = 1$	then $P(E = 1) = .6$
If $C_1 = 0 \wedge C_2 = 0 \wedge C_3 = 0 \wedge C_4 = 0$	then $P(E = 1) = .8$

- The premisses α_i must be mutually exclusive and exhaustive to avoid conflicting rules and covering every CPT parameter
- This representation can be very efficient but also exponential if the CPT lacks structure

Deterministic CPTs

- A *deterministic* CPT has only either 0 or 1 probabilities
  - They are common in practice
  - When a node has a deterministic CPT, the node is said to be *functionally determined* by its parents
- Deterministic CPTS can be represented compactly using propositional sentences

$$\Gamma_i \Leftrightarrow E = e_i$$

  - We have one rule for each value e_i of E
  - The premisses Γ_i are mutually exclusive and exhaustive
- For example, for the example table

$$(X = \text{ok} \wedge A = \text{high}) \vee X = \text{stuckat0} \Leftrightarrow C = \text{low}$$
$$(X = \text{ok} \wedge A = \text{low}) \vee X = \text{stuckat1} \Leftrightarrow C = \text{high}$$

A	X	C	$\theta_{c\|a,x}$
high	ok	high	0
low	ok	high	1
high	stuckat0	high	0
low	stuckat0	high	0
high	stuckat1	high	1
low	stuckat1	high	1

Conclusion

- In this lecture, we discusses different Bayesian network models
  - Our focus was on modelling and different queries we can pose
  - Differently from several Machine Learning models, Bayesian networks models are comprehensible and not tied to a single task
  - We also discussed methods for large CPT representation. These methods solve a modelling issue. But, many times we have to expand them to make inference
- Tasks
  - Read Chapter 5 from the textbook (Darwiche)