Bayesian Network Representation

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Exploiting Independence Properties
- Knowledge Engineering

Parameters for Independent r.v.s

- Each X_i represents outcome of toss of coin i
  - Assume coin tosses are marginally independent
  - i.e., $(X_i \perp X_j)$, therefore

$$P(X_1,..,X_n)=P(X_1)P(X_2)..P(X_n)$$

- If we use standard parameterization of the joint distribution, the independence structure is obscured and required 2^n parameters
- However we can use a more natural set of parameters: n parameters $\theta_1,..\theta_n$

Conditional Parameterization

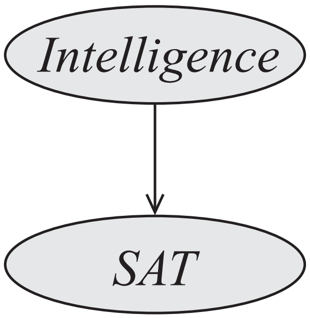
- Ex: Company is trying to hire recent graduates
- Goal is to hire intelligent employees
  - No way to test intelligence directly
  - But have access to Student's SAT score
    - Which is informative but not fully indicative
- Two random variables
  - Intelligence: $Val(I)=\{i^1,i^0\}$, high and low
  - Score: $Val(S)=\{s^1,s^0\}$, high and low
- Joint distribution has 4 entries

  Need three parameters

I	S	$P(I,S)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24.

Alternative Representation: Conditional Parameterization

- $P(I, S) = P(I)P(S \mid \mathrm{I})$
  - Representation more compatible with causality
    - Intelligence influenced by Genetics, upbringing
    - Score influenced by Intelligence
- Note: BNs are not required to follow causality but they often do



- Need to specify $P(I)$ and $P(S/I)$

i^0	i^1
0.7	0.3

I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

  - Three binomial distributions (3 parameters) needed
    - One marginal, two conditionals $P(S/I=i^0)$, $P(S/i=i^1)$

Conditional Parameterization and Conditional Independences

- Conditional Parameterization is combined with Conditional Independence assumptions to produce very compact representations of high dimensional probability distributions

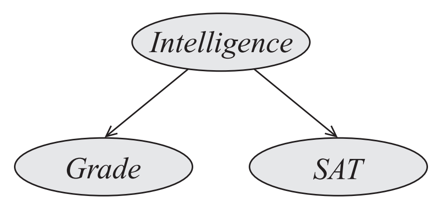
Naïve Bayes Model

- Conditional Parameterization combined with Conditional Independence assumptions
  - $Val(G)=\{g^1, g^2, g^3\}$ represents grades A, B, C

$P(G|I)$

I	g^1	g^2	g^3
i^0	0.2	0.34	0.46
i^1	0.74	0.17	0.09

- SAT and Grade are independent given Intelligence (assumption)
  - Knowing intelligence, SAT gives no information about class grade $P \models (S \perp G \mid I)$

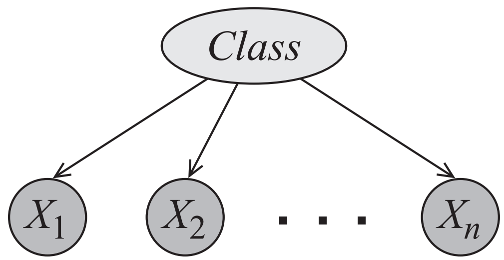


- Assertions
  - From probabilistic reasoning $P(I,S,G) = P(S,G \mid I)P(I)$
  - From assumption $P(S,G \mid I) = P(S \mid I)P(G \mid I)$
  - Combining $P(I,S,G) = P(S \mid I)P(G \mid I)P(I)$

$$\begin{aligned} P(i^1, s^1, g^2) &= P(i^1)P(s^1 \mid i^1)P(g^2 \mid i^1) \\ &= 0.3 \cdot 0.8 \cdot 0.17 = 0.0408. \end{aligned}$$

Three binomials, two 3-value multinomials: 7 params
More compact than joint distribution

