Undirected Graphical Models

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Directed versus Undirected graphical models
- Components of a Markov Network
- Independence Properties
- Parameterization
- Gibbs Distributions and Markov Networks
- Reduced Markov Networks

Directed vs Undirected

- Bayesian Networks= Directed Graphical Models
  - Useful in many real-world domains
- Undirected Graphical Models
  - When no natural directionality exists betw. variables
  - Offer a simpler perspective on directed graphs
    - Independence structure
    - Inference task
  - Also called Markov Networks or MRFs
- Can combine directed and undirected graphs
  - Unlike in previous discussion on BNs attention restricted to discrete state spaces

Directed vs Undirected

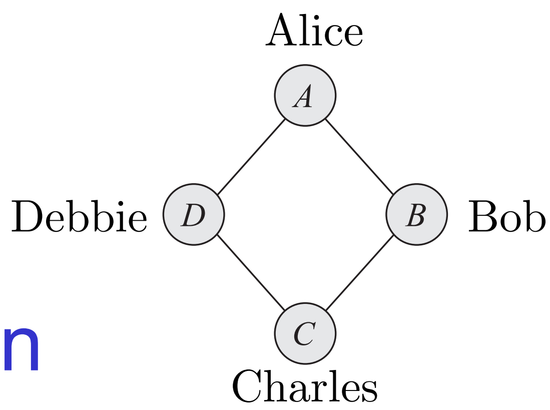
- Bayesian Networks= Directed Graphical Models
  - Useful in many real-world domains
- Undirected Graphical Models
  - When no natural directionality exists betw. variables
  - Offer a simpler perspective on directed graphs
    - Independence structure
    - Inference task
  - Also called Markov Networks or MRFs
- Can combine directed and undirected graphs
- Attention restricted to discrete state spaces

Example to Motivate Undirected Graphs

1. Four students study in pairs to work on homework

Alice and Bob are friends
Bob and Charles study together
Charles and Debbie argue with each other
Debbie and Alice study together
Alice and Charles can't stand each other
Bob and Debbie had relationship ended badly

A Social Network with
$(A \perp C | \{B, D\})$ and $(B \perp D | A, C)$



2. Professor may have mis-spoken

e.g., on a machine learning topic

3. Students may have figured out the problem

e.g., by thinking about issue
or by studying textbook

4. Students transmit this understanding to his/her study partner

Modeling Influences Using a BN

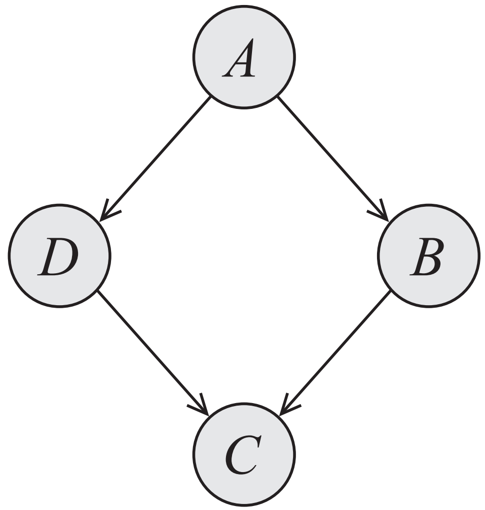
Probability of misconception of one person
depends on whether their study partner has a misconception

