Machine Learning (CS 181):
18. Graphical Models and Bayesian Networks

David C. Parkes and Sasha Rush

Spring 2017

Contents

Contents

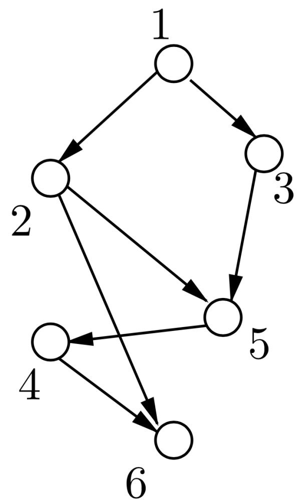
Graphical Probabilistic Models

We have already seen graphical probabilistic models for topic models and hidden Markov models.

Graphical models are useful for the following reasons:

- Provide a compact representation of a joint distribution on a large number of random variables
- Enable tractable inference (i.e., what is the probability of some variables given others?); e.g., we saw this with the forward-backward algorithm for HMMs.

Directed Graphical Models: Bayesian Networks



Wainwright & Jordan'08

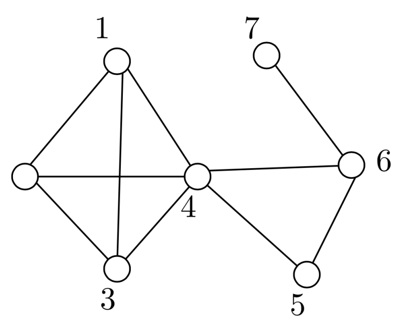
A Bayesian network is a directed, acyclic graph on random variables.

Defines a joint distribution that is the product of local dependencies:

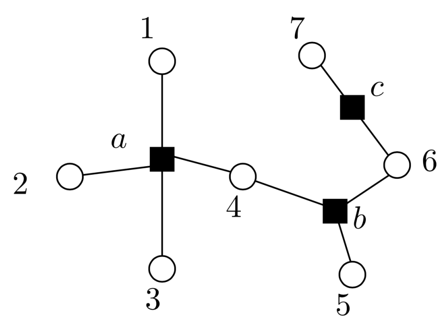
$$p(\mathbf{x}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_5 \mid x_2, x_3)p(x_4 \mid x_5)p(x_6 \mid x_2, x_4)$$

Undirected Models: MRFs (and Factor Graphs)

Markov Random Field



Factor graph



Wainwright & Jordan'08

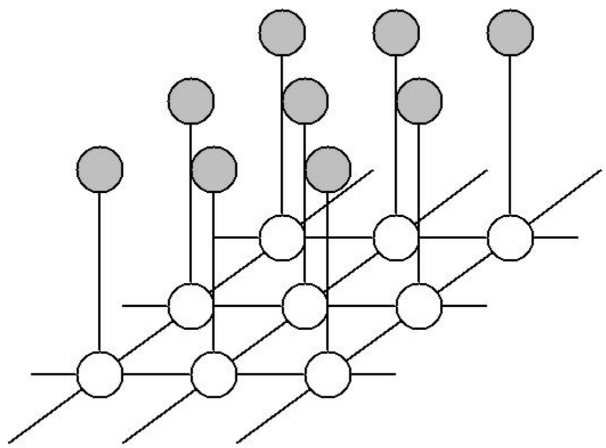
A Markov random field is an undirected graph on random variables.

Defines a joint distribution that is the product of potential functions ψ (≥ 0) over maximal cliques:

$$p(\mathbf{x}) \propto \psi_a(x_1, x_2, x_3, x_4)\psi_b(x_4, x_5, x_6)\psi_c(x_6, x_7)$$

Equivalent factor graph representation, in which the 'factors' are shown as small squares and represent potential functions.

Example 1: Image Restoration (1 of 4)



(Peter Orchard)

MRF ('Ising model'). Observed data $x_j \in \{-1, +1\}$ represents pixels. Latent variables $z_j \in \{-1, +1\}$ are the true pixel values.

Correlation between latent neighbors: $\psi_a(x_j, x_k) = e^{-\beta x_j x_k}$ ($\beta > 0$).Correlation between latent pixel and observed pixel $\psi_b(x_j, z_j) = e^{-\eta x_j z_j}$ ($\eta > 0$).

Example 1: Image Restoration (2 of 4)

P. Orchard

The original image (continuous, gray scale version of the model).

