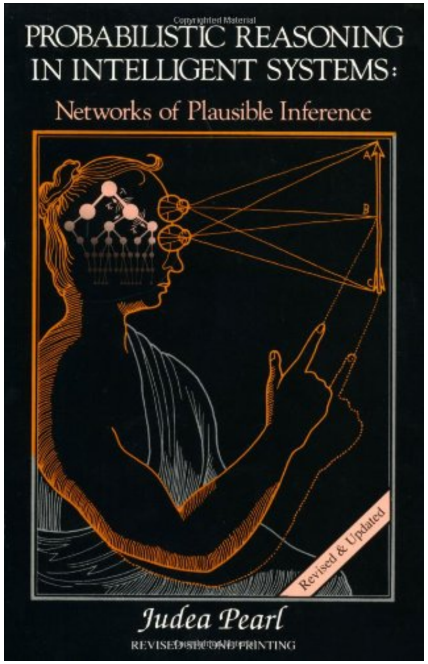
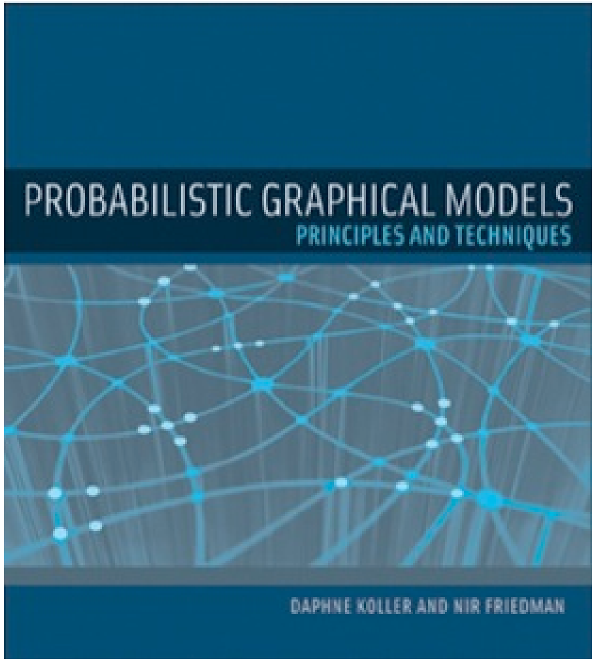
Bayesian networks (BNs)

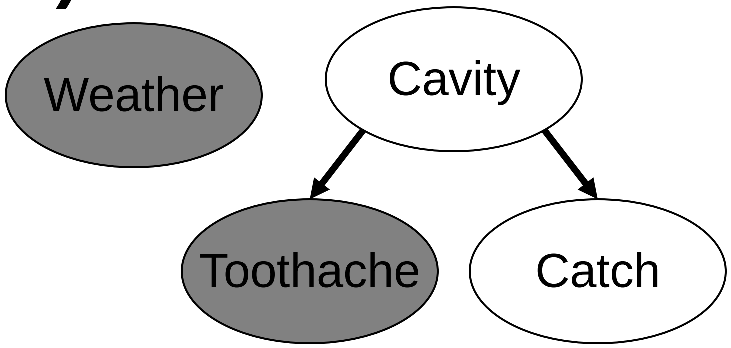
- A type of *graphical model*
- A BN states conditional independence relationships between random variables
- Compact specification of full joint distributions





Syntax

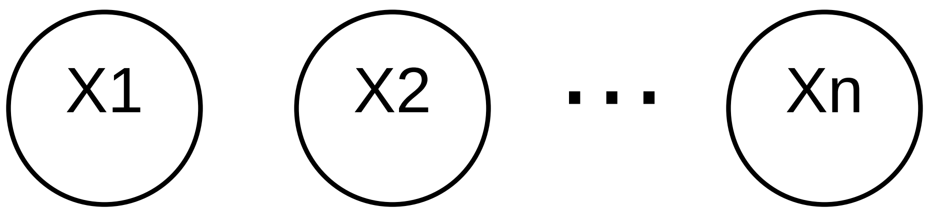
- **Directed Acyclic Graph (DAG)**
- **Nodes:** random variables
  - Can be assigned (observed) or unassigned (unobserved)



- **Arcs:** interactions
  - An arrow from one variable to another indicates direct influence
  - Encode conditional independence
    - *Weather* is independent of the other variables
    - *Toothache* and *Catch* are conditionally independent given *Cavity*
  - Must form a directed, *acyclic* graph

