Bayesian Networks (= Belief Networks)

Sven Koenig, USC

Russell and Norvig, 3rd Edition, Sections 14.1-14.4

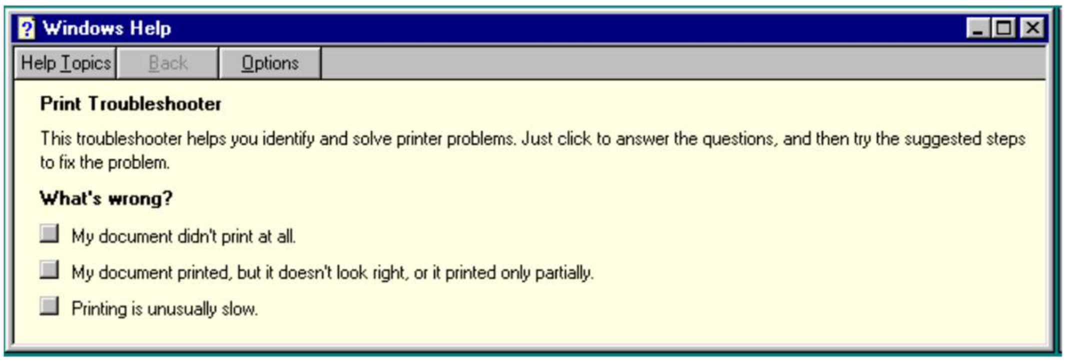
These slides are new and can contain mistakes and typos.
Please report them to Sven (skoenig@usc.edu).

Rule-Based Systems (= Production Systems)

- We now start with probabilistic knowledge representation and reasoning.
- Conclusions are often not certain
  - if OfficeMachine(x) then HasEnergySource(x, WallOutlet)
  - If OfficeMachine(x) then it is highly likely that HasEnergySource(x, WallOutlet)

Bayesian Networks

- Windows 95: diagnosis of printing problems



Bayesian Networks

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

S1	S2	S3	…	D1	D2	D3	…	P(S1, S2, S3, …, D1, D2, D3, …)
true	true	true	…	true	true	true	…	0.0000001
…	…	…	…	…	….	…	…	
false	false	false	…	false	false	false	…	0.0000002

- When the doctor observes presence of S1 and absence of S3, calculate
  - P(D1 | S1, NOT S3) = P(D1, S1, NOT S3) / P(S1, NOT S3)
  - P(D2 | S1, NOT S3)
  - P(D3 | S1, NOT S3)
  - …

Bayesian Networks

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

S1	S2	S3	…	D1	D2	D3	…	P(S1, S2, S3, …, D1, D2, D3, …)
true	true	true	…	true	true	true	…	0.0000001
…	…	…	…	…	….	…	…	
false	false	false	…	false	false	false	…	0.0000002

- We need to acquire too many probabilities from the expert.
- Many of the probabilities are very close to zero and thus hard to specify by experts.

Bayesian Networks

- Medical diagnosis
  - S1, S2, …: symptoms (e.g. high temperature) or causes of diseases (e.g. age)
  - D1, D2, …: diseases (e.g. flu, kidney stone, …)

S1	S2	S3	…	D1	D2	D3	…	P(S1, S2, S3, …, D1, D2, D3, …)
true	true	true	…	true	true	true	…	0.0000001
…	…	…	…	…	….	…	…	
false	false	false	…	false	false	false	…	0.0000002

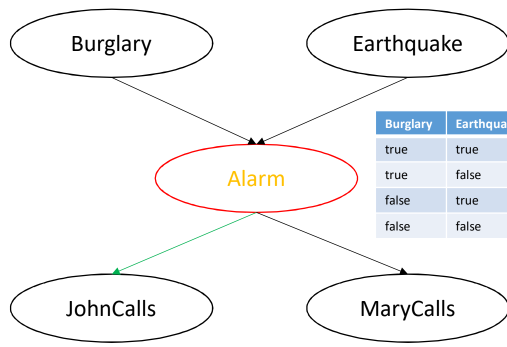
- Bayesian networks make use of conditional independence to specify such a joint probability distribution without these problems.
- Can’t we just assume, for example, pairwise independence?
  No, if diseases were independent from symptoms, then there would be no need to observe any symptoms to perform a medical diagnosis!