BN for General Naiive Bayes Model



$$P(C, X_1, .. X_n) = P(C)\prod_{i=1}^{n} P(X_i \mid C)$$

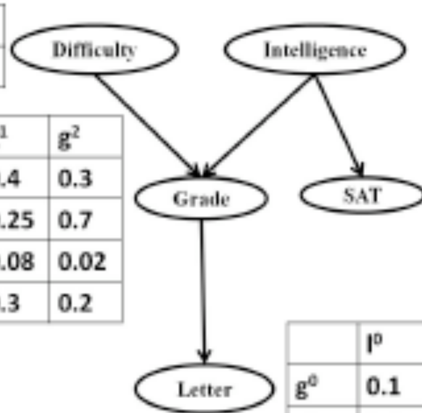
Encoded using a very small number of parameters

Linear in the number of variables

Application of Naiive Bayes Model

- Medical Diagnosis
  - Pathfinder expert system for lymph node disease (Heckerman et.al., 1992)
- Full BN agreed with human expert 50/53 cases
- Naiive Bayes agreed 47/53 cases

“Student” Bayesian Network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^0	g^1	g^2
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

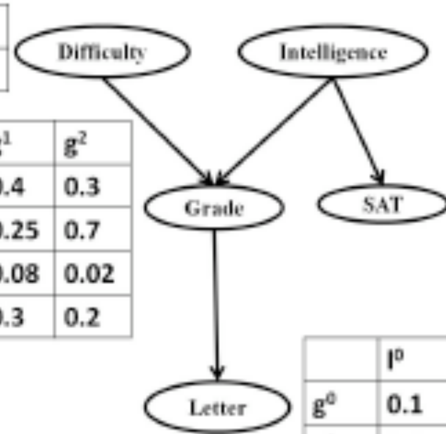
	l^0	l^1
g^0	0.1	0.9
g^1	0.4	0.6
g^2	0.99	0.01

- Represents joint probability distribution over multiple variables
  - BNs represent them in terms of graphs and conditional probability distributions(CPDs)
    - Resulting in great savings in no of parameters needed

Joint distribution from Student BN

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3



	g^0	g^1	g^2
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^0	0.1	0.9
g^1	0.4	0.6
g^2	0.99	0.01

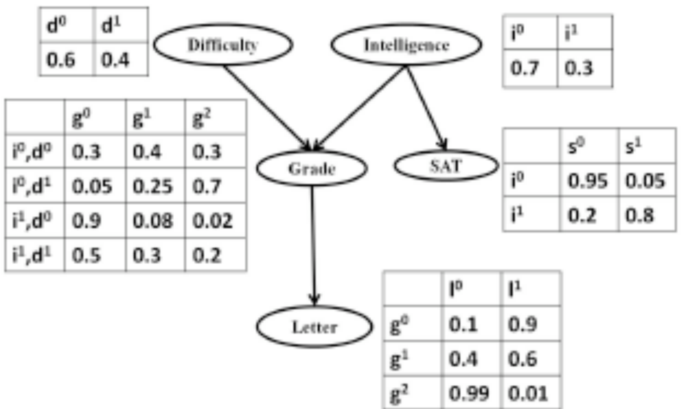
- CPDs: $P(X_i \mid pa(X_i))$
- Joint Distribution:

$$P(X) = P(X_1, .. X_n)$$

$$P(X) = \prod_{i=1}^{N} P(X_i \mid pa(X_i))$$

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

Example of Probability Query



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^0	g^1	g^2
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^0	0.1	0.9
g^1	0.4	0.6
g^2	0.99	0.01

$$P(Y = y_i \mid E = e) = \frac{P(Y = y_i, E = e)}{P(E = e)}$$

Posterior Marginal Probability of Evidence

- Posterior Marginal Estimation: $P(I=i^1|L=l^0,S=s^1)=?$
- Probability of Evidence: $P(L=l^0,s=s^1)=?$
  - Here we are asking for a specific probability rather than a full distribution

Computing the Probability of Evidence

Probability Distribution of Evidence

$$P(L,S) = \sum_{D,I,G} P(D,I,G,L,S)$$ Sum Rule of Probability

$$= \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L \mid G)P(S \mid I)$$ From the Graphical Model

