# Graphical models

Sunita Sarawagi
IIT Bombay
http://www.cse.iitb.ac.in/~sunita

# Probabilistic modeling

- Given: several variables: $x_1, \ldots x_n$ , $n$ is large.
- Task: build a joint distribution function $\Pr(x_1, \ldots x_n)$
- Goal: Efficiently represent, estimate, and answer inference queries on the distribution
- Basic premise
  - Explicit joint distribution is dauntingly large
  - Queries are simple marginals (sum or max) over the joint distribution.

# Example

- Variables are attributes are people.

| Age | Income | Experience | Degree | Location |
|---|---|---|---|---|
| 10 ranges | 7 scales | 7 scales | 3 scales | 30 places |

- An explicit joint distribution over all columns not tractable: number of combinations: $10 \times 7 \times 7 \times 3 \times 30 = 44100$.
- Queries: Estimate fraction of people with
  - Income $>$ 200K and Degree="Bachelors",
  - Income $<$ 200K, Degree="PhD" and experience $>$ 10 years.
  - Many, many more.

# Alternatives to an explicit joint distribution

- Assume all columns are independent of each other: bad assumption
- Use data to detect pairs of highly correlated column pairs and estimate their pairwise frequencies
  - Many highly correlated pairs
    income $\not\perp\!\!\!\perp$ age, income $\not\perp\!\!\!\perp$ experience, age$\not\perp\!\!\!\perp$experience
  - Ad hoc methods of combining these into a single estimate
- Go beyond pairwise correlations: conditional independencies
  - income $\not\perp\!\!\!\perp$ age, but income $\perp\!\!\!\perp$ age | experience
  - experience $\perp\!\!\!\perp$ degree, but experience $\not\perp\!\!\!\perp$ degree | income

Graphical models make explicit an efficient joint distribution from these independencies

# More examples of CIs

H I

ML opt

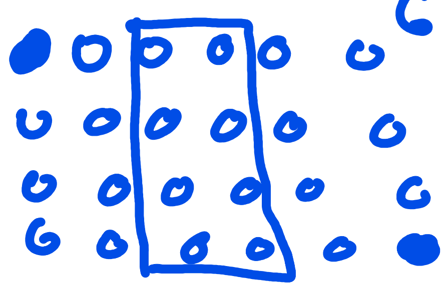
$ML \perp\!\!\!\perp opt \mid H, I$

$ML \not\perp\!\!\!\perp opt$

- The grades of a student in various courses are correlated but they become CI given attributes of the student (hard-working, intelligent, etc?)
- Health symptoms of a person may be correlated but are CI given the latent disease.

  fever $\perp\!\!\!\perp$ sore throat | flu, omicron
- Words in a document are correlated, but may become CI given the topic.
- Pixel color in an image become CI of distant pixels given near-by pixels.

6 x4



# Graphical models

Model joint distribution over **several** variables as a product of smaller factors that is

1. *Intuitive* to represent and visualize
   - Graph: represent structure of dependencies
   - Potentials over subsets: quantify the dependencies
2. *Efficient* to query
   - given values of any variable subset, reason about probability distribution of others.
   - many efficient exact and approximate inference algorithms

Graphical models = graph theory + probability theory.

# Graphical models in use

- Roots in statistical physics for modeling interacting atoms in gas and solids [ 1900]
- Early usage in genetics for modeling properties of species [ 1920]
- AI: expert systems ( 1970s-80s)
- Now many new applications:
  - Error Correcting Codes: Turbo codes, impressive success story (1990s)
  - Robotics and Vision: image denoising, robot navigation.
  - Text mining: information extraction, duplicate elimination, hypertext classification, help systems
  - Bio-informatics: Secondary structure prediction, Gene discovery
  - Data mining: probabilistic classification and clustering.

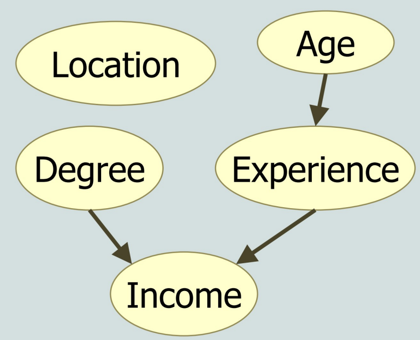
# Representation

Structure of a graphical model: Graph + Potential

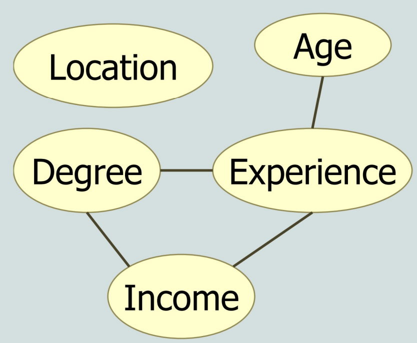
## Graph

- Nodes: variables $\mathbf{x} = x_1, \ldots x_n$
  - Continuous: Sensor temperatures, income
  - Discrete: Degree (one of Bachelors, Masters, PhD), Levels of age, Labels of words
- Edges: direct interaction
  - Directed edges: Bayesian networks
  - Undirected edges: Markov Random fields