Bayesian Networks

- Directed acyclic graph, where nodes are random variables, links are direct influences between random variables, and conditional probability tables specify probabilities

P(Burglary)
P(B) = 0.001

P(Earthquake)
P(E) = 0.002



Expresses unmodeled causes, e.g. trucks passing by, etc.

Burglary	Earthquake	P(Alarm | Burglary, Earthquake)
true	true	P(A | B, E) = 0.95
true	false	P(A | B, NOT E) = 0.94
false	true	P(A | NOT B, E) = 0.29
false	false	P(A | NOT B, NOT E) = 0.001

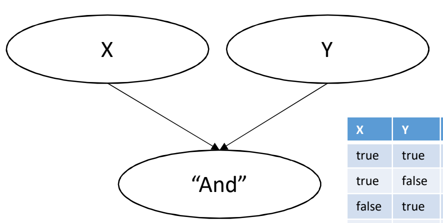
Remember that P(J \| A) + P(J \| NOT A) does not need to equal 1!

Alarm	P(JohnCalls | Alarm)
true	P(J | A) = 0.90
false	P(J | NOT A) = 0.05

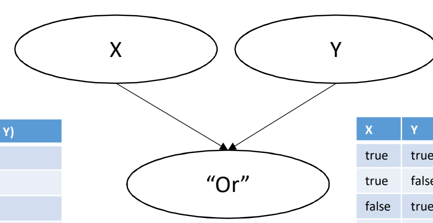
Alarm	P(MaryCalls | Alarm)
true	P(M | A) = 0.70
false	P(M | NOT A) = 0.01

Bayesian Networks

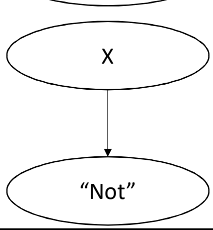
- Can Bayesian networks represent all Boolean functions? – Yes. f(Feature_1, …, Feature_n) ≡ some propositional sentence



X	Y	P("And" | X, Y)
true	true	1.0
true	false	0.0
false	true	0.0
false	false	0.0



X	Y	P("Or" | X, Y)
true	true	1.0
true	false	1.0
false	true	1.0
false	false	0.0



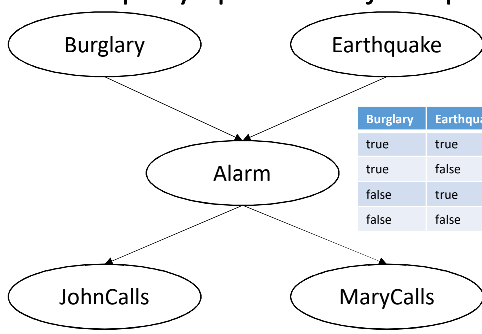
X	P("Not" | X)
true	0.0
false	1.0

Bayesian Networks

- A Bayesian network uniquely specifies a joint probability table

P(Burglary)
0.001

P(Earthquake)
0.002



Burglary	Earthquake	P(Alarm | Burglary, Earthquake)
true	true	0.95
true	false	0.94
false	true	0.29
false	false	0.001

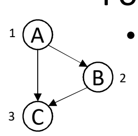
Alarm	P(JohnCalls | Alarm)
true	0.90
false	0.05

Alarm	P(MaryCalls | Alarm)
true	0.70
false	0.01

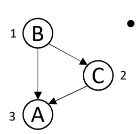
- P(B, E, A, J, M) = P(B) P(E) P(A | B, E) P(J | A) P(M | A)
  for all assignments of truth values to B, E, A, J and M
- P(B, NOT E, NOT A, J, NOT M) = 0.001 (1-0.002) (1-0.94) 0.05 (1 – 0.01)

Bayesian Networks

- A joint probability table does not uniquely specify a Bayesian network since each way of factoring the joint probability distribution corresponds to one Bayesian network structure. Each resulting Bayesian network represents the joint probability distribution correctly for suitably calculated conditional probability tables.
- For example, there are 6 ways of factoring P(A, B, C), including



  - P(A, B, C) = P(C | B, A) P(B, A) = P(C | B, A) P(B | A) P(A) (called the chain rule)
    for all assignments of truth values to A, B and C
    (corresponding to: first picking A, then picking B and finally picking C, each time conditioning the picked random variable on all random variables picked earlier)



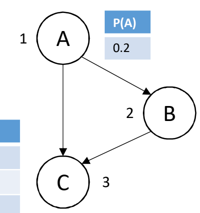
  - P(A, B, C) = P(A | B, C) P(B, C) = P(A | B, C) P(C | B) P(B)
    for all assignments of truth values to A, B and C
    (corresponding to: first picking B, then picking C and finally picking A, each time conditioning the picked random variable on all random variables picked earlier)

Bayesian Networks

- The Bayesian network structure determines how many probabilities need to be specified for the conditional probability tables.
- Let's choose P(A, B, C) = P(C | B, A) P(B | A) P(A).