Example 1: Image Restoration (3 of 4)

P. Orchard

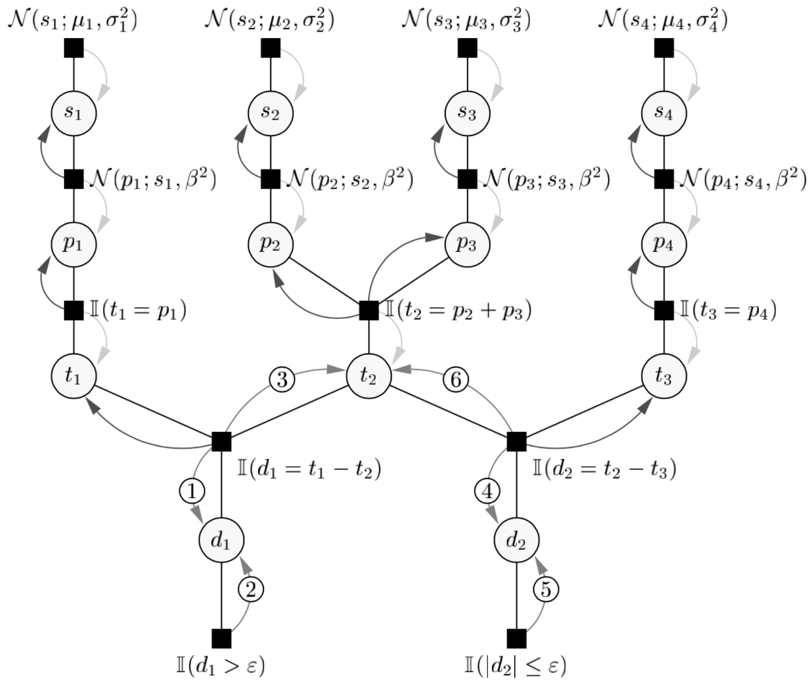
The image with added Gaussian noise.

Example 1: Image Restoration (4 of 4)

P. Orchard

Works well on surfaces of slowly varying intensity (e.g., road, side of buildings, sky.) Fails with thin, sharp features; e.g., telephone cables almost completely removed!

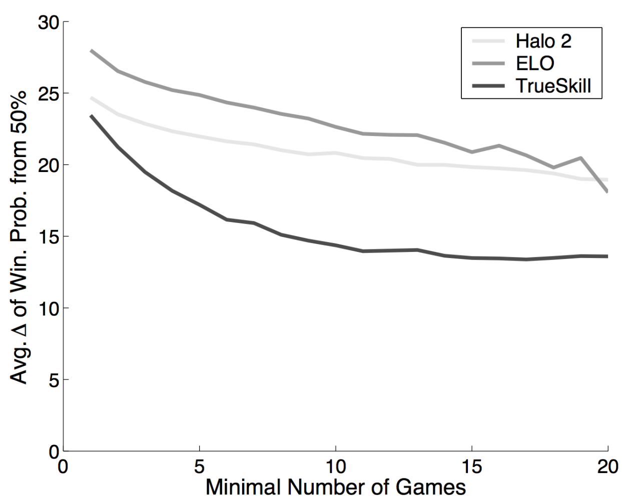
Example 2: TrueSkill (1 of 2) (Herbrich, Minka and Graepel, 2007)



3 teams, $\{1\}$, $\{2,3\}$,$\{4\}$. For graph for: team 1 wins, teams 2 and 3 draw. (ignore the directed edges.)

Variables: s_j: player skill, p_j: player performance, t_j: team performance, d_j: difference in performance.

Example 2: TrueSkill (2 of 2)