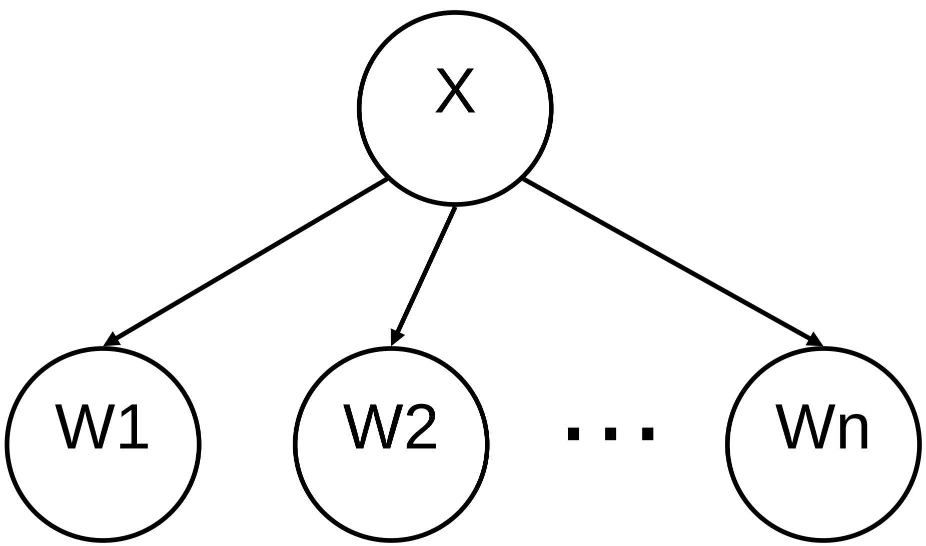
Example: N independent coin flips

- Complete independence: no interactions



Example: Naïve Bayes document model

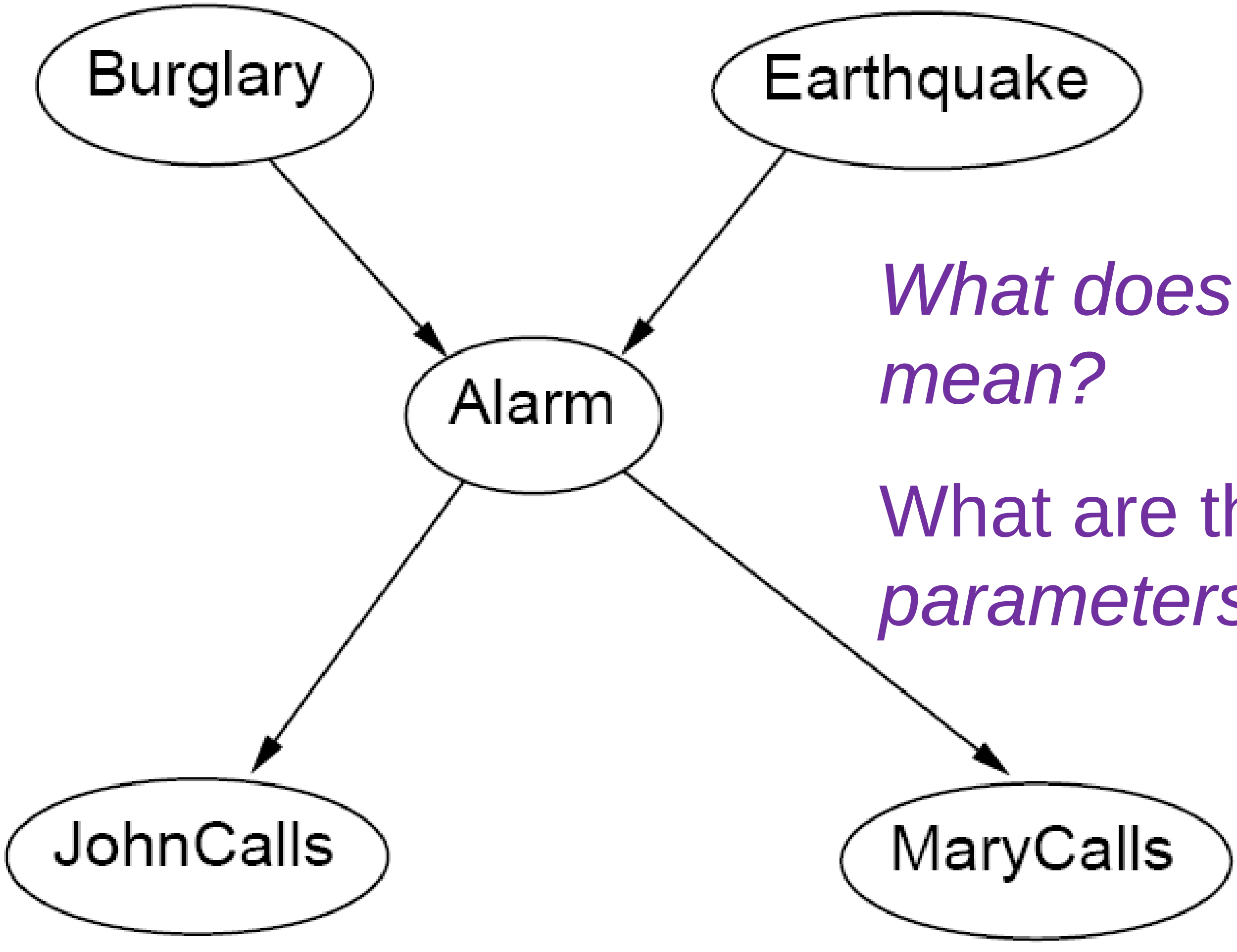
- Random variables:
  - X: document class
  - W1, …, Wn: words in the document



Example: Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- What are the direct influence relationships?
  - 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

Example: Burglar Alarm



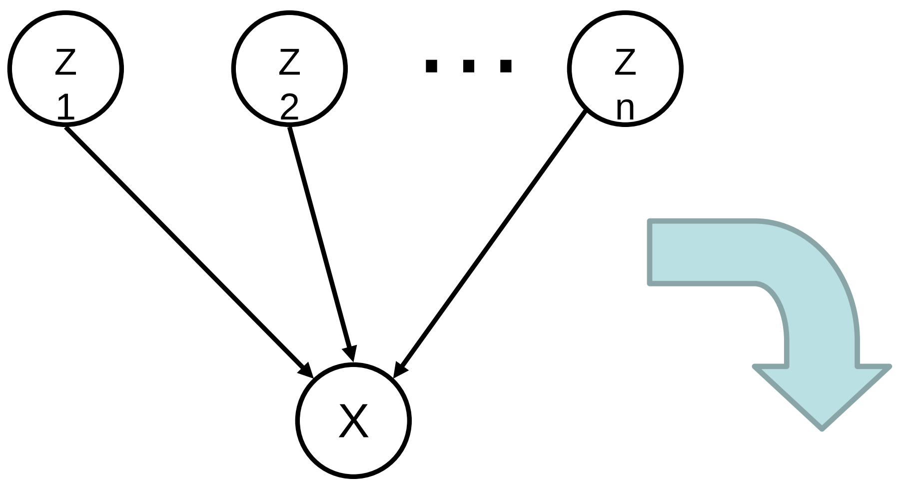
What does this mean?

What are the *model parameters*?