A	B	C	P(A, B, C)
true	true	true	0.054
true	true	false	0.126
true	false	true	0.002
true	false	false	0.018
false	true	true	0.432
false	true	false	0.288
false	false	true	0.032
false	false	false	0.048

P(A)
0.2



A	P(B | A)
true	0.9
false	0.9

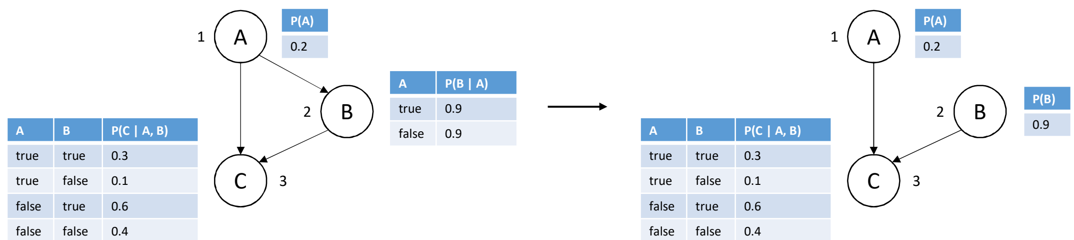
A	B	P(C | A, B)
true	true	0.3
true	false	0.1
false	true	0.6
false	false	0.4

Bayesian Networks

- Here: P(B | A) = P(B | NOT A).
- Thus, A and B are independent since
  - P(B) = P(B AND A) + P(B AND NOT A) =
    P(B | A) P(A) + P(B | NOT A) P(NOT A) =
    P(B | A) P(A) + P(B | A) P(NOT A) =
    P(B | A) (P(A) + P(NOT A)) =
    P(B | A)

Bayesian Networks

- This allows us to simplify the Bayesian network, which requires the specification of only 6 probabilities for all conditional probability tables rather than 7 probabilities for the joint probability table.



P(A)
0.2

A	P(B | A)
true	0.9
false	0.9

A	B	P(C | A, B)
true	true	0.3
true	false	0.1
false	true	0.6
false	false	0.4

Need to specify 7 probabilities for all conditional probability tables

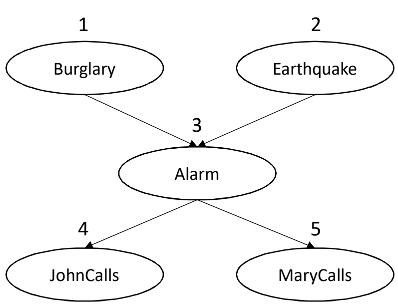
P(A)
0.2

P(B)
0.9

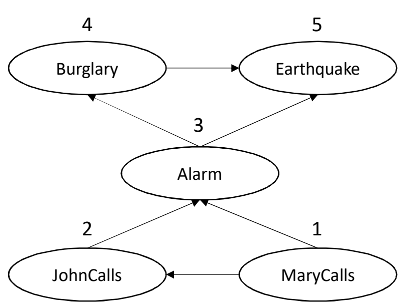
A	B	P(C | A, B)
true	true	0.3
true	false	0.1
false	true	0.6
false	false	0.4

Need to specify only 6 probabilities for all conditional probability tables

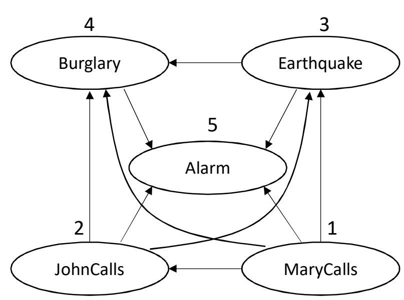
Bayesian Networks



Need to specify 10 probabilities for all conditional probability tables



Need to specify 13 probabilities for all conditional probability tables



Need to specify 31 probabilities for all conditional probability tables

Bayesian Networks

- The Bayesian network structure (that is, the ordering of the random variables) makes a difference for how many probabilities need to be specified for all conditional probability tables.
- We try to find a good ordering by ordering the random variables from causes to effects, which typically works well.
- Example: put first the causes of diseases (e.g. "age"), then the diseases (e.g. "flu"), then the symptoms of the diseases (e.g. "cough"). Note that this cannot be done perfectly since "weight gain" might be the cause of a disease but also a symptom of a disease.