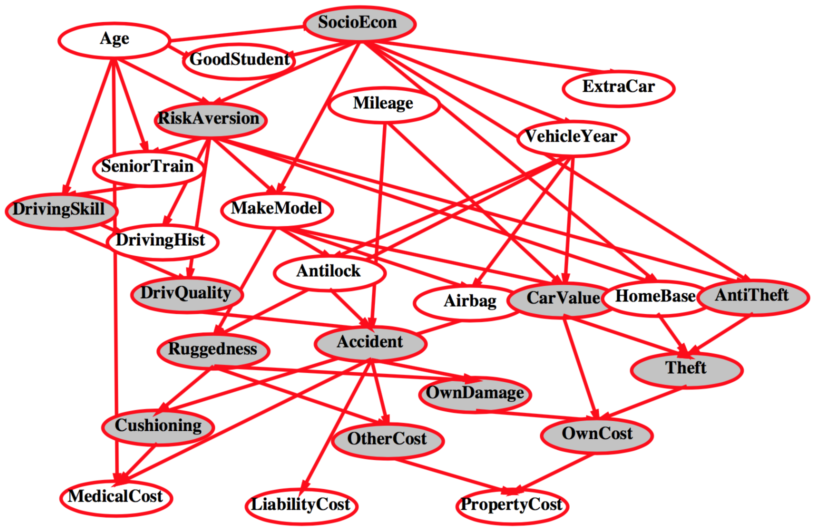


Average deviation of winning prob. from 50%. Lower deviation is better.

ELO = Chess rating system.

- Perceived quality of rating system is in terms of winning ratio (50% is ideal!): if too high, then opposition too weak (and vice versa)
- Deployed in Xbox 360 Live, providing automatic player rating and matchmaking. Processes hundreds of thousands games per day.

Example 3: Car Insurance



(P. Domingos)

Note: The latent variables are shown shaded (sorry!).
Predict the "cost" variables; e.g., the distribution on MedicalCost for an adolescent driving a car with 50,000 miles and no anti-lock brakes.

Bayesian Networks

- Directed, acyclic graphs
- We focus on discrete random variables (but they can be continuous in general).

BNs provide a natural and compact representation of joint distributions, and can also support efficient inference and learning.

Markov random fields are only introduced above as another important family, and for the example applicatons. We don't study them in detail!

Contents

A Notation for Sets of Random Variables

M. Paskin

It is helpful when working with large, complex models to have a good notation for sets of random variables.

- Let X_j denote the jth random variable, and for $A \subseteq \{1, \ldots, m\}$, let $X_A = \{X_i \ : \ i \in A\}$.
- Let X denote the set of all RVs.
- Let $\mathbf{x}_A$ denote a vector of realized values for variables X_A; with $\mathbf{x} = [x_1, \ldots, x_m]^\top$ as always.

Example. If $A = \{1, 5\}$ then $X_A = \{X_1, X_5\}$.

Kinds of questions we're interested in

- Given evidence $E \subseteq \{1, \ldots, m\}$, s.t. values $\mathbf{x}_E$ are observed, what is

$$p(x_q \,|\, \mathbf{x}_E; \mathbf{w})$$

  for query $q \in \{1, \ldots, m\} \setminus E$, and a Bayesian network with parameters $\mathbf{w}$?
- More generally, compute $p(\mathbf{x}_Q \,|\, \mathbf{x}_E; \mathbf{w})$ for $Q \subseteq \{1, \ldots, m\} \setminus E$.
- How can we learn a Bayesian Network representation of a joint distribution?

Note: We will not carry around parameters $\mathbf{w}$ explicitly for the rest of these notes.

Independence is rare in complex systems

M. Paskin

Independence (e.g., two coin flips):

$$P(X_1) = P(X_1 \,|\, X_2)$$

- Independence is useful, because it allows us to reason about aspects of a system in isolation. But quite rare!
- A generalization is conditional independence, where two aspects become independent once we observe a third aspect.
- This often does arise and can lead to significant representational and computational savings.

Reasoning about conditional independence

M. Paskin

- X_1 and X_2 are conditionally independent given X_3 if and only if

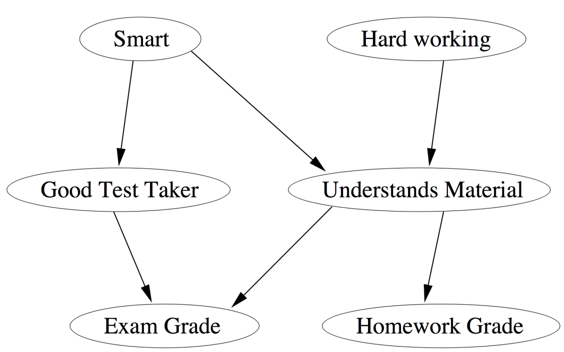
$$p(X_1 \mid X_3) = p(X_1 \mid X_2, X_3)$$

- Write this as $I(X_1, X_2 \mid X_3)$. Examples:
  - operation of a car's starter motor and radio are conditionally independent given the status of the battery
  - symptoms are conditionally independent given the disease
  - future and past are conditionally independent given the present