Semantics

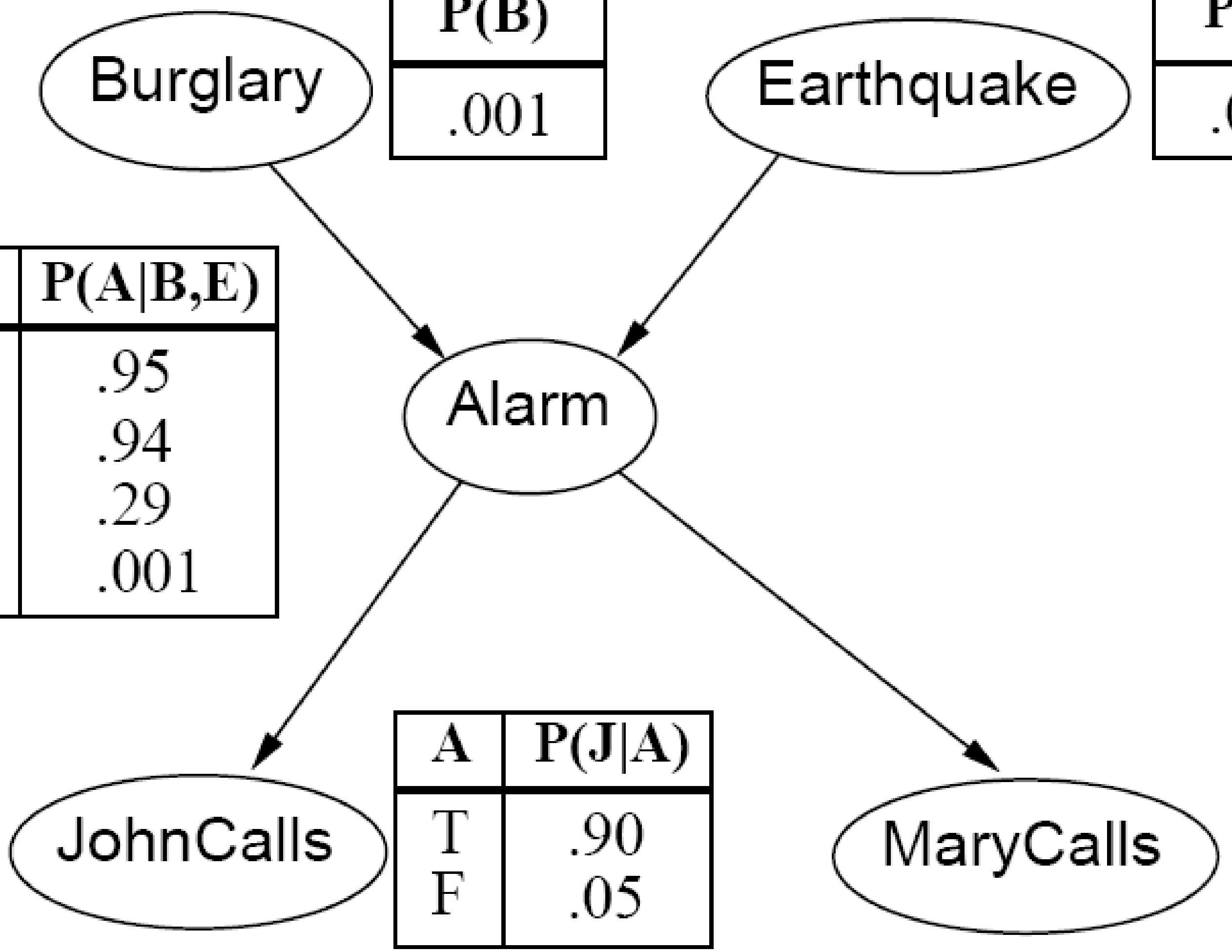
- A BN represents a full joint distribution in a compact way.

- We need to specify a *conditional probability* distribution for each node given its parents:
  P (X | Parents(X))



P (X | Z1, …, Zn)

Example: Burglar Alarm



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

The joint probability distribution

- Key property: each node is conditionally independent of its non-descendents given its parents
- Suppose the nodes X1, …, Xn are sorted in topological order
- For each node Xi, we know P(Xi | Parents(Xi))
- To get the joint distribution P(X1, …, Xn), use chain rule:

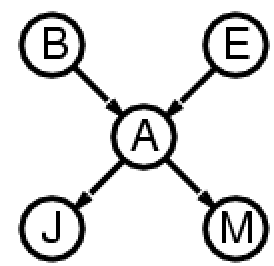
$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid X_1,\ldots,X_{i-1}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i \mid Parents(X_i)\right)$$

The joint probability distribution

$$P(X_1,\ldots,X_n)=\prod_{i=1}^{n}P\left(X_i \mid P \qquad a(X_i)\right)$$

- For example, P(j, m, a, ¬b, ¬e)
- = P(¬b) P(¬e) P(a | ¬b, ¬e) P(j | a) P(m | a)



Conditional independence

- Key property: X is conditionally independent of every *non-descendant node* given its parents
- Causal Chain

X: Low pressure

Y: Rain

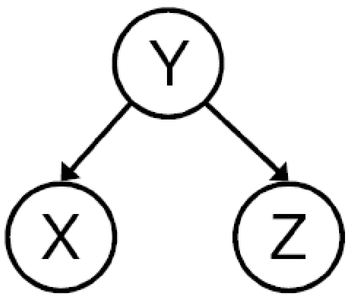
Z: Traffic

- Are X and Z independent?
- Is Z independent of X given Y?

$$P(Z \mid X, Y) = \frac{P(X, Y, Z)}{P(X, Y)} = \frac{P(X)P(Y \mid X)P(Z \mid Y)}{P(X)P(Y \mid X)} = P(Z \mid Y)$$

Conditional independence

Common cause



Y: Project due

X: Newsgroup busy

Z: Lab full

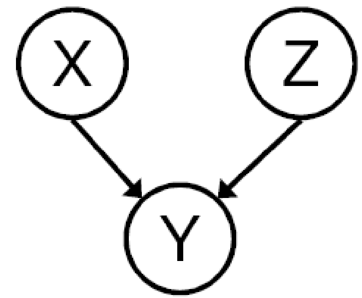
Are X and Z independent?

No

Are they conditionally independent given Y?

Yes

Common effect



X: Raining

Z: Ballgame

Y: Traffic

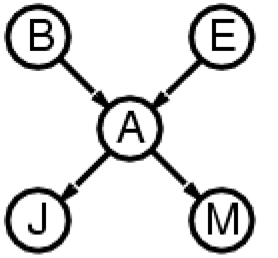
Are X and Z independent?

Yes

Are they conditionally independent given Y?

No

Compactness



- Suppose we have a Boolean variable Xi with k Boolean parents. How many rows does its conditional probability table have?
  - 2^k rows for all the combinations of parent values
  - Each row requires one number p for Xi = true
- If each variable has no more than k parents, how many numbers does the complete network require?
  - $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution
- How many nodes for the burglary network?

1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 25-1 = 31)

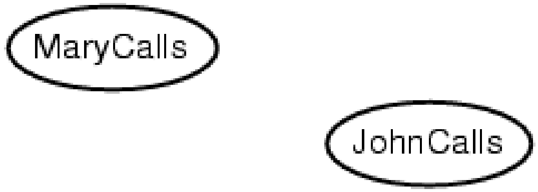
Constructing Bayesian networks

1. Choose an ordering of variables X1, … , Xn
2. For i = 1 to n
   - add Xi to the network
   - select parents from X1, … ,Xi-1 such that
     P(Xi | Parents(Xi)) = P(Xi | X1, ... Xi-1)

Example

- Suppose we choose the ordering M, J, A, B, E

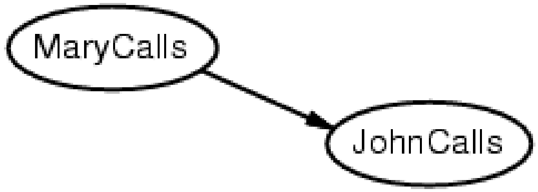
P(J | M) = P(J)?