Alice and Charles never speak *directly*

Thus A and C should be conditionally independent given B and D

We need to model $(A \perp C \mid \{B,D\})$

Consider this Proposed Bayesian Network



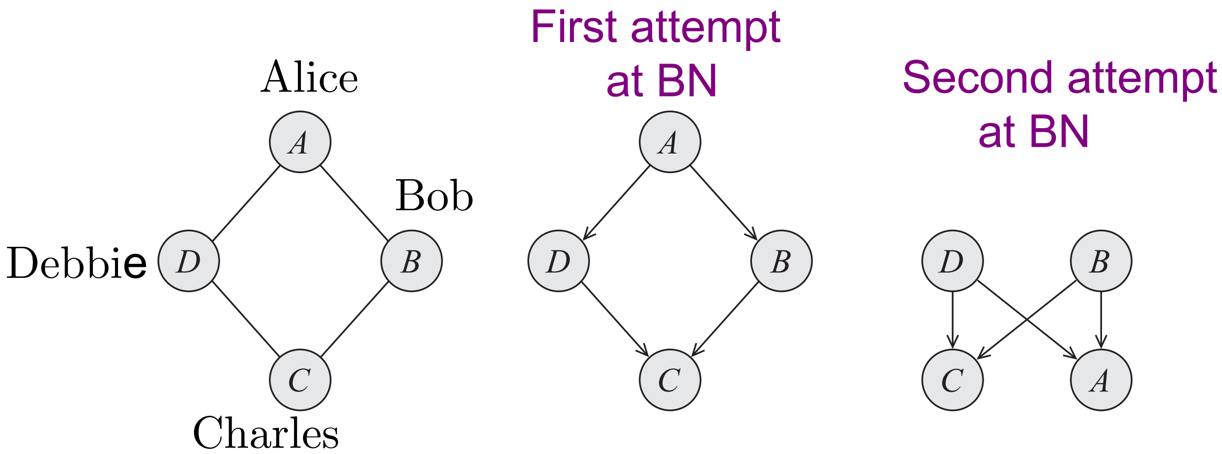
It does model $(A \perp C \mid \{B,D\})$
since the path between A and C is blocked when B ,D are knowr

But also means B and D are independent given only A
since V-structure through C implies blockage when C is not known
But dependent given both A and C
since V-structure through C implies no blockage when C is known

Lack of Perfect Map in BN

Misconception Example

We need to model $(A \perp C \mid B,D),(B \perp D \mid A,C)$



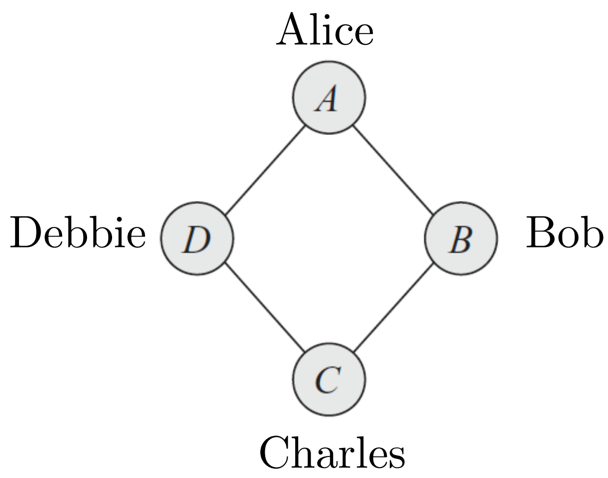
In both (b) and (c) $(A \perp C \mid B,D)$ holds
but $(B \not\perp D \mid A,C)$ due to *v-structure* $D \rightarrow C \leftarrow B$ in d-separation

No perfect map since independences imposed by BN are inappropriate for the distribution

In a perfect map the graph precisely captures the independencies in the given distribution

Drawbacks of BN in Example

- Independences imposed by BN are inappropriate for the distribution
- Interaction between the variables are symmetric



An attempt at Modeling the Misconception Problem

Four binary random variables representing whether or not student has misconception

$$X \in \{A, B, C, D\}$$

x^1 : student has the misconception

x^0 : Student does not have a misconcpetion

Probabilities assuming four variables are independent

a^0	a^1
0.3	0.7

b^0	b^1
0.2	0.8

etc

a^0 = has misconception
a^1 = no misconception

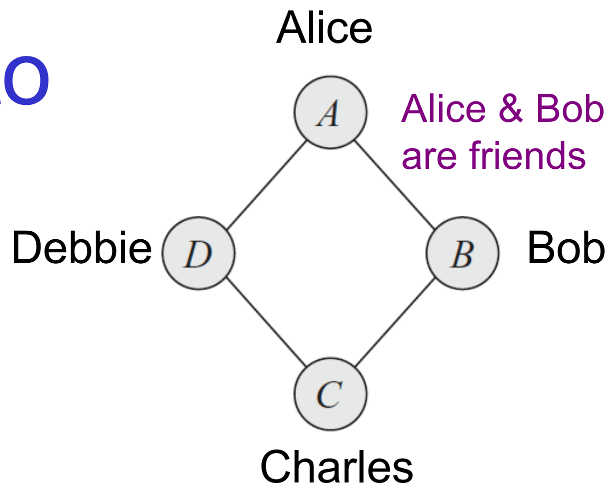
To get contextual interactions between variables we need a MN