An intuitive test for $I(X_1, X_2 \mid X_3)$:

> *Imagine you know the value of X_3 and you are trying to guess the value of X_1. Would opening an envelope containing the value of X_2 help you guess X_1? If not, then $I(X_1, X_2 \mid X_3)$.*

Bayesian Network: Classroom example



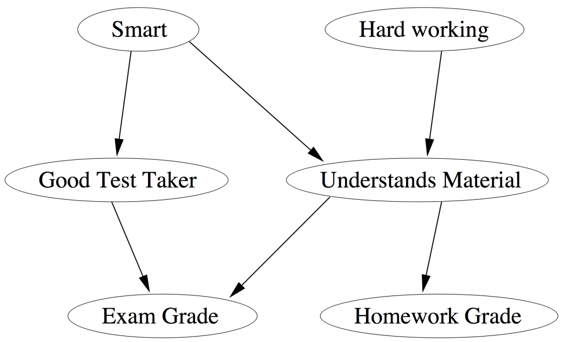
S	HW	Understands Material *true*	Understands Material *false*
true	*true*	0.95	0.05
true	*false*	0.6	0.4
false	*true*	0.6	0.4
false	*false*	0.2	0.8

UM	Homework Grade A	B	C	D	F
true	0.7	0.25	0.03	0.01	0.01
false	0.2	0.3	0.4	0.05	0.05

- Set of variables $\{S, HW, GTT, U, EG, HG\}$. Each has a finite domain (e.g., $Dom(S) = \{T, F\}$, $Dom(HG) = \{A, B, C, D, E\}$). Let $Pa(X_j)$ denote set of parents of X_j.
- A Bayesian Network captures qualitative (via graph structure) and quantitative information (via conditional probability tables).

Local Semantics

- Local independence: every variable is conditionally independent of its non-descendants given (only) its parents
- Say that if an edge exists between X_1 and X_2, the variables are "directly related."

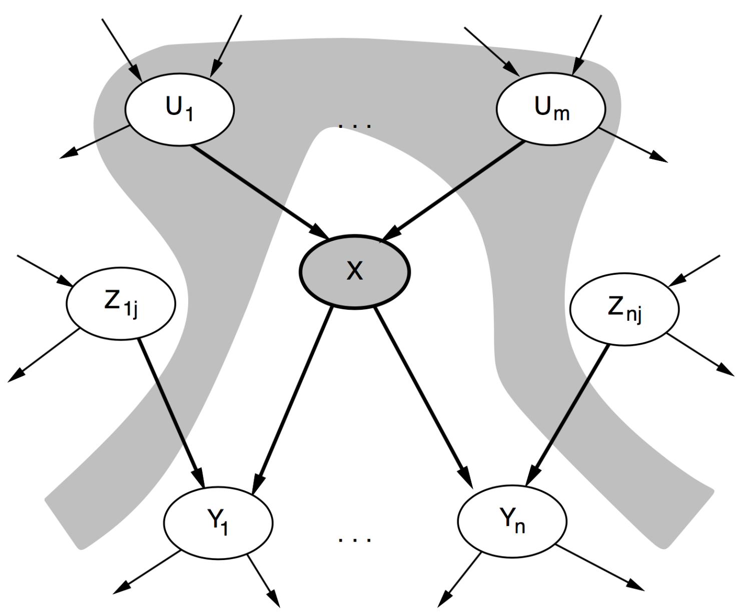


Examples:
$I(EG, S \mid GTT, U)$
$I(GTT, HW \mid S)$
$I(U, GTT \mid S, HW)$

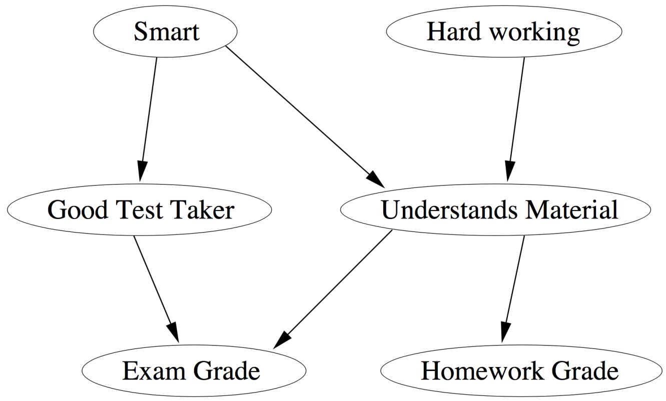
Local Semantics

P. Domingos

Local independence: every variable is conditionally independent of its non-descendants given (only) its parents.