Example

- Suppose we choose the ordering M, J, A, B, E

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



Example

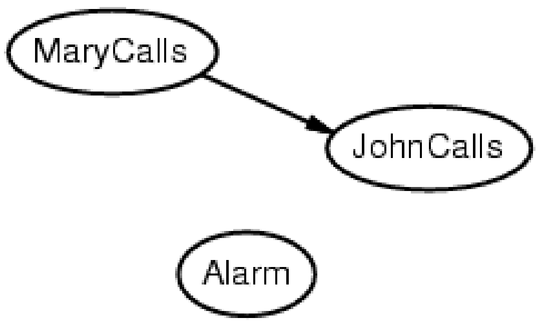
- Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)? No

P(A | J, M) = P(A)?

P(A | J, M) = P(A | J)?

P(A | J, M) = P(A | M)?



Example

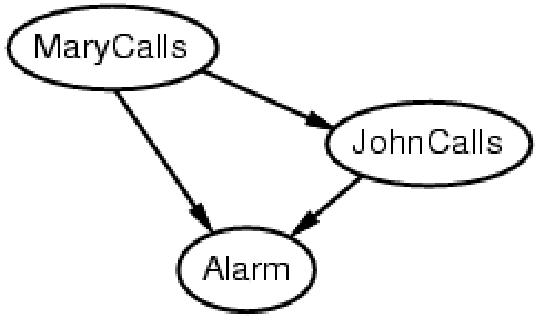
- Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)? No

P(A | J, M) = P(A)? No

P(A | J, M) = P(A | J)? No

P(A | J, M) = P(A | M)? No



Example

- Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)? No

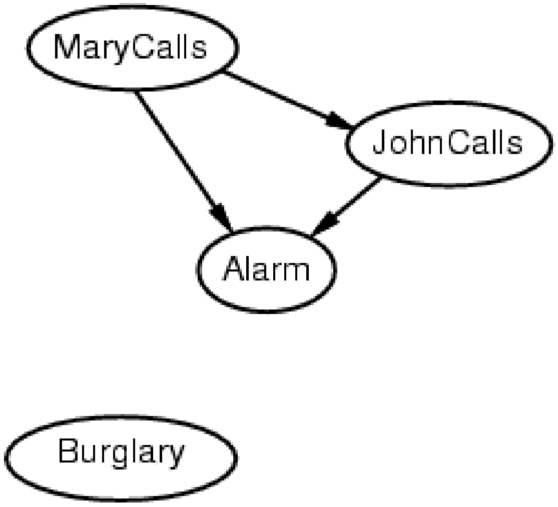
P(A | J, M) = P(A)? No

P(A | J, M) = P(A | J)? No

P(A | J, M) = P(A | M)? No

P(B | A, J, M) = P(B)?

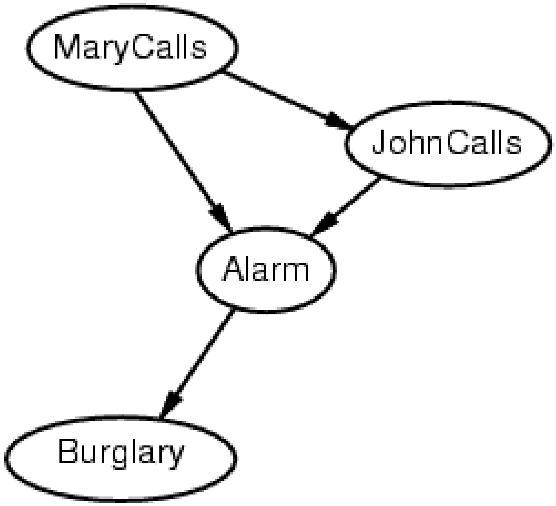
P(B | A, J, M) = P(B | A)?



Example

- Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)?	No
P(A | J, M) = P(A)?	No
P(A | J, M) = P(A | J)?	No
P(A | J, M) = P(A | M)?	No
P(B | A, J, M) = P(B)?	No
P(B | A, J, M) = P(B | A)?	Yes



Example

- Suppose we choose the ordering M, J, A, B, E

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

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

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

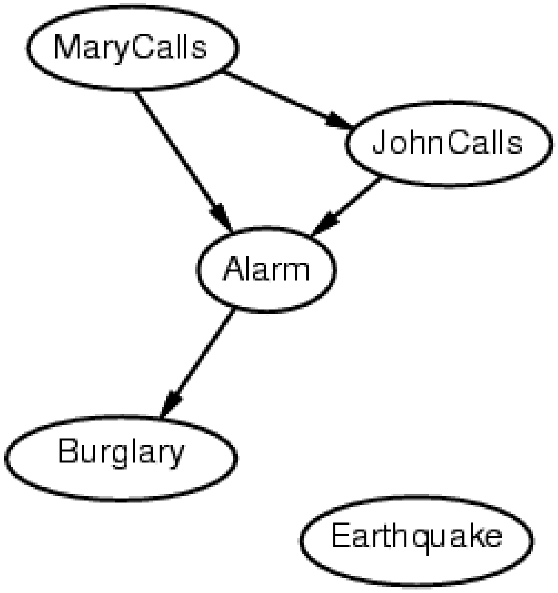
$P(A \mid J, M) = P(A \mid M)$? No

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

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

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

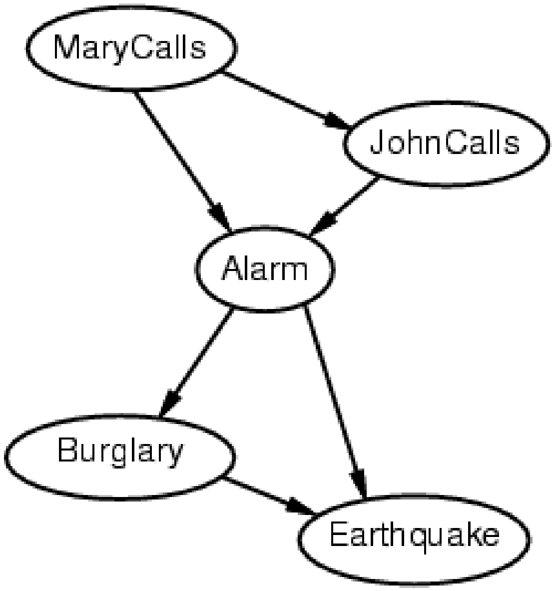
$P(E \mid B, A, J, M) = P(E \mid A, B)$?