Capturing affinities between variables

- Let $\boldsymbol{D}$ be a set of random variables
- A factor ϕ is a function from $Val(\boldsymbol{D})$ to R
  - where Val is the set of values that $\boldsymbol{D}$ can take
- A factor is non-negative if all its entries are non-negative
- The set of variables $\boldsymbol{D}$ is called the scope of the factor, denoted $Scope[\Phi]$
- In our example, $\phi_1(A,B)$: $Val(A,B) \rightarrow R^+$
  - Higher the value, more compatible they are

Example of a Factor

- Factor is not normalized to sum to one
- Values need not be in $[0,1]$
- $\phi_1(A,B)$ asserts that A and B
  - are more likely to agree than disagree
  - More weight to when they agree than when they disagree
    - Translates to higher probability for disagreeing values



Alice & Bob are friends

$\phi_1[A,B]$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

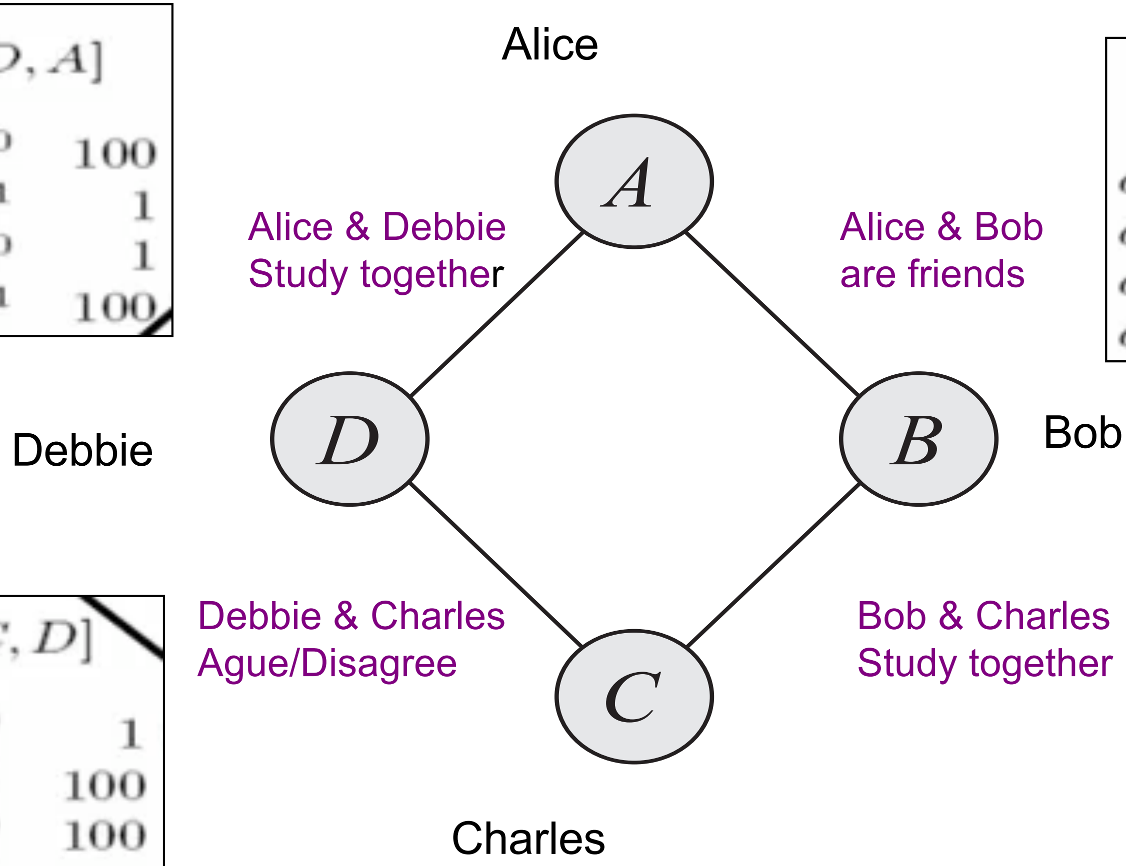
a^0 = has misconception
a^1 = no misconception