### Directed




### Undirected



# Representation

## Potentials: $\psi_c(\mathbf{x}_c)$

- Scores for assignment of values to subsets $c$ of directly interacting variables.
- Which subsets? What do the potentials mean?
  - Different for directed and undirected graphs

## Probability

Factorizes as product of potentials

$$\Pr(\mathbf{x} = x_1, \ldots x_n) \propto \prod \psi_S(\mathbf{x}_S)$$

# Directed graphical models: Bayesian networks

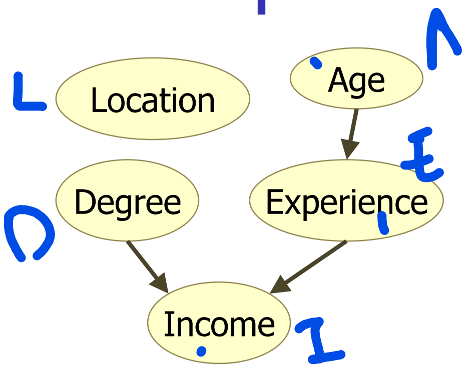
- Graph $G$: directed acyclic
  - Parents of a node: $\text{Pa}(x_i)$ = set of nodes in $G$ pointing to $x_i$
- Potentials: defined at each node in terms of its parents.

$$\psi_i(x_i, \text{Pa}(x_i)) = \Pr(x_i|\text{Pa}(x_i))$$

- Probability distribution

$$\Pr(x_1 \ldots x_n) = \prod_{i=1}^{n} \Pr(x_i|pa(x_i))$$

# Example of a directed graph




## $\psi_1(L) = \Pr(L)$

| NY | CA | London | Other |
|---|---|---|---|
| 0.2 | 0.3 | 0.1 | 0.4 |

## $\psi_2(A) = \Pr(A)$

| 20–30 | 30–45 | $>$ 45 |
|---|---|---|
| 0.3 | 0.4 | 0.3 |

or, a Guassian distribution $(\mu, \sigma) = (35, 10)$

## $\psi_2(E, A) = \Pr(E|A)$

Experience

| | 0–10 | 10–15 | $>$ 15 |
|---|---|---|---|
| **20–30** | 0.9 | 0.1 | 0 |
| **30–45** | 0.4 | 0.5 | 0.1 |
| $>$ **45** | 0.1 | 0.1 | 0.8 |

Pr(E| Age=20-30)

## $\psi_2(I, E, D) = \Pr(I|D, \cancel{A}\,E)$

3 dimensional table, or a histogram approximation.

## Probability distribution

$$\text{Pa}(\mathbf{x} = L, D, I, A, E) = \Pr(L)\Pr(D)\Pr(A)\Pr(E|A)\Pr(I|D, E)$$

# Conditional Independencies

- Given three sets of variables $X$, $Y$, $Z$, set $X$ is conditionally independent of $Y$ given $Z$ ($X \perp\!\!\!\perp Y | Z$) iff

$$\boxed{\Pr(X|Y,Z) = \Pr(X|Z)}$$

(G)

- Local conditional independencies in BN: for each $x_i$ G ← DAG

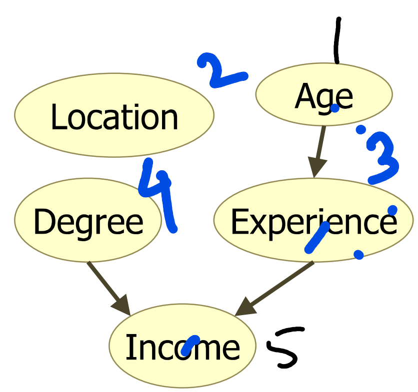
Local-CI

$$x_i \perp\!\!\!\perp ND(x_i) | Pa(x_i)$$

ND($x_i$)

ND(E) = A L, D

- $L \perp\!\!\!\perp E, D, A, I$
- $A \perp\!\!\!\perp L, D$
- $E \perp\!\!\!\perp L, D | A$
- $I \perp\!\!\!\perp L, A | E, D$



# CIs and Fractorization

## Theorem

*Given a distribution $P(x_1, \ldots, x_n)$ and a DAG $G$, if $P$ satisfies Local-CI induced by $G$, then $P$ can be factorized as per the graph.*
*Local-CI($P, G$) $\implies$ Factorize($P, G$)*

## Proof.

- $x_1, x_2, \ldots, x_n$ topographically ordered (parents before children) in $G$.
- Local CI($P, G$): $P(x_i|x_1, \ldots, x_{i-1}) = P(x_i|Pa_G(x_i))$
- Chain rule:
  $P(x_1, \ldots, x_n) = \prod_i P(x_i|x_1, \ldots, x_{i-1}) = \prod_i P(x_i|Pa_G(x_i))$
- $\implies$ Factorize($P, G$)

$\square$

Also as Theorem 3.1 in KF book

# CIs and Fractorization

## Theorem

*Given a distribution* $P(x_1, \ldots, x_n)$ *and a DAG* $G$*, if* $P$ *can be factorized as per* $G$ *then* $P$ *satisfies Local-CI induced by* $G$*.*
*Factorize*$(P, G) \implies$ *Local-CI*$(P, G)$

Proof skipped. (Refer Theorem 3.2 in KF book.)