Probability of Evidence

$$P(L=l^0, s=s^1) = \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L=l^0 \mid G)P(S=s^1 \mid I)$$

More Generally

$$P(E=e) = \sum_{X \setminus E} \prod_{i=1}^{n} P(X_i \mid pa(X_i))\mid_{E=e}$$

- An intractable problem
  - #P complete
- Tractable when tree-width is less than 25
  - Most real-world applications have higher tree-width
- Approximations are usually sufficient (hence sampling)
  - When $P(Y=y|E=e)=0.29292$, approximation yields 0.3

Genetic Inheritance and Bayesian Networks

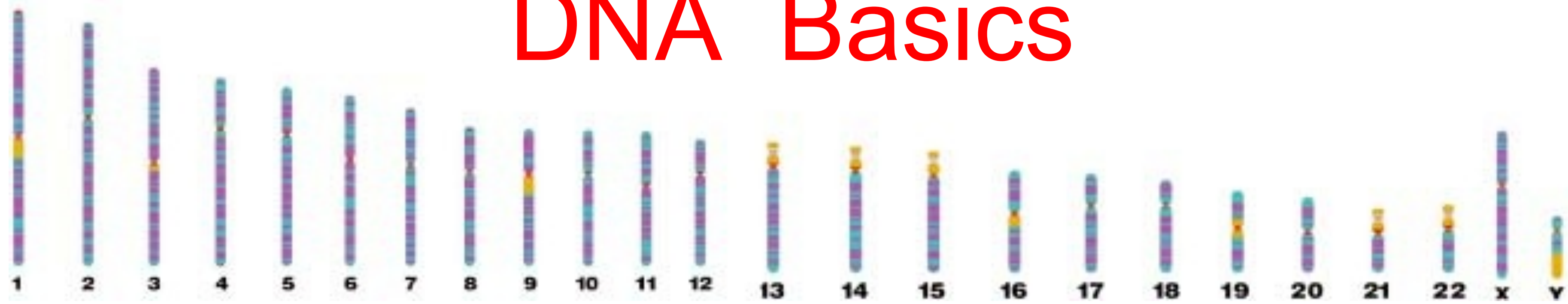
Genetics Pedigree Example

- One of the earliest uses of Bayesian Networks
  - Before general framework was defined
- Local independencies are intuitive
- Model transmission of certain properties such as blood type from parent to child

Phenotype and Genotype

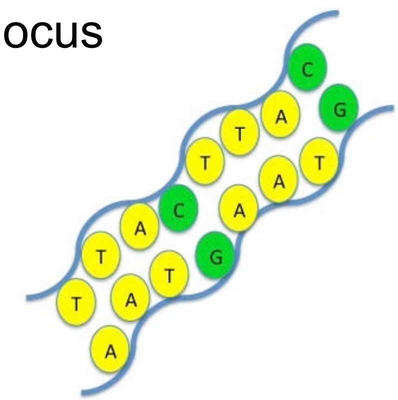
- Some background on genetics needed to model properly
- Blood type is an observable quantity that depends on the genetic makeup
  - Called a phenotype
- Genetic makeup of a person is called a genotype

DNA Basics



Genome has 46 chromosomes (22 are repeated plus XX and XY)

Actual Electron Photo micrograph



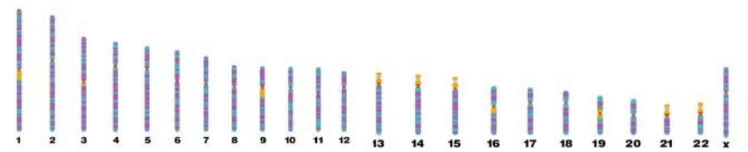
Single Chromosome: ~10^8 base-pairs
Genome: sequence of $3\text{x}10^9$ base-pairs
(nucleotides A,C,G,T)
Represents full set of chromosomes

Large portions of DNA have no survival function (98.5%) and have variations useful for identification