b^0 = has misconception
b^1 = no misconception

Factors for Misconception Example

$\phi_4[D, A]$		
d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10



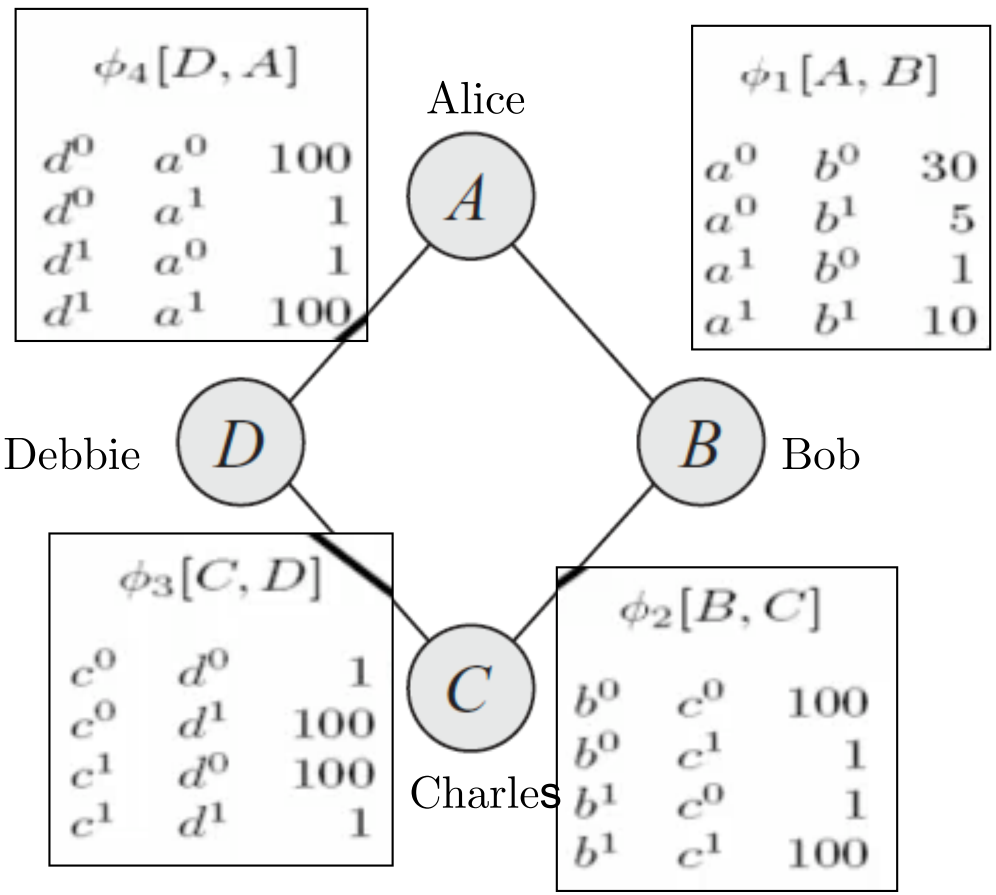
$\phi_3[C, D]$		
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$		
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Pairwise Compatibility Factors

a^0 = has misconception
a^1 = no misconception

b^0 = has misconception
b^1 = no misconception



$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_2[B, C]$

b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

Charles and Debbie are incompatible

Most likely instantiations are when they disagree

Factor is similar to CPD

For each combination there is a value

Combining Local Models into Global

- As in BNs parameterization of Markov network defines local interactions
- We combine local models by multiplying them

$$\phi_1(A,B)\cdot\phi_2(B,C)\cdot\phi_3(C,D)\cdot\phi_4(D,A)$$

Alice, Bob and Debbie have misconception
Charles does not

$$\phi_1(a^1,b^1)\cdot\phi_2(b^1,c^0)\cdot\phi_3(c^0,d^1)\cdot\phi_4(d^1,a^1)=$$
$$10.1.100.100=100{,}000$$