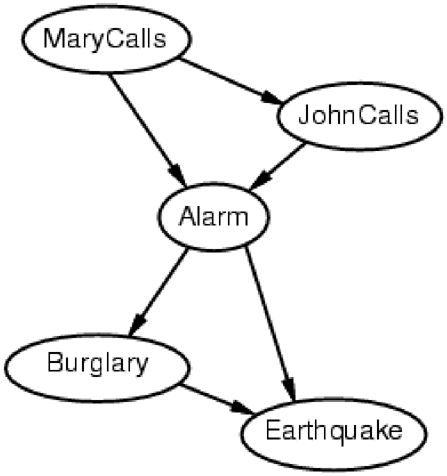
Example

- Suppose we choose the ordering M, J, A, B, E

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



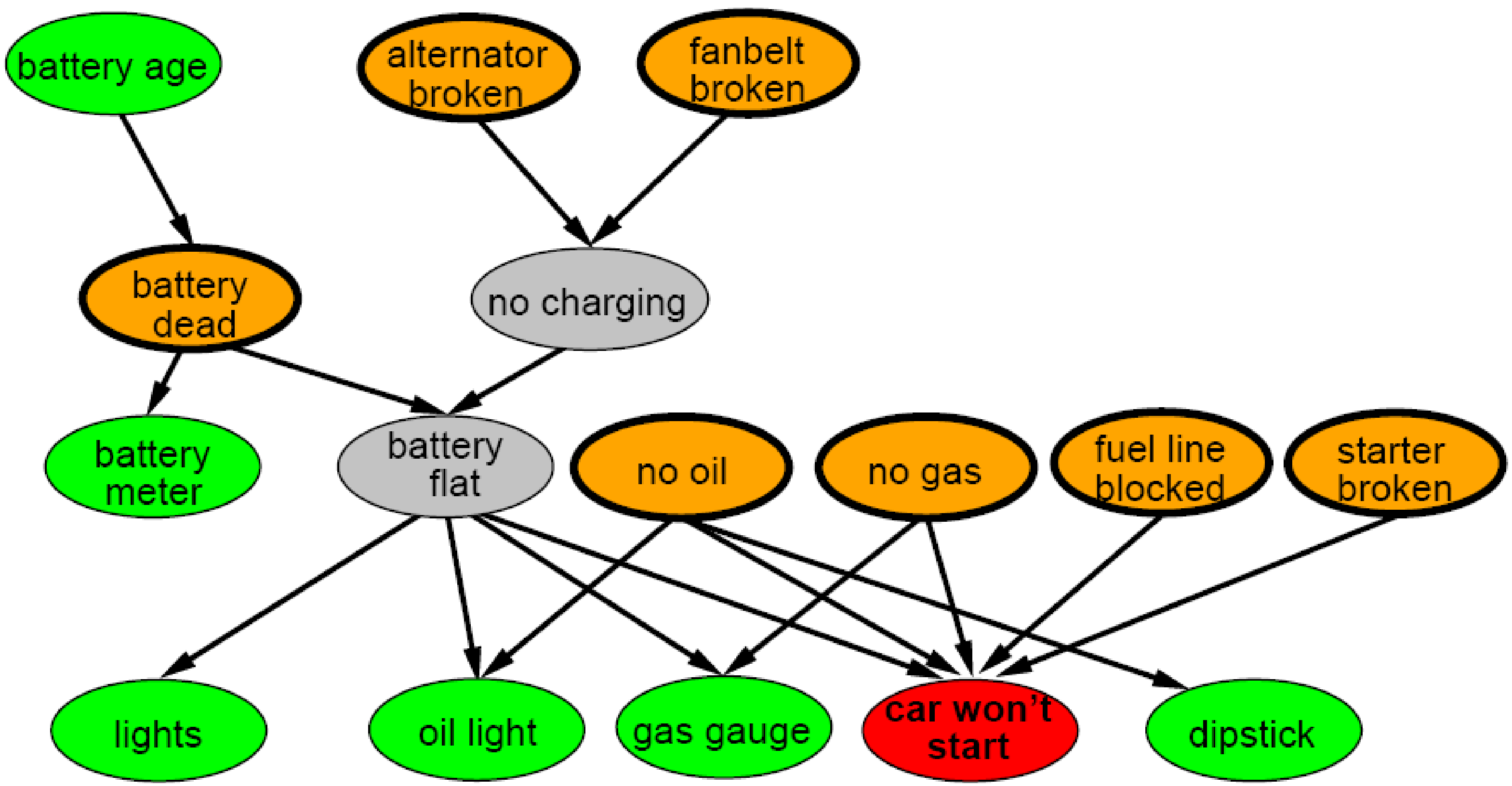
Example contd.



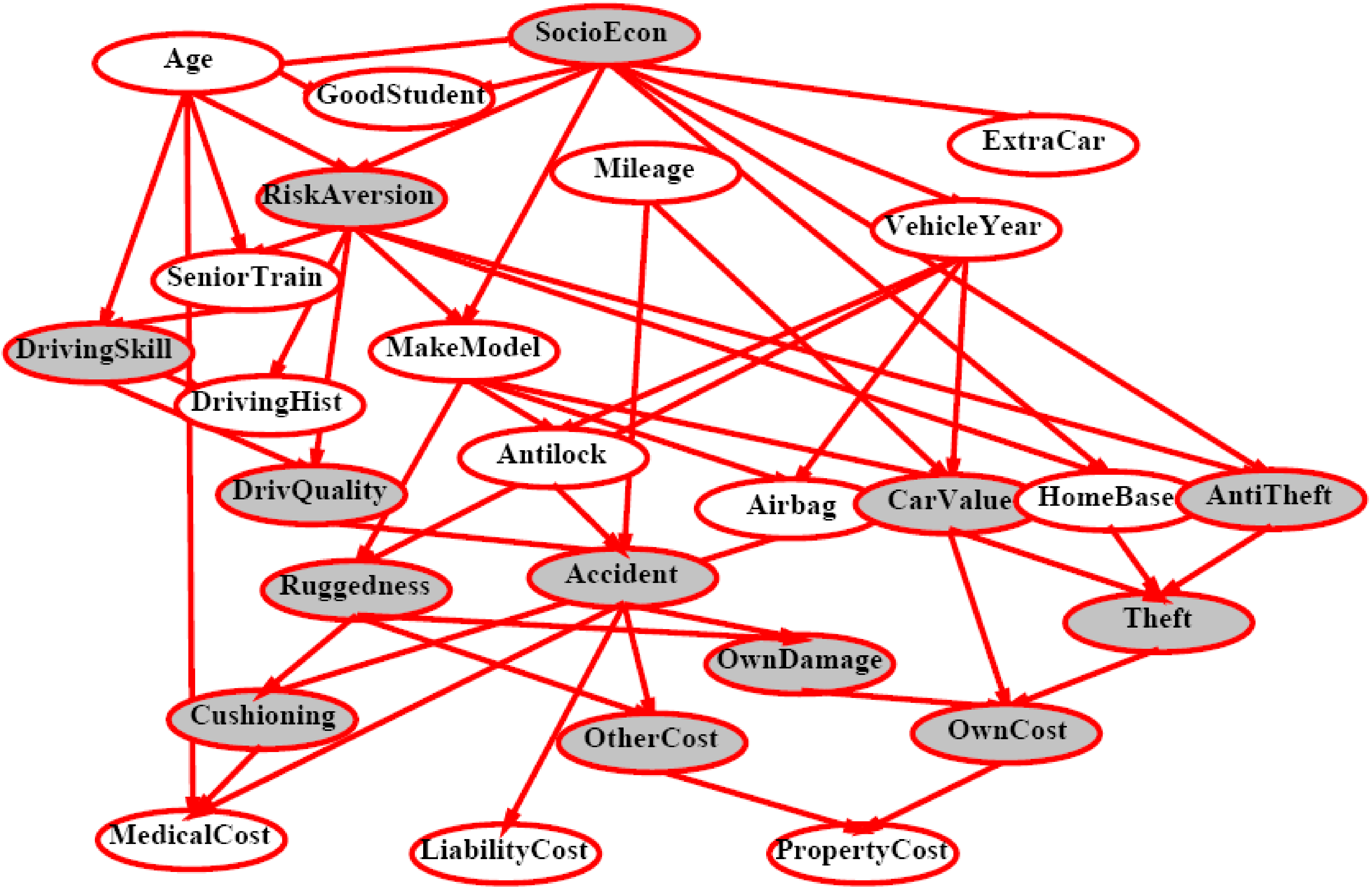
- Deciding conditional independence is hard in noncausal directions
  - The causal direction seems much more natural, but is not mandatory
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