TH01 is a location on short arm of chromosome 11:
short tandem repeats (STR) of same base pair AATG
Variant forms (alleles) different for different individuals

Genetic Model

- Human genetic material



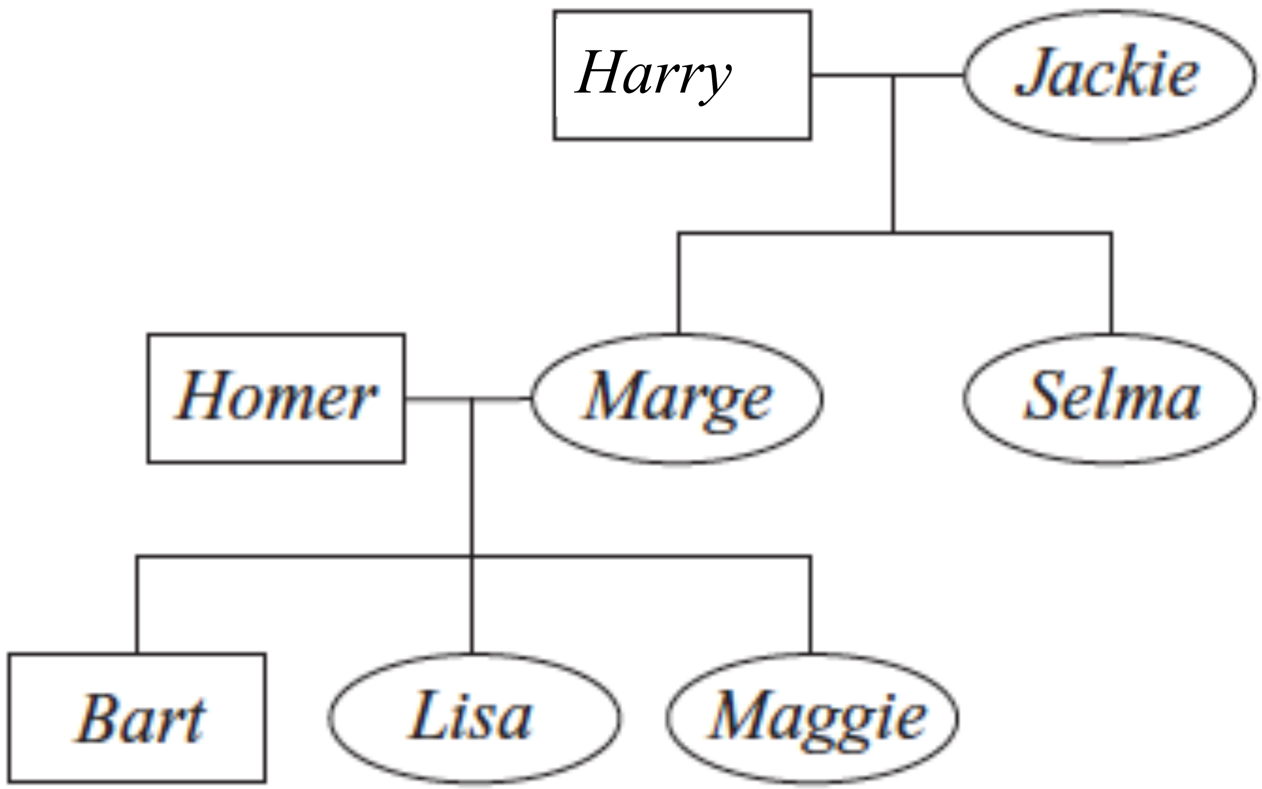
  - 22 pairs of *autosomal* chromosomes
  - One pair of *sex* chromosomes (X and Y)
- Each chromosome contains genetic material that determine person's properties

- Locus: Region of chromosome of interest
  - Blood type is a particular locus
- Alleles: Variants of locus
  - Blood type has three variants: A, B, O