Where we're going: More complex relationships



We don't yet know how to answer the following:

- $I(EG, HG \mid S, HW)$?
- $I(GTT, U \mid S, EG)$?

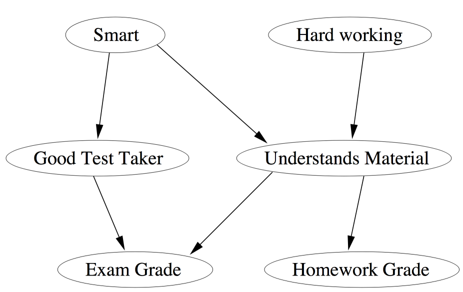
Global Semantics

Chain rule: For an ordering of variables, say X_1, X_2, X_3, X_4, we have

$$p(\mathbf{x}) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)$$

Definition (topological ordering)

In a topological ordering of a directed graph, if variable X_j is a parent of X_k, then X_j is before X_k in the ordering.



BNs are acyclic $\Rightarrow$ topological orderings exist.

A possible topological ordering is S, GTT, HW, U, EG, HG.

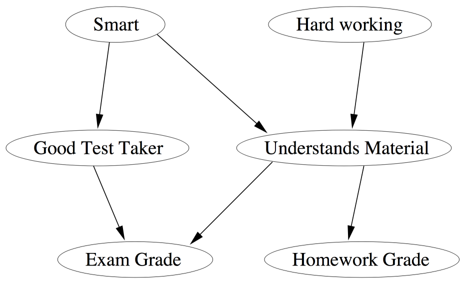
Global Semantics

Theorem

The joint probability distribution that corresponds to a BN is

$$p(\mathbf{x}) = \prod_{j=1}^{m} p(x_j \mid Pa(X_j)).$$

Proof: cyclic ordering + chain rule. In our example:



ordering S, GTT, HW, U, EG, HG.

$$\begin{aligned} p(X) &= p(S)p(GTT \mid S)p(HW \mid S, GTT)p(U \mid S, GTT, HW) \\ &\quad p(EG \mid S, GTT, HW, U)p(HG \mid S, GTT, HW, U, EG) \\ &= p(S)p(HW)p(GTT|S)p(U|S, HW)p(EG|GTT, U)p(HG \mid U) \end{aligned}$$

Generative Process

How can we sample from the distribution defined by a Bayesian network?

Generative process:

- For each node X_j in topological order:
  - Let $\mathbf{u}$ be the previously generated values for $Pa(X_j)$.
  - Sample a value for X_j according to the distribution $p(x_j \mid Pa(X_j) = \mathbf{u})$.

This provides another way to interpret a Bayesian Network.

Contents

Constructing a Bayesian Network

The basic approach is as follows:

1. Choose an ordering over the variables. (Say $X_1, \ldots, X_m$).
2. For each variable X_j in order:
   (a) Find a minimum subset of the preceding variables $X_1, \ldots, X_{j-1}$ for which $p(x_j \,|\, x_1, \ldots, x_{j-1}) = p(x_j \,|\, Pa(X_j))$. Make this subset the parents of X_j.
   (b) Determine the conditional probability table for X_j given its parents (e.g., via learning from data.)

The BN constructed for <u>any</u> such ordering is correct. What varies is the compactness of the resulting network.

Exercise: Alarms, Earthquakes, Burglars

(J. Pearl)

I'm at work, neighbor John calls to say my alarm is ringing, but Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