- Convert to a legal distribution

  by performing a normalization

Assignment				Unnormalized
a^0	b^0	c^0	d^0	300000
a^0	b^0	c^0	d^1	300000
a^0	b^0	c^1	d^0	300000
a^0	b^0	c^1	d^1	30
a^0	b^1	c^0	d^0	500
a^0	b^1	c^0	d^1	500
a^0	b^1	c^1	d^0	5000000
a^0	b^1	c^1	d^1	500
a^1	b^0	c^0	d^0	100
a^1	b^0	c^0	d^1	1000000
a^1	b^0	c^1	d^0	100
a^1	b^0	c^1	d^1	100
a^1	b^1	c^0	d^0	10
a^1	b^1	c^0	d^1	100000
a^1	b^1	c^1	d^0	100000
a^1	b^1	c^1	d^1	100000

Normalized Joint Distribution

$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)$$

where

$$Z = \sum_{a,b,c,d} \phi_1(a,b)\cdot\phi_2(b,c)\cdot\phi_3(c,d)\cdot\phi_4(d,a)$$

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

Z=7,201,840

Z is a normalizing constant called the Partition function

"Partition" originates from Markov Random Fields in statistical physics

- Energy of a physical system of interacting atoms

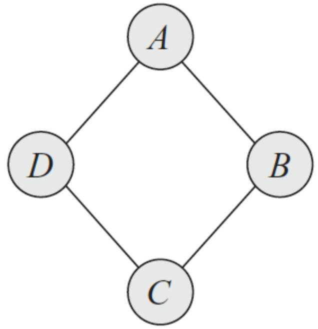
"Function" is because Z is a function of the parameters

Answering Queries

- We can obtain any desired probability from the joint distribution as usual

$P(b^0)=0.268$: Bob is 26% likely to have a misconception

$P(b^1|c^0)=0.06$: if Charles does not have the misconception, Bob is only 6% likely to have misconception.



- Most probable joint probability (from table): $P(a^0,b^1,c^1,d^0)=0.69$
  - Alice,Debby have no misconception, Bob,Charles have misconception

$\phi_1[A,B]$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_2[B,C]$		
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3[C,D]$		
c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_4[D,A]$		
d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

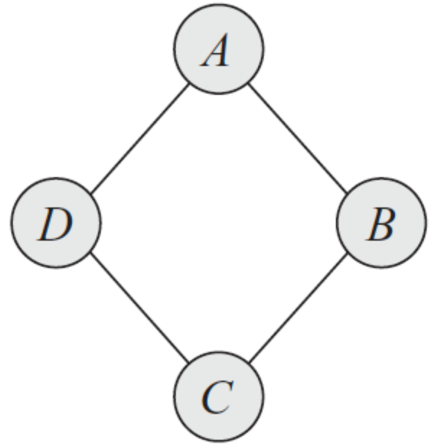
Benefit of MN representation

- Flexibility in representing interactions between variables
- E.g., if we want to change interaction between A and B simply modify the entries in that factor
- Flip side of flexibility is that the effects of these changes are not intuitive

Factorization and Independence

- Tight connection between factorization and independencies (as in BNs)
  - P supports $(X \perp Y \mid Z)$ iff we can write distribution as $P(X) = \phi_1(X,Z)\phi_2(Y,Z)$
    - In our example, we can write

$$P(A,B,C,D) = \underbrace{\left[\frac{1}{Z}\phi_1(A,B)\cdot\phi_2(B,C)\cdot\right]}_{\text{Factor with } \{B,\{A,C\}\}}\underbrace{\phi_3(C,D)\cdot\phi_4(D,A)}_{\text{Factor with } \{D,\{A,C\}\}}$$



    - Therefore $(B \perp D \mid A, C)$
    - By grouping factors with $\{A,\{B,D\}\}$ and $\{C,\{B,D\}\}$
    - We get $(A \perp C \mid B, D)$
    - Precisely independences we could not get with a BN