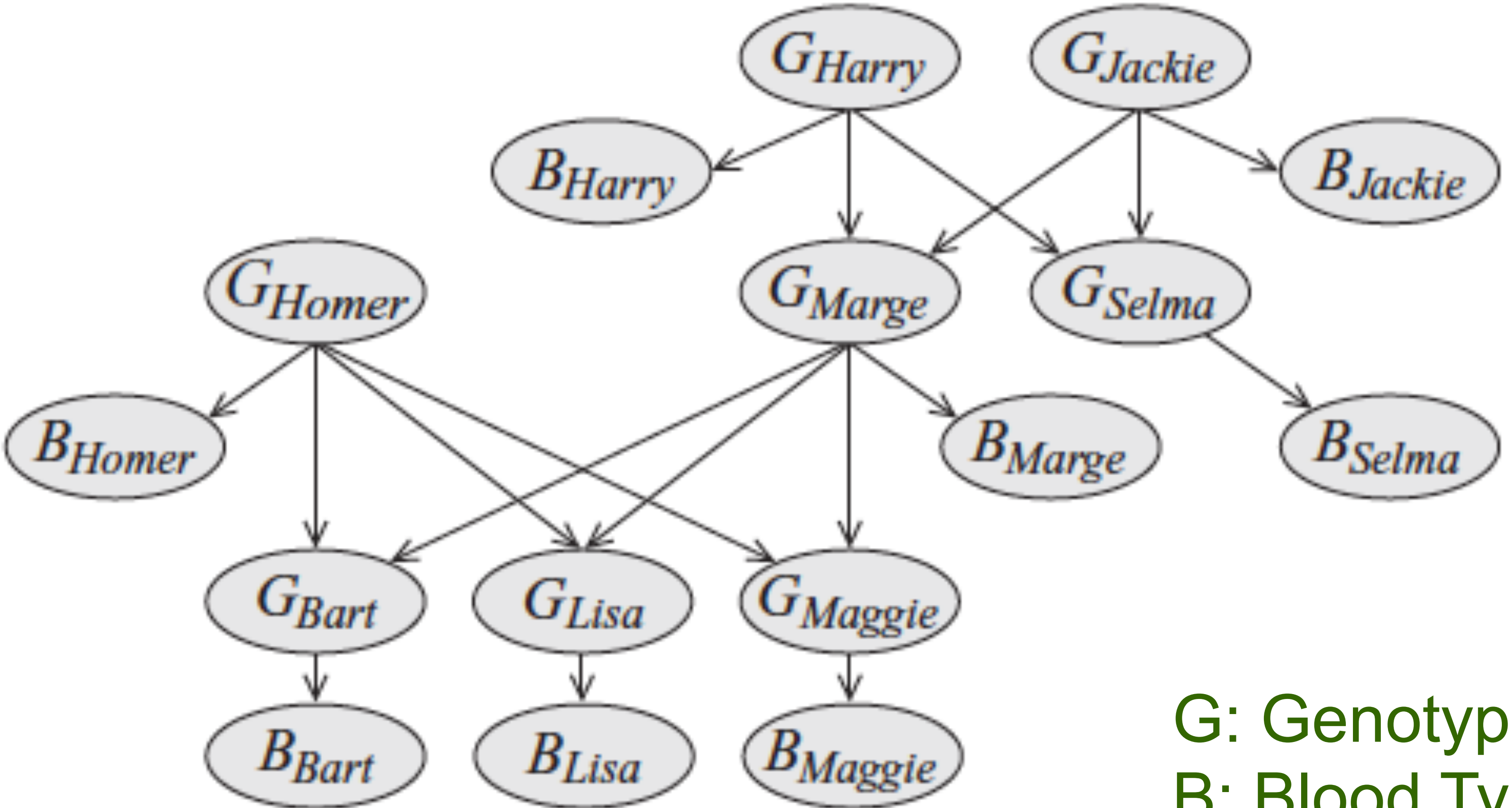
Independence Assumptions

- Arise from biology
- Once we know
  - Genotype of a person
    - additional evidence about other members of family will not provide new information about blood-type
  - Genotype of both parents
    - Determine what is passed to off-spring
    - Additional ancestral information not needed
- These independencies can be captured in BN for a family tree