Bayesian Networks

- How to create a Bayesian network with a domain expert
  - Ask the expert for the random variables
  - Ask the expert to order the random variables from cause to effect
  - Repeatedly
    - Create a node for the next random variable in the ordering
    - For each previously created node
      - If the expert states that there should be a link from the previously created node to the newly created node (because there is a "direct influence" from the previously created node to the newly created node), create the link
  - Ask the expert for all probabilities in the conditional probability tables

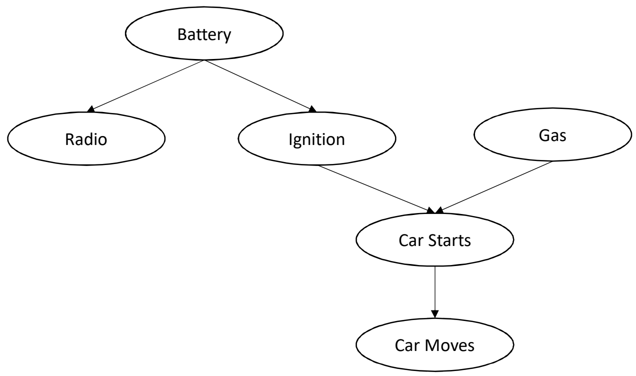
Bayesian Networks

- Warning: The links in a Bayesian network do not need to go from causes to effects in order for the Bayesian network to be correct!
- The links going from causes to effects just help to keep the number of edges and thus the number of probabilities in all conditional probability tables small, which makes it easier to acquire them from an expert and also makes reasoning with them faster.
- In other words, it is smart but not necessary to make the links go from causes to effects.

Bayesian Networks

- A node is conditionally independent of its non-descendants, given its parents.
- A node is conditionally independent of all other nodes, given its parents, children and children's parents (that is, given its Markov blanket).

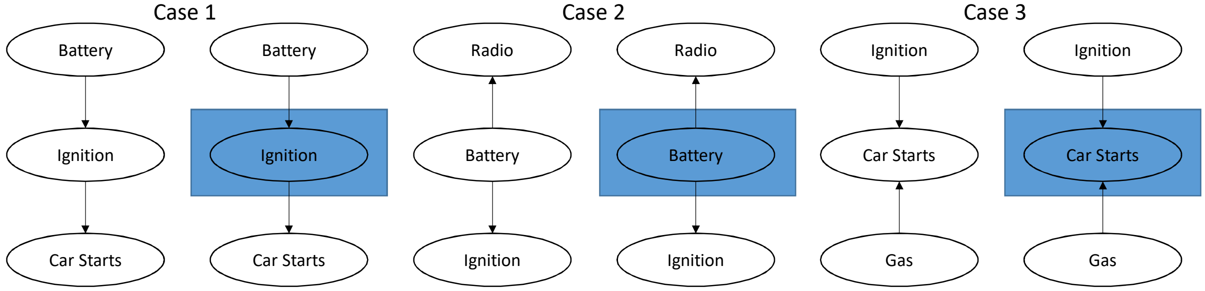
Bayesian Networks: D-Separation




= value of random variable is known

Bayesian Networks: D-Separation

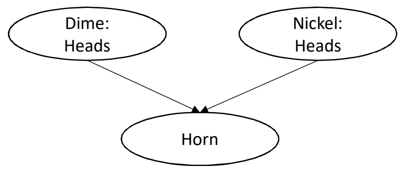
Case 1 | Case 2 | Case 3



"Battery" and "Car Starts" are **not** guaranteed to be independent	"Battery" and "Car Starts" are (guaranteed to be) conditionally independent given "Ignition"	"Radio" and "Ignition" are **not** guaranteed to be independent	"Radio" and "Ignition" are (guaranteed to be) conditionally independent given "Battery"	"Ignition" and "Gas" are (guaranteed to be) independent (provided that **both** "Car Starts" and "Car Moves" are not given)	"Ignition" and "Gas" are **not** guaranteed to be conditionally independent given "Car Starts" (and/or "Car Moves")
	Path blocked		**Path blocked**	**Path blocked**	