- Independencies correspond to separation properties in graph over which P factorizes

Parameterization

- Need to associate graph with parameters
- Not as intuitive as CPDs in Bayesian networks
  - Factors do not correspond to probabilities or CPDs
    - Not intuitively understandable
  - Hard to elicit from people
  - Significantly harder to estimate from data
    - As seen with algorithms to learn MNs

Factors have no constraints

$\phi_1[A,B]$		
a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

- A factor ϕ is a function from $Val(D)$ to R
  - Subsumes notion of both joint distribution and CPD
    - A joint distribution over D is a factor over D
      - » Specifies a real no for every value of D
    - A conditional distrib. $P(X|U)$ is a factor over $\{X\} \cup U$
- While CPDs and joint distributions should satisfy normalization constraints (sum to one)
  - There are no constraints on parameters in a factor

Factors

- Factor describes *compatibilities* between different values of variables in its scope
- A graph can be parameterized by associating a set of factors with it
- One idea: associate a factor with each edge
  - This is insufficient for a full distribution as shown in example next

Pairwise Factors

- Associating parameters with each edge is insufficient to specify arbitrary joint distribution
- Consider a fully connected graph
  - There are no independence assumptions
- If all variables are binary
  - Each factor over an edge has 4 parameters
  - Total number of parameters is $4\ {}^{n}\mathrm{C}_2$ (a quadratic in n)
  - An arbitrary distribution needs 2^n-1 parameters
- Pairwise factors have insufficient parameters
  - Parameters cannot be just associated with edges

Factors over arbitrary subsets

- More general than factors over edges (pairwise)
- To provide formal definition we need idea of a factor product

Definition of Factor Product

Let X, Y and Z be three disjoint sets of variables

$\phi_1(X,Y)$ and $\phi_2(Y,Z)$ are two factors

Factor product $\phi_1 \times \phi_2$ is a factor

$\psi\colon Val(X,Y,Z) \rightarrow R$

as follows:

$\psi(X,Y,Z) = \phi_1(X,Y).\phi_2(Y,Z)$

$\phi_1(A,B)$

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

$\phi_2(B,C)$

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

$\psi(A,B,C)$

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

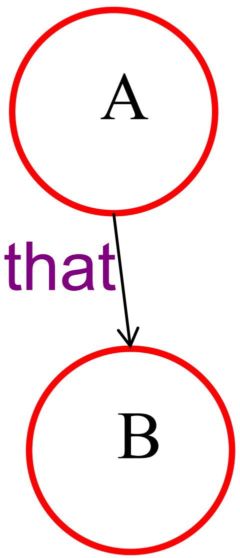
Two factors are multiplied in a way that matches-up the common part Y

In the example, factors do not correspond to either probabilities or conditional probabilities

Factor Product and Bayesian network

- Both joint distributions and CPDs are factors
  - since they are mappings of $Val(D)$ to R
  - Chain rule of Bayesian networks defines joint distribution as product of CPD factors
  - Example, in computing $P(A,B)=P(A)P(B|A)$

    We multiply entries in the $P(A)$ and $P(B\,|A)$ tables that have the same value for A



- Letting $\phi_{X_i}(X_i,Pa_{X_i})$ represent $P(X_i|PaX_i)$
  - we have

$$P(X_1,..X_n)=\prod_i \phi_{X_i}$$

Gibbs Distributions

- Generalize the idea of factor product to define an undirected parameterization of a distribution
- A distribution P_Φ is a Gibbs distribution parameterized by a set of factors $\boxed{\Phi = \{\phi_1(D_1),\ldots,\phi_K(D_K)\}}$
  - If it is defined as follows

D_i are sets of random variables

$$P_\Phi(\mathrm{X}_1,..\mathrm{X}_n) = \frac{1}{Z}\tilde{P}(\mathrm{X}_1,..\mathrm{X}_n)$$

where

$$\tilde{P}(\mathrm{X}_1,..\mathrm{X}_n) = \prod_{i=1}^{m} \phi_i(D_i)$$

is an unnomalized measure and

$$Z = \sum_{\mathrm{X}_1,..\mathrm{X}_n} \tilde{P}(\mathrm{X}_1,..\mathrm{X}_n) \text{ is a normalizing constant}$$

called the partition function

The factors do not necessarily represent the marginal distributions of the variables in their scopes