A small family tree



BN for Genetic Inheritance



G: Genotype
B: Blood Type

Autosomal Chromosome

- In each pair,
  - Paternal: inherited from father
  - Maternal: inherited from mother
- Person's genotype is an ordered pair (X,Y)
  - with each having three possible values (A,B,O)
  - there are nine values such as (A,B)
- Blood type phenotype is a function of both copies
  - E.g., genotype (A,O) blood type is A
  - (O,O) → O

CPDs for Genetic Inheritance

- Penetrance Model $P(B(c)|G(c))$
  - Probabilities of different phenotypes given person's genotype
    - Deterministic for bloodtype
- Transmission Model $P(G(c)|G(p),G(m))$
  - Each parent equally likely to transmit either of two alleles to child
- Genotype Priors $P(G(c))$
  - Genotype frequencies in population

Real models more complex

- Phenotypes for late-onset diseases are not a deterministic function of genotype
  - A particular genotype may have a higher probability of a disease
- Genetic makeup of individual determined by many genes
- Some phenotypes depend on many genes
- Multiple phenotypes depend on many genes

Modeling multi-locus inheritance

- Inheritance patterns of different genes not independent of each other
- Need to take into account adjacent loci
- Introduce selector variables $S(l,c,m)$
    - 1 if locus l in c's maternal chromosome inherited from c's *maternal grandmother*
    - 2 if locus inherited from c's *maternal grandfather*
- Model correlations of variables of adjacent loci l and l'

Use of Genetic Inheritance Model

- Extensively used in

1. In genetic counseling and prediction
2. In linkage analysis

Genetic Counseling and Prediction

- Take phenotype with known loci and observed phenotype and genotype data for individuals
  - to infer genotype and phenotype for another person (planned child)
- Genetic data
  - Direct measurements of relevant disease loci or nearby loci which are correlated with disease loci

Linkage Analysis

- Harder task
- Identifying disease genes from pedigree data using several pedigrees
  - Several individuals exhibit disease phenotype
  - Available data
    - Phenotype information for many individuals in pedigree
    - Genotype information for known location in chromosome
  - Use inheritance model to evaluate likelihood
  - Pinpoint area linked to disease to further analyze genes in that area
    - Allows focusing on 1/10,000 of genome

Sparse BN in genetic inheritance

- Allow reasoning about large pedigree and multiple loci
- Allow use of *model learning* algorithms to understand recombination rates in different regions and penetration probabilities for different diseases

Graphs and Distributions

- Relating two concepts:
  - Independencies in distributions
  - Independencies in graphs
- I-Map is a relationship between the two

Independencies in a Distribution

- Let P be a distribution over X
- $I(P)$ is set of conditional independence assertions of the form $(X \perp Y|Z)$ that hold in P

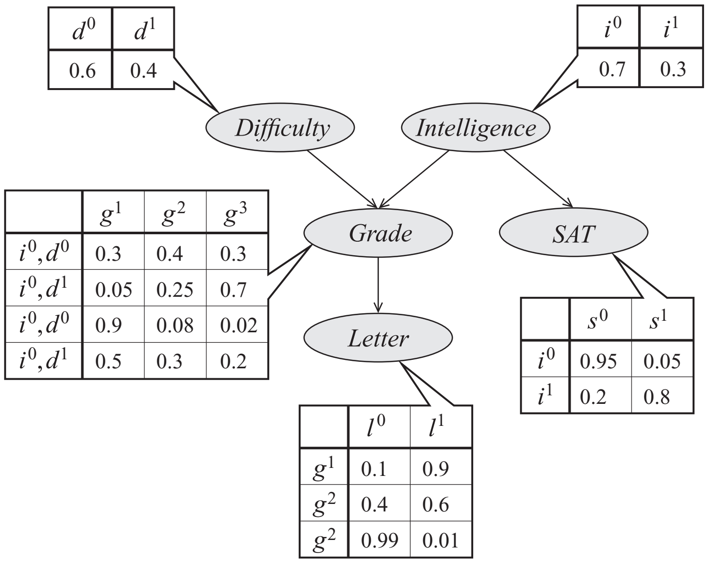
X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and Y are independent in P, e.g.,