Bayesian Networks: D-Separation

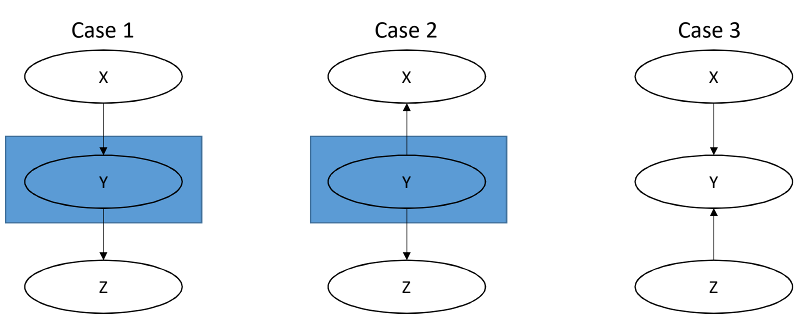
- Example for Case 3:
  In the neighboring room, someone flips both a dime and a nickel. Then, they sound a horn if and only if exactly one of the two coins comes up heads.



- "Dime: Heads" and "Nickel: Heads" are independent.
- However, they are not conditionally independent given "Horn" since
  P(Dime: Heads | Horn) = ½ but
  P(Dime: Heads | Nickel: Heads, Horn) = 0.

[blue box] = value of random variable is known

Bayesian Networks: D-Separation



Case 1	Case 2	Case 3
Path between X and Z is blocked provided that Y is given	**Path between X and Z is blocked provided that Y is given**	**Path between X and Z is blocked provided that neither Y nor any of its descendants are given**

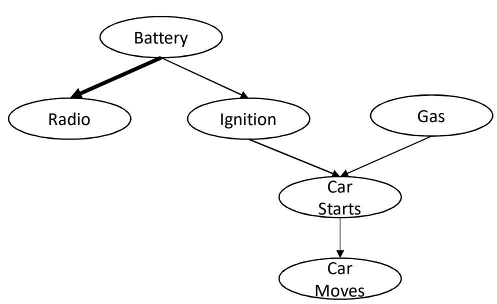
Bayesian Networks: D-Separation

- X and Y are conditionally independent given E if and only if every undirected path (that is, one can go either with or against the directed edges) between them is blocked in at least one part each.

Water Analogy

Bayesian Networks: D-Separation

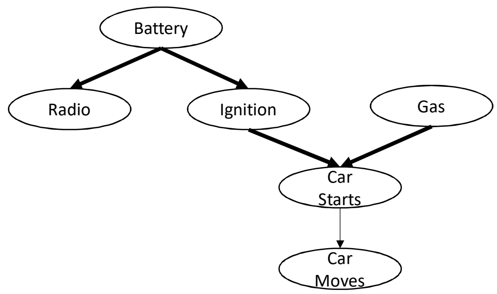
- Example: Are “Radio” and “Battery” independent?



- Perhaps not: There is only one undirected path between “Radio” and “Battery”, and this path is not blocked. (A path that consists of one link only cannot be blocked.) Thus, it depends on the conditional probability tables whether they are independent.

Bayesian Networks: D-Separation

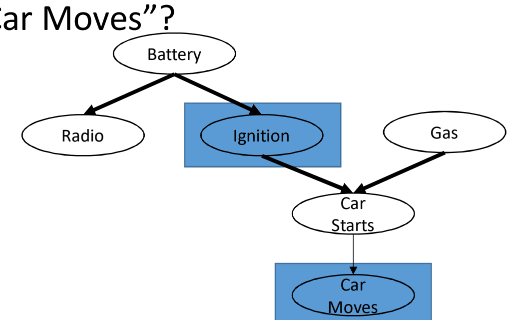
- Example: Are “Radio” and “Gas” independent?



- Yes: There is only one undirected path between “Radio” and “Gas”, and this path is blocked because its part “Ignition → Car Starts ← Gas” is blocked. (This is the only blocked part.)