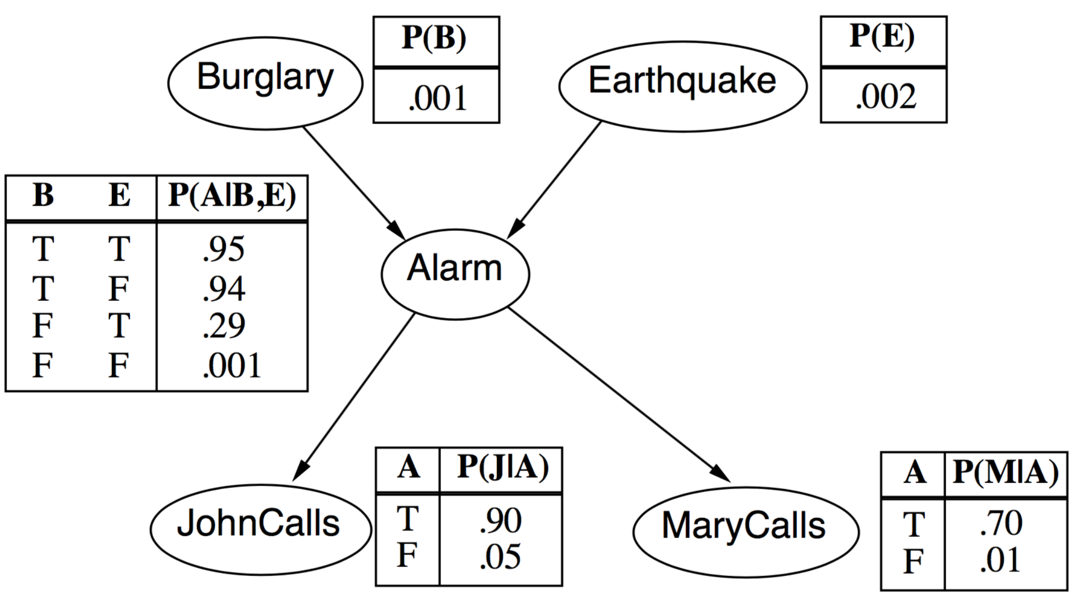
- Variables: Burglar, Earthquake, Alarm, John, Mary
- Causal knowledge (tends to be useful for determining an ordering):
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Suggests 'causal ordering' of B, E, A, J, M.

Ordering B, E, A, J, M

(P. Domingos)

Bayesian Network:



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

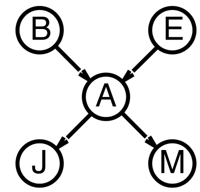
A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness

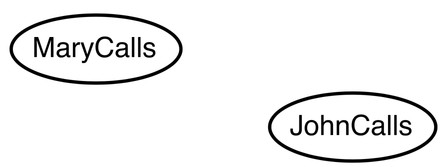
(P. Domingos)

- A CPT for binary X_j with d binary parents has 2^d rows.
- If each variable has at most d parents, then network requires $O(n \cdot 2^d)$ parameters; i.e., it grows linearly with n.
- For the burglary network, we have $1+1+4+2+2 = 10$ parameters.



Suppose we Choose Ordering M, J, A, B, E

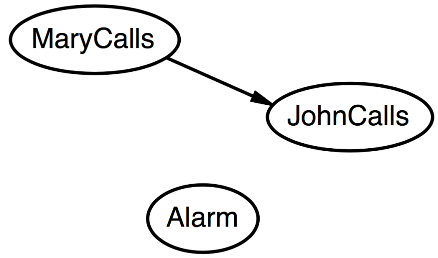
(P. Domingos)



$p(J \mid M) = p(J)$?

Suppose we Choose Ordering M, J, A, B, E

(P. Domingos)

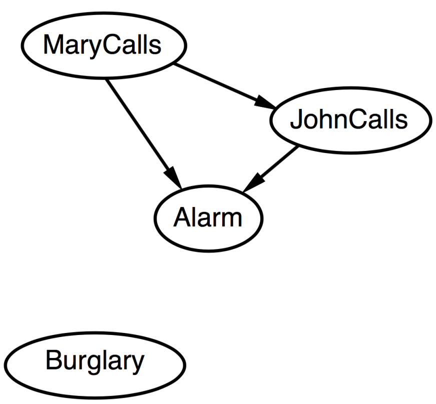


$p(J \mid M) = p(J)$? No

$p(A \mid J, M) = p(A \mid J)$? $p(A \mid J, M) = p(A)$?

Suppose we Choose Ordering M, J, A, B, E

(P. Domingos)



$p(J \mid M) = p(J)$? No

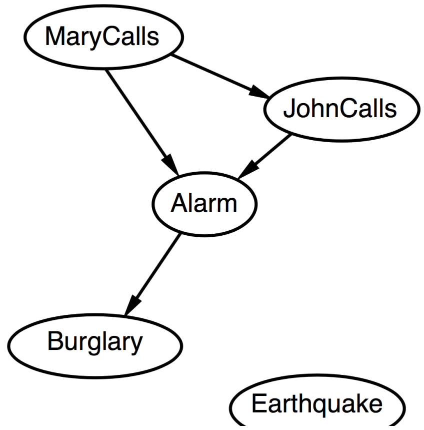
$p(A \mid J, M) = p(A \mid J)$? No. $p(A \mid J, M) = p(A)$? No

$p(B \mid A, J, M) = p(B \mid A)$?