A more realistic Bayes Network: Car diagnosis

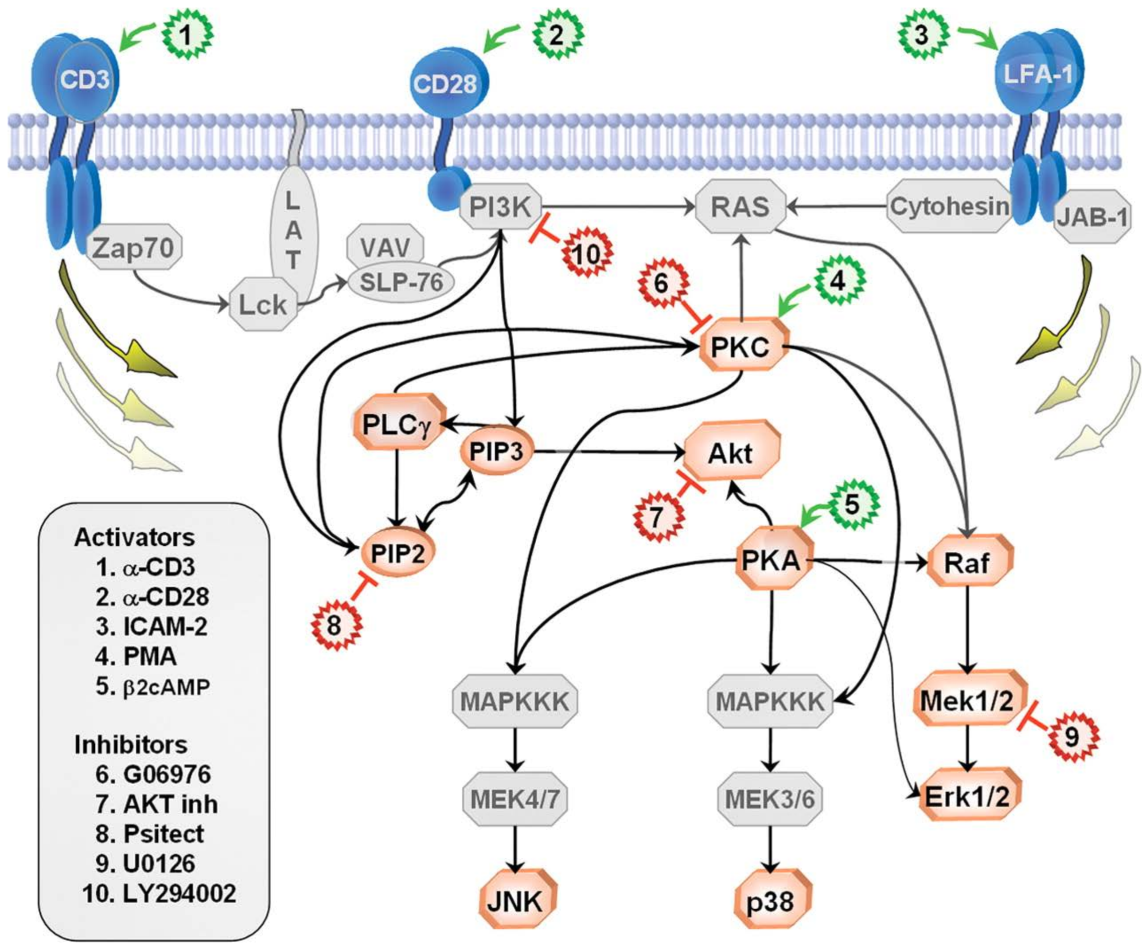
- Initial observation: car won't start
- Orange: "broken, so fix it" nodes
- Green: testable evidence
- Gray: "hidden variables" to ensure sparse structure, reduce parameteres



Car insurance



In computational biology…

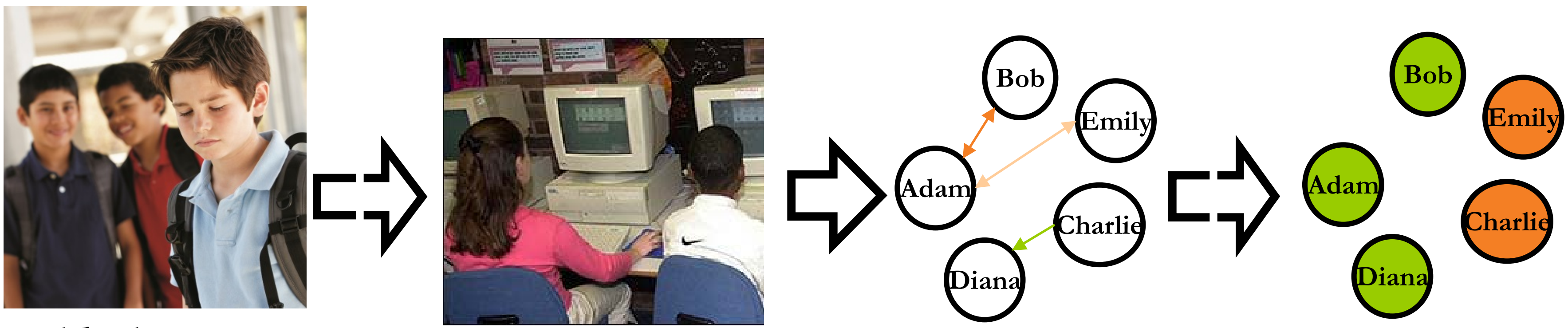


Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan

(22 April 2005) *Science* **308** (5721), 523.