$P(x^1)=0.48+0.12=0.6$
$P(y^1)=0.32+0.48=0.8$
$P(x^1,y^1)=0.48=0.6 \times 0.8$

Thus $(X \perp Y|\phi) \in I(P)$

Independencies in a Graph



- Graph $\mathcal{G}$ with CPDs is equivalent to a set of independence assertions

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

- Local Conditional Independence Assertions (starting from leaf nodes):

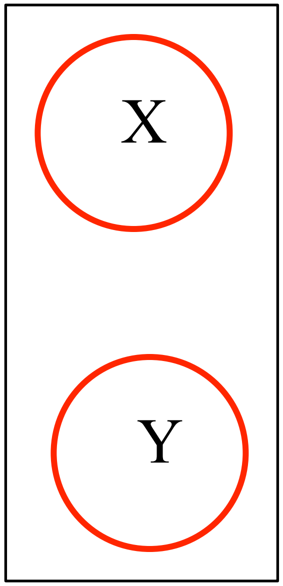
$I(G) = \{(L \perp I,D,S \mid G),$	L is conditionally independent of all other nodes given parent G
$(S \perp D,G,L \mid I),$	S is conditionally independent of all other nodes given parent I
$(G \perp S \mid D,I),$	Even given parents, G is NOT independent of descendant L
$(\mathrm{I} \perp D \mid \phi),$	Nodes with no parents are marginally independent
$(D \perp I,S \mid \phi)\}$	D is independent of non-descendants I and S

- Parents of a variable shield it from probabilistic influence
    - Once value of parents known, no influence of ancestors
- Information about descendants can change beliefs about a node

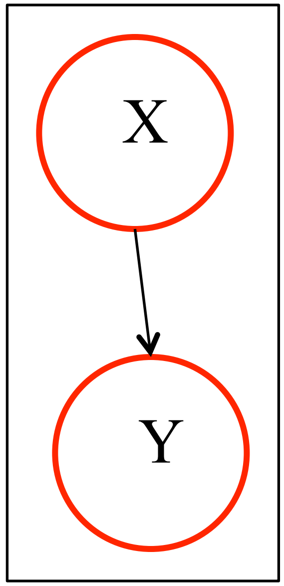
I-MAP

- Let G be a graph associated with a set of independencies $I(G)$
- Let P be a probability distribution with a set of independencies $I(P)$
- Then G is an I-map of I if $I(G) \subseteq I(P)$
- From direction of inclusion
  - distribution can have more independencies than the graph
  - Graph does not mislead in independencies existing in P

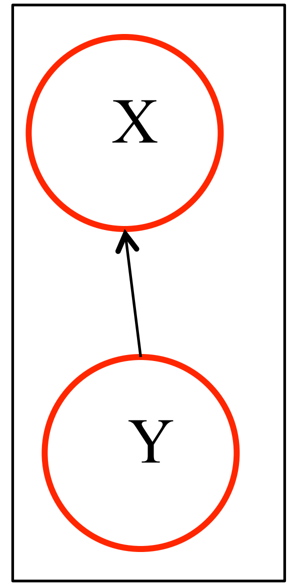
Example of I-MAP



G_0 encodes $X \perp Y$ or $I(G_0)=\{X \perp Y\}$



G_1 encodes no Independence or $I(G_1)=\{\Phi\}$



G_2 encodes no Independence $I(G_2)=\{\Phi\}$

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

X and Y are independent in P, e.g.,

G_0 is an I-map of P
G_1 is an I-map of P
G_2 is an I-map of P

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

X and Y are not independent in P
Thus $(X \perp Y) \setminus \in I(P)$

G_0 is not an I-map of P
G_1 is an I-map of P
G_2 is an I-map of P

If G is an I-map of P then it captures some of the independences, not all

I-map to Factorization

- A Bayesian network G encodes a set of conditional independence assumptions $I(G)$
- Every distribution P for which G is an I-map should satisfy these assumptions
  - Every element of $I(G)$ should be in $I(P)$
- This is the key property to allowing a compact representation