# Drawing a BN starting from a distribution

Given a distribution $P(x_1, \ldots, x_n)$ to which we can ask any CI of the form "Is $X \perp\!\!\!\perp Y | Z$?" and get a yes, no answer.
Goal: Draw a minimal, correct BN $G$ to represent $P$.

- A DAG G is correct if all Local-CIs that are implied in G hold in P.
- A DAG G is minimal if we cannot remove any edge(s) from G and still get a correct BN for P.

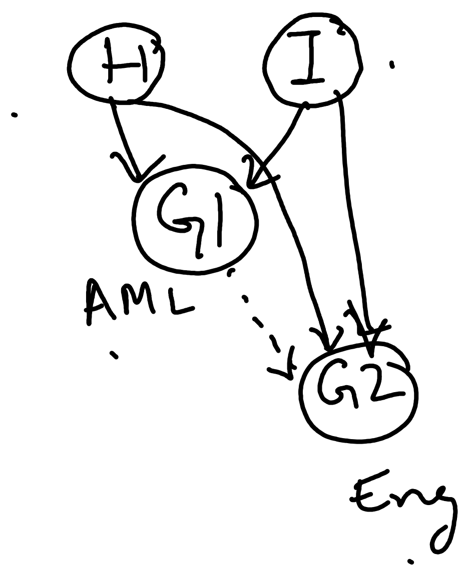
# Algorithm for drawing a BN from CIs

$x_1, \ldots, x_n$ = Choose an ordering of variables
For $i = 1 \ldots n$

- $S$=smallest subset of $Q_i = \{x_1, \ldots x_{i-1}\}$ such that $x_i \perp\!\!\!\perp Q_i - S | S$
- Make each variable in $S$ a parent of $x_i$

# Examples

H, I, G1, G2




$I \perp\!\!\!\perp H \mid$ ✓

$G1 \perp\!\!\!\perp \{H, I\}$ ✗

$G1 \perp\!\!\!\perp H \mid I$ ✗

$G1 \perp\!\!\!\perp I \mid H$ ✗

$G2 \perp\!\!\!\perp \{H, I, G1\}$ ✗

$\boxed{G2 \perp\!\!\!\perp G1 \mid H, I}$ ?

# Examples

Diseases & symptoms.

# Why minimal

## Theorem

*$G$ constructed by the above algorithm is minimal, that is, we cannot remove any edge from the BN while maintaining the correctness of the BN for $P$*

## Proof.

By construction. A subset of ND of each $x_i$ were available when parent of $U$ were chosen minimally. $\square$

# Why Correct

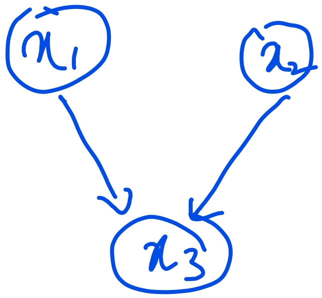
## Theorem

*$G$ constructed by the above algorithm is correct, that is, the local-CIs induced by $G$ hold in $P$*

## Proof.

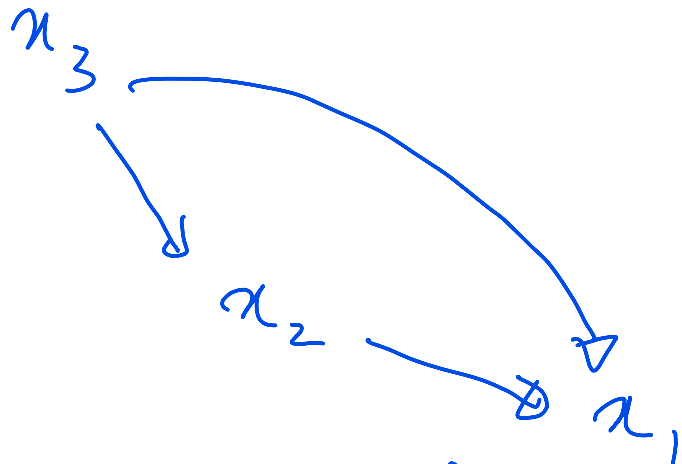
The construction process makes sure that the factorization property holds. Since factorization implies local-CIs, the constructed BN satisfied the local-CIs of $P$ □

# Order is important




Hidden

order $x_3, x_2, x_1$ for creating BN




correct and minimal but not optimal.

# Examples of CIs that hold in BN but not covered by local-CI