Influence of Factors

- Factors do not represent marginal probabilities of the variables within their scope (as we will see)
- A factor is only one contribution to the overall joint distribution
- Distribution as a whole has to take into consideration contributions from all factors involved
- Example given next

Factors ≠ Marginal Probabilities

Joint Distribution

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

Marginal distribution over Alice,Bob

$\boldsymbol{a^0}$	$\boldsymbol{b^0}$	**0.13**
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

Factor over Alice,Bob

$\phi_1[A, B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

Most likely: $a^0, b^1(disagree)$ $a^0, b^0(agree)$

Because probability takes into account influence of other factors

$\phi_2[B, C]$

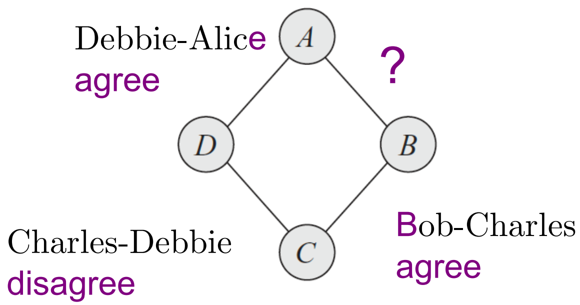
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3[C, D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

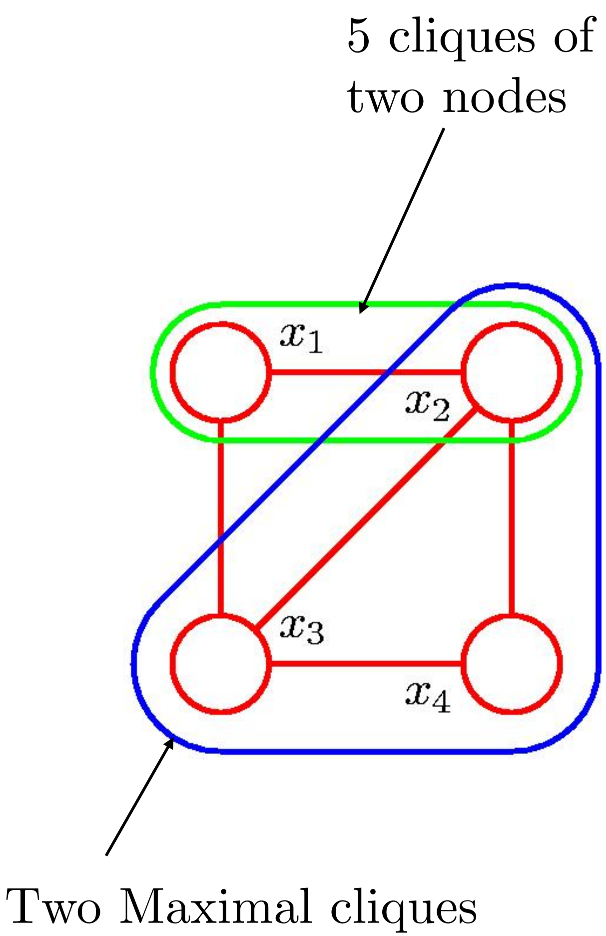
$\phi_4[D, A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100



Relating parameterization to graph

- Clique
  - Subset of nodes in graph such that there exists a link between all pairs of nodes in subset
    - Set of nodes in clique are fully connected
- Maximal Clique
  - Not possible to include any other nodes in the graph in the set without ceasing to be a clique



Gibbs distribution and Graph

- A distribution P_Φ with $\Phi = \{\phi_1(D_1),..,\phi_K(D_K)\}$
  - factorizes over a Markov network $\mathcal{H}$ if each D_k is a complete subgraph (clique) of $\mathcal{H}$
- Factors that parameterize network are called clique potentials
- Number of factors are reduced, e.g.,