$p(B \mid A, J, M) = p(B)$?

Suppose we Choose Ordering M, J, A, B, E

(P. Domingos)



$p(J \mid M) = p(J)$? No

$p(A \mid J, M) = p(A \mid J)$? No. $p(A \mid J, M) = p(A)$? No

$p(B \mid A, J, M) = p(B \mid A)$? Yes

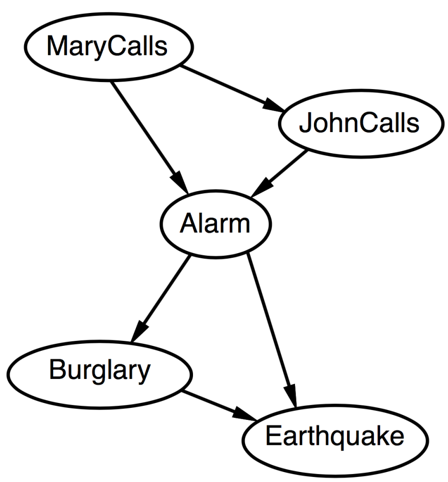
$p(B \mid A, J, M) = p(B)$? No

$p(E \mid B, A, J, M) = p(E \mid B, A)$?

$p(E \mid B, A, J, M) = p(E \mid A)$?

Suppose we Choose Ordering M, J, A, B, E

(P. Domingos)



$p(J \mid M) = p(J)$? No

$p(A \mid J, M) = p(A \mid J)$? $p(A \mid J, M) = p(A)$? No

$p(B \mid A, J, M) = p(B \mid A)$? Yes

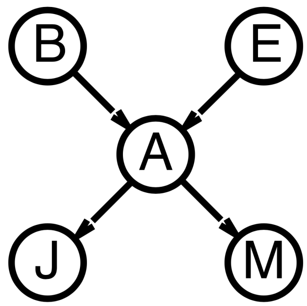
$p(B \mid A, J, M) = p(B)$? No

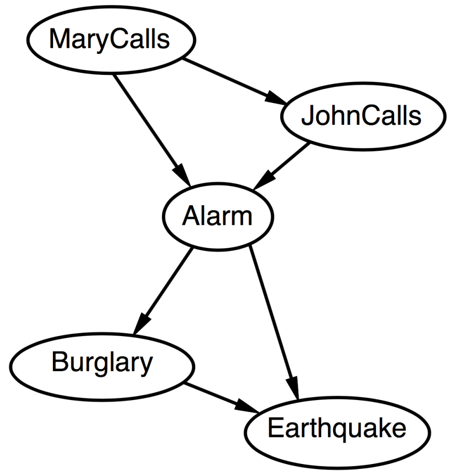
$p(E \mid B, A, J, M) = p(E \mid B, A)$? Yes

$p(E \mid B, A, J, M) = p(E \mid A)$? No

Suppose we Choose Ordering M, J, A, B, E

(P. Domingos)





- Both BayesNets are correct, but the ordering matters!
- Network is less compact in second ordering: $1 + 2 + 4 + 2 + 4 = 13$ parameters.
- Deciding conditional independence is also harder.

Contents

Learning Bayesian Networks

Data $D = \{\mathbf{x}_i\}_{i=1}^n$. With known structure:

- If complete data (all variables observed), then use MLE to estimate parameters in conditional probability tables. Use EM if we have incomplete data.

With unknown structure, and complete data:

- For an initial structure, estimate the parameters and the compute the likelihood of the data for the structure
- Adopt a local search through structures, to modify the current structure to find the adjacent structure with the best likelihood (e.g., add, delete or reverse an edge). Use regularization to avoid overfitting.

Contents

Conclusion

- Bayesian networks provide a compact representation of distributions on lots of variables.
- Focused here on the local and global semantics, and how to construct from an order.
- Markov random fields also provide a popular, probabilistic graphical model (more details out of scope!)

Next lecture: reasoning patterns, d-separation, and inference— what is probability of x_q given evidence $\mathbf{x}_E$?