Bayesian Networks: D-Separation

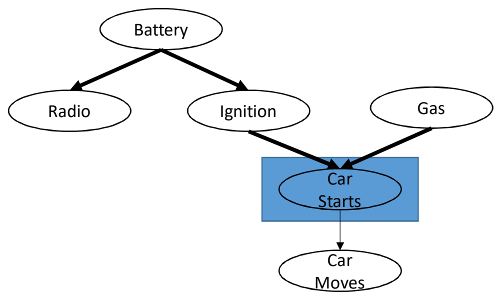
- Example: Are “Radio” and “Gas” conditionally independent given “Ignition” and “Car Moves”?



- Yes: There is only one undirected path between “Radio” and “Gas”, and this path is blocked because its part “Battery → Ignition → Car Starts” is blocked. (This is the only blocked part.)

Bayesian Networks: D-Separation

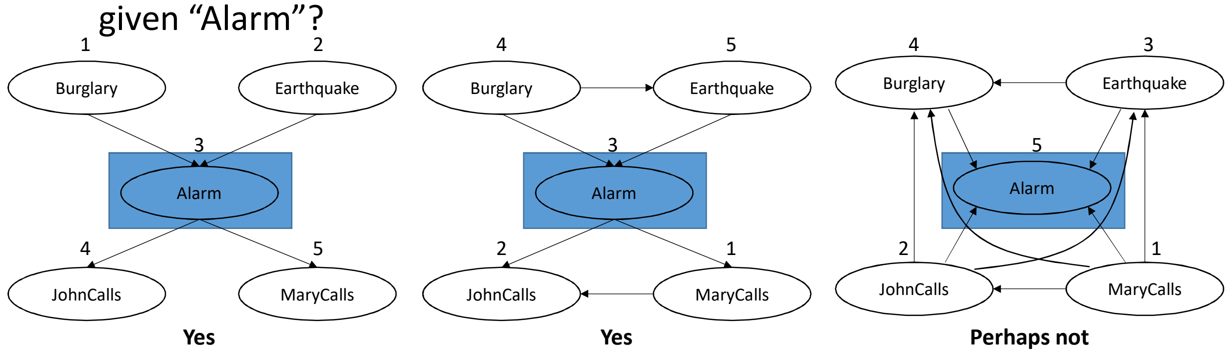
- Example: Are "Radio" and "Gas" conditionally independent given "Car Starts"?



- Perhaps not: There is only one undirected path between "Radio" and "Gas", and this path is not blocked anywhere. Thus, it depends on the conditional probability tables whether they are conditionally independent.

Bayesian Networks: D-Separation

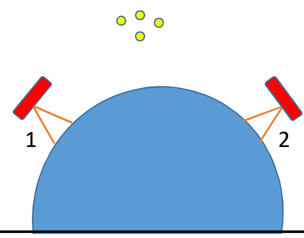
- Example: Are "Burglary" and "JohnCalls" conditionally independent given "Alarm"?



- This is another reason why we like Bayesian networks with few edges: one can read off more (conditional) independence relationships from the Bayesian network structure.

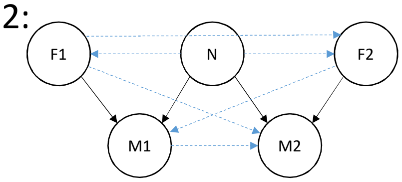
Bayesian Networks

- Two astronomers, in different parts of the world, make measurements M1 and M2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small possibility of error by up to one star. Each telescope can also (with a slightly smaller probability) be badly out of focus (events F1 and F2), in which case the scientist will undercount by three or more stars (Problem 14.12 in Russell and Norvig).



Bayesian Networks

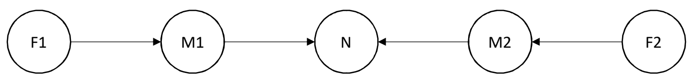
- Two astronomers, in different parts of the world, make measurements M1 and M2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small possibility of error by up to one star. Each telescope can also (with a slightly smaller probability) be badly out of focus (events F1 and F2), in which case the scientist will undercount by three or more stars.
- You want to generate the following Bayesian network since F1 and N cause M1 and F2 and N cause M2, so a good ordering is (for example) N, F1, F2, M1 and M2:



Dashed links should NOT be put in because there is no direct influence and they thus do not need to be put in!