Four Maximal Cliques:
$\{A,B\},\{B,C\}$
$\{C,D\},\{D,A\}$
with possible parameter settings:

$\phi_1[A,B]$

a^0	b^0	30
a^0	b^1	5
a^1	b^0	1
a^1	b^1	10

$\phi_2[B,C]$

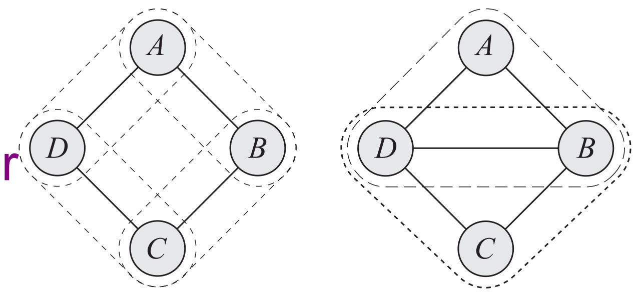
b^0	c^0	100
b^0	c^1	1
b^1	c^0	1
b^1	c^1	100

$\phi_3[C,D]$

c^0	d^0	1
c^0	d^1	100
c^1	d^0	100
c^1	d^1	1

$\phi_4[D,A]$

d^0	a^0	100
d^0	a^1	1
d^1	a^0	1
d^1	a^1	100



(a) (b)

Two Maximal Cliques:
$\{A,B,D\}$,
$\{B,C,D\}$

Maximal Cliques have a Disadvantage

- No of parameters reduced by allowing factors only for maximal cliques
  - But obscures structure present
- Example: All n variables are binary
  - Pairwise cliques
    - Each factor over an edge has 4 parameters
    - Total number of parameters is $4\ {}^{n}\mathrm{C}_2$
      - Quadratic number of parameters
  - An arbitrary distribution (Single clique)
    - needs 2^n-1 parameter
      - Exponential number of parameters

Factors as Cliques

- Functions of cliques
- Set of variables in clique C is denoted x_C
- Joint distribution is written as a product of *potential functions*

$$p(\mathrm{x}) = \frac{1}{Z}\prod_C \psi_C(\mathrm{x}_C)$$

- Where Z , the *partition function*, is a normalization constant

$$Z = \sum_{\mathrm{x}}\prod_C \psi_C(\mathrm{x}_C)$$

Pairwise Markov Network

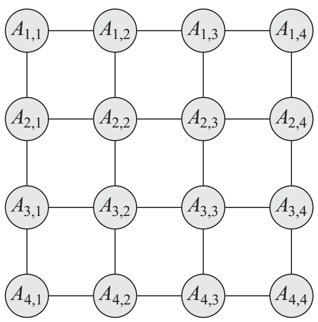
- Subclass of Markov networks commonly encountered, e.g.,
  1. Ising model and Boltzmann machines
  2. Computer Vision

- All factors are over single variables Or pairs of variables

- A pairwise Markov network over graph $\mathcal{H}$
  - has a set of node potentials $\phi(X_i)$ and edge potentials $\phi(X_i, X_j)$

- Although simple they pose a challenge for inference algorithms

- Special Case when arranged as a grid



Reduced Markov Networks

a^1	b^1	c^1	$0.5 \cdot 0.5 = 0.25$
a^1	b^1	c^2	$0.5 \cdot 0.7 = 0.35$
a^1	b^2	c^1	$0.8 \cdot 0.1 = 0.08$
a^1	b^2	c^2	$0.8 \cdot 0.2 = 0.16$
a^2	b^1	c^1	$0.1 \cdot 0.5 = 0.05$
a^2	b^1	c^2	$0.1 \cdot 0.7 = 0.07$
a^2	b^2	c^1	$0 \cdot 0.1 = 0$
a^2	b^2	c^2	$0 \cdot 0.2 = 0$
a^3	b^1	c^1	$0.3 \cdot 0.5 = 0.15$
a^3	b^1	c^2	$0.3 \cdot 0.7 = 0.21$
a^3	b^2	c^1	$0.9 \cdot 0.1 = 0.09$
a^3	b^2	c^2	$0.9 \cdot 0.2 = 0.18$

Factor is reduced to the context $C = c^1$