I-map to Factorization

- From chain rule of probability

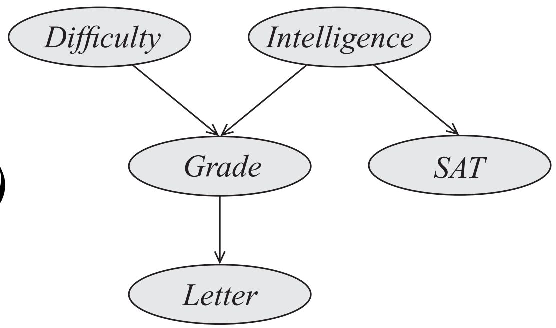
  $P(I,D,G,L,S)=P(I)P(D|I)P(G|I,D)P(L|I,D,G)P(S|I,D,G,L)$

  – Relies on no assumptions

  – Also not very helpful

    - Last factor requires evaluation of 24 conditional probabilities

- Apply conditional independence assumptions induced from the graph



  $D \perp I \in I(P)$ therefore $P(D|I)=P(D)$

  $(L \perp I,D) \in I(P)$ therefore $P(L|I,D,G)=P(L|G)$

  – Thus we get

  $$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

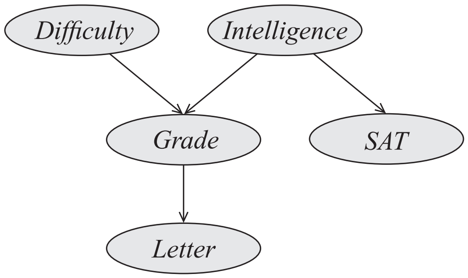
    - Which is a factorization into local probability models

  – Thus we can go from graphs to factorization of P

Factorization to I-map

- We have seen that we can go from the independences encoded in $\mathcal{G}$, i.e., $I(\mathcal{G})$, to Factorization of P
- Conversely, Factorization according to G implies associated conditional independences
  - If P factorizes according to $\mathcal{G}$ then $\mathcal{G}$ is an I-map for P
  - Need to show that if P factorizes according to G then $I(\mathcal{G})$ holds in P
  - Proof by example

Example that independences in $\mathcal{G}$ hold in P



- P is defined by set of CPDs
- Consider independences for S in G, i.e.,

$$P(S \perp D,G,L|I)$$

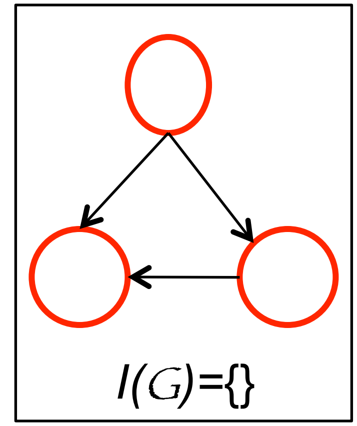
- Starting from factorization induced by graph

$$P(D,I,G,S,L) = P(D)P(I)P(G \mid D,I)P(S \mid I)P(L \mid G)$$

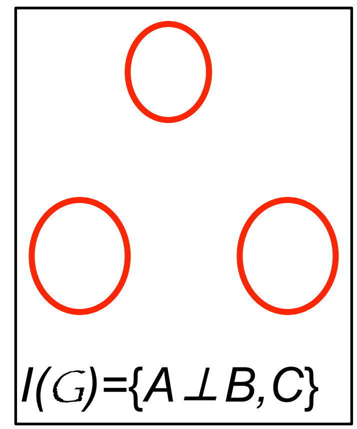
- Can show that $P(S|I,D,G,L)=P(S|I)$
- Which is what we had assumed for P

Perfect Map

- I-map
  - All independencies in $I(G)$ present in $I(P)$
  - Trivial case: all nodes interconnected



- D-Map
  - All independencies in $I(P)$ present in $I(G)$
  - Trivial case: all nodes disconnected



- Perfect map
  - Both an I-map and a *D*-map
  - Interestingly not all distributions P over a given set of variables can be represented as a perfect map
    - Venn Diagram where *D* is set of distributions that can be represented as a perfect map