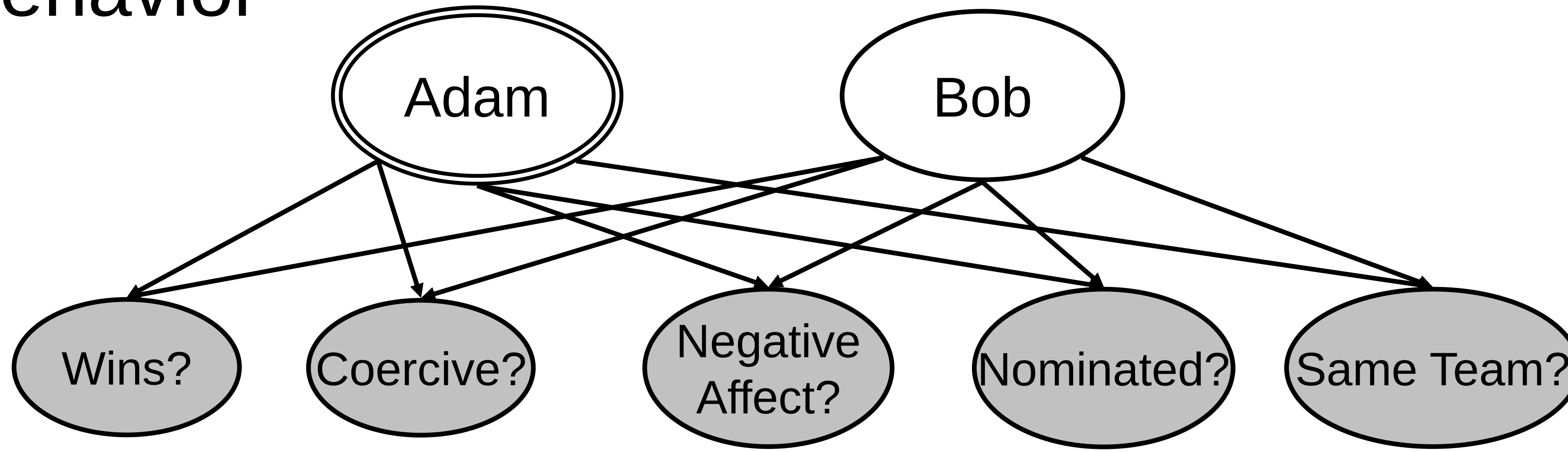
Identifying Aggressive Behavior



World Phenomena (e.g., bullying) — Game — Inference — Inferred Roles

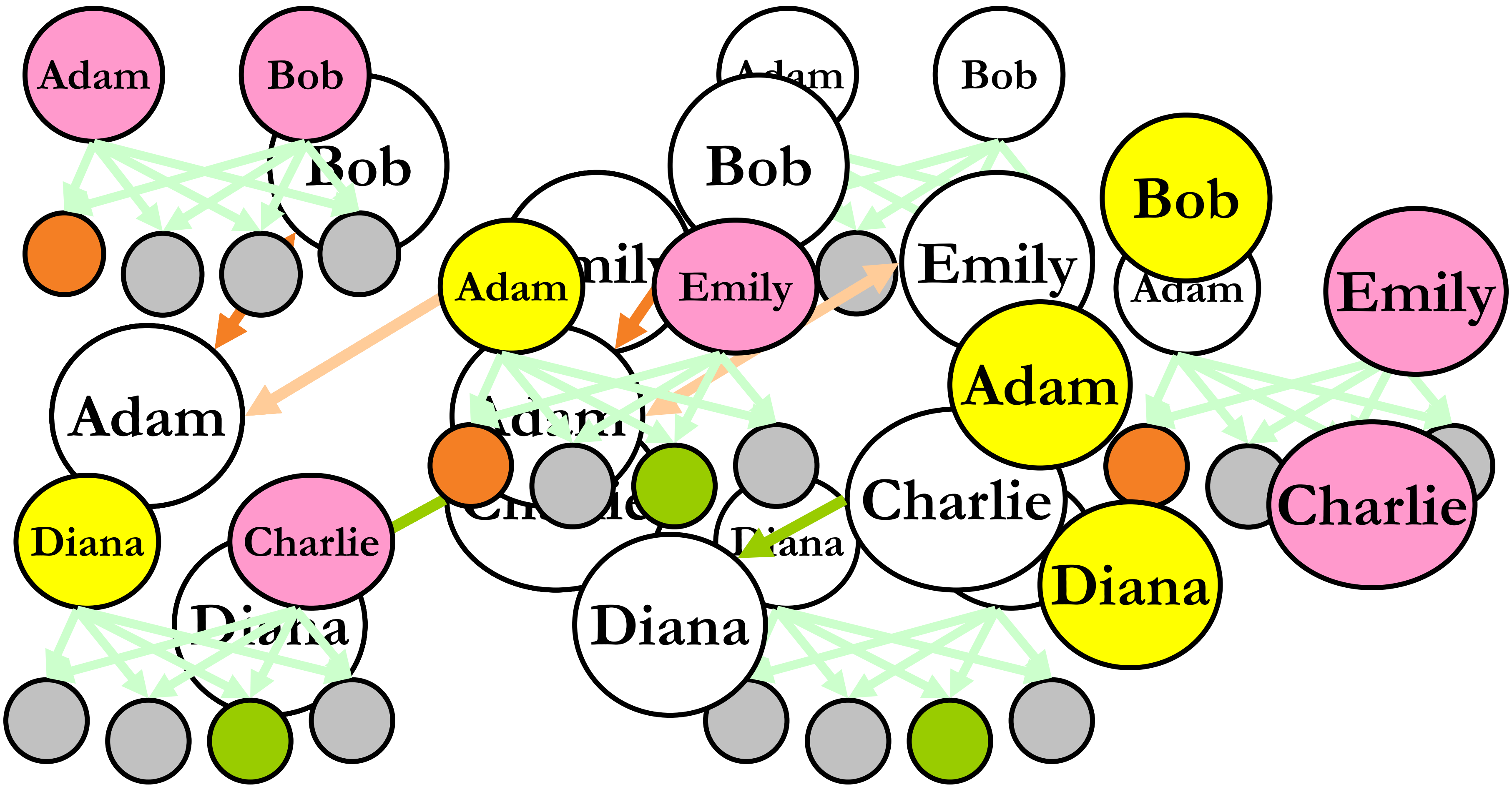
Mancilla-Caceres, J.F., Pu, W., Amir, E., and D. Espelage. *Identifying Bullies with a Computer Game.* In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12). 2012

Pairwise Model for Identification of Aggressive Behavior



- **Wins** = 1 if *Adam* got more coins from *Bob* than *Bob* from *Adam*.
- **Coercive** = 1 if *Adam* sent more coercive than prosocial messages to *Bob*.
- **Negative Affect** = 1 if *Adam* sent more negative affect messages than positive ones to *Bob*.
- **Nominated** = 1 if *Adam* nominated *Bob* to be on the same team, -1 if the nomination was negative, 0 otherwise.
- **Same Team** = 1 if *Adam* and *Bob* belong to the same team.