Bayesian Networks

- Two astronomers, in different parts of the world, make measurements M1 and M2 of the number of stars N in some small region of the sky, using their telescopes. Normally, there is a small possibility of error by up to one star. Each telescope can also (with a slightly smaller probability) be badly out of focus (events F1 and F2), in which case the scientist will undercount by three or more stars.
- Argue that the following Bayesian network structure is incorrect (that is, there are no conditional probability tables for it that result in a Bayesian network that models the described situation correctly):



Bayesian Networks

- You cannot argue that the links do not go from causes to effects.
- You cannot argue that independence relationships present in the described situation are not present in the Bayesian network since they could be correctly present in the conditional probability tables. In other words, Bayesian network topologies can express only the presence of independence relationships, not their absence.

Bayesian Networks

- Instead, you need to argue that the independence relationships present in the Bayesian network structure are not present in the described situation, for example:
  - D-separation states that, in the Bayesian network structure, F1 and N are conditionally independent given M1. However, if M1 is known to be 1000 in the described situation, then learning that N is 2000 increases the probability that F1 is true to one. Thus, F1 and N are not necessarily conditionally independent given M1.
  - D-separation states that, in the Bayesian network structure, M1 and M2 are independent if N is not given. However, if F1 and F2 are known to be false in the described situation, then learning that M1 is 1000 increases the probability that N is in the range 999-1001 to one, which in turn increases the probability that M2 is in the range 998-1002 to one. Thus, M1 and M2 are not necessarily independent if N is not given.

Bayesian Networks

- There are a number of algorithms that can calculate conditional probabilities, such as P(D1 | S1, NOT S3), for a given Bayesian network. There is also good software available where one sets known values, e.g. S1 to true and S3 to false, and then queries other nodes, e.g. D1 to obtain P(D1 | S1, NOT S3).
- In the following, we are content to perform a couple of probability calculations by hand.

Bayesian Networks

A	P(C|A)
true	0.8
false	0.3

P(A)
0.4

A	P(B|A)
true	0.8
false	0.3

B	P(D|B)
true	0.8
false	0.3

- Easy probability calculations:
  - P(B | NOT A) = 0.3
  - P(NOT B | A) = 1 – P(B | A)
  - P(NOT B | NOT A) = 1 – P(B | NOT A) = 0.7
  - P(C) = P(A, C) + P(NOT A, C) = P(C | A) P(A) + P(C | NOT A) P(NOT A) = 0.8 0.4 + 0.3 0.6 = 0.5
  - P(A | C) = P(A, C) / P(C) = P(C | A) P(A) / P(C) = 0.8 0.4 / 0.50 = 0.64

Bayesian Networks

A	P(C|A)
true	0.8
false	0.3

P(A)
0.4

A	P(B|A)
true	0.8
false	0.3

B	P(D|B)
true	0.8
false	0.3

- Probability calculations that make use of d-separation:
  - P(D | A) = P(A, D) / P(A) =
    (P(A, B, D) + P(A, NOT B, D)) / P(A) =
    P(D | A, B) P(A, B) / P(A) + P(D | A, NOT B) P(A, NOT B) / P(A) =
    P(D | A, B) P(B | A) + P(D | A, NOT B) P(NOT B | A) =
    P(D | B) P(B | A) + P(D | NOT B) P(NOT B | A) = 0.8 0.8 + 0.3 0.2 = 0.7,
    where P(D | A, B) = P(D | B) and P(D | A, NOT B) = P(D | NOT B) follows from d-separation

Bayesian Networks

A	P(C|A)
true	0.8
false	0.3

P(A)
0.4

A	P(B|A)
true	0.8
false	0.3

B	P(D|B)
true	0.8
false	0.3

- Probability calculations that make use of d-separation:
  - P(B, C) = P(A, B, C) + P(NOT A, B, C) =
    P(B, C | A) P(A) + P(B, C | NOT A) P(NOT A) =
    P(B | A) P(C | A) P(A) + P(B | NOT A) P(C | NOT A) P(NOT A) =
    0.8 0.8 0.4 + 0.3 0.3 0.6 = 0.31,
    where P(B, C | A) = P(B | A) P(C | A) and P(B, C | NOT A) = P(B | NOT A) P(C | NOT A) follows from d-separation

Bayesian Networks

- Whenever you need to calculate probabilities in exams, you can try to simply transform the given Bayesian network into a joint probability table and then calculate the probabilities from the joint probability table, which is typically conceptually very easy. In real life, however, the probability tables are often way to large to do this efficiently, which is why we learned about Bayesian networks in the first place!

Bayesian Networks

- Want to play around with Bayesian networks?
- Go here: http://aispace.org/bayes/