Global Inference from Pairwise Interactions.



In computer vision

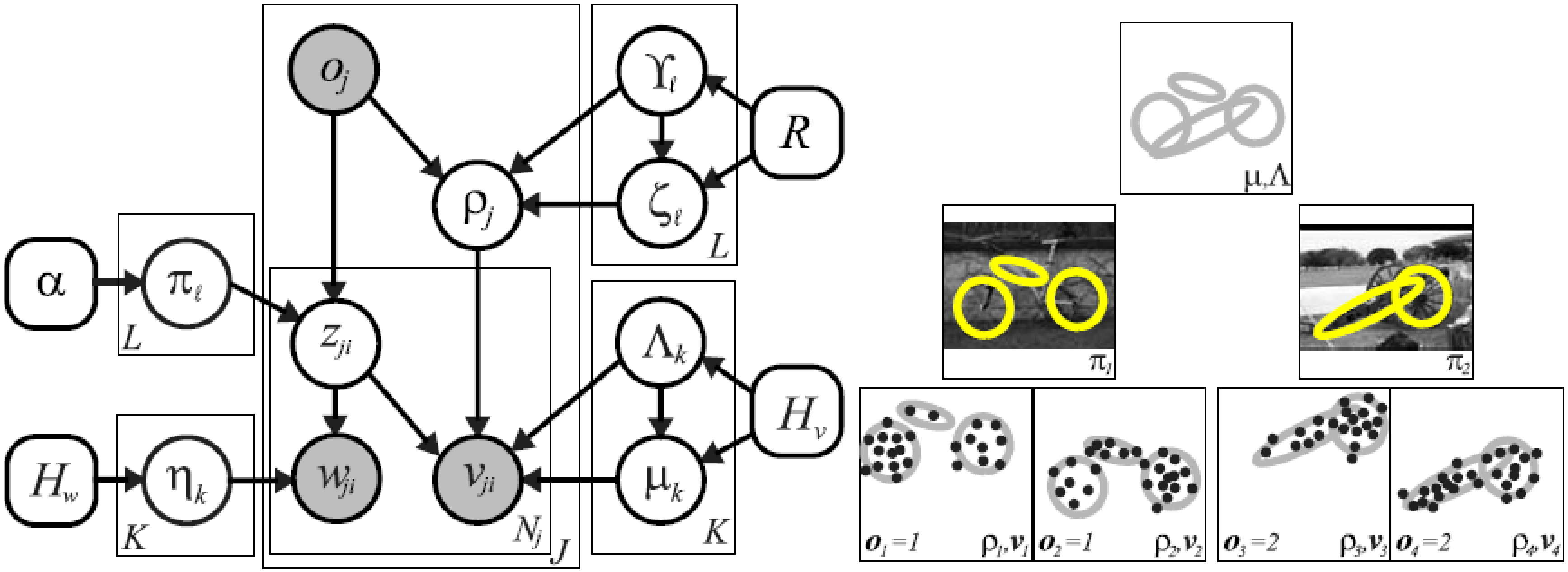


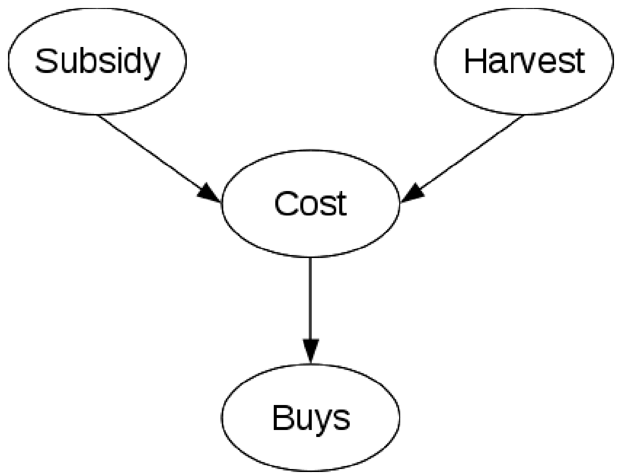
Fig. 3 A parametric, fixed-order model which describes the visual appearance of L object categories via a common set of K shared parts. The j^{th} image depicts an instance of object category o_j, whose position is determined by the reference transformation ρ_j. The appearance w_{ji} and position v_{ji}, relative to ρ_j, of visual features are determined by assignments $z_{ji} \sim \pi_{o_j}$ to latent parts. The cartoon example illustrates how a wheel part might be shared among two categories, *bicycle* and *cannon*. We show feature positions (but not appearance) for two hypothetical samples from each category

Describing Visual Scenes Using Transformed Objects and Parts

E. Sudderth, A. Torralba, W. T. Freeman, and A. Willsky.

International Journal of Computer Vision, No. 1-3, May 2008, pp. 291-330.

Continuous Variables Example



$$P(c|h, subsidy) = N(a_t h + b_t, \sigma_t^2)(c) = \frac{1}{\sigma_t\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_t h+b_t)}{\sigma_t}\right)^2}$$

$$P(c|h, \neg subsidy) = N(a_f h + b_f, \sigma_f^2)(c) = \frac{1}{\sigma_f\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{c-(a_f h+b_f)}{\sigma_f}\right)^2}$$

$$P(buys \mid Cost = c) = \frac{1}{1 + exp(-2\frac{-c+\mu}{\sigma})}.$$

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + conditional probability tables
- Generally easy for domain experts to construct

Probabilistic inference

- A general scenario:
  - Query *variables:* **X**
  - *Evidence* (*observed*) variables: **E** = **e**
  - *Unobserved* variables: **Y**
- If we know the full joint distribution P(**X**, **E**, **Y**), how can we perform inference about **X**?

$$P(\boldsymbol{X} \mid \boldsymbol{E} = \boldsymbol{e}) = \frac{P(\boldsymbol{X}, \boldsymbol{e})}{P(\boldsymbol{e})} \propto \sum_{\boldsymbol{y}} P(\boldsymbol{X}, \boldsymbol{e}, \boldsymbol{y})$$

- Problems
  - Full joint distributions are too large
  - Marginalizing out Y may involve too many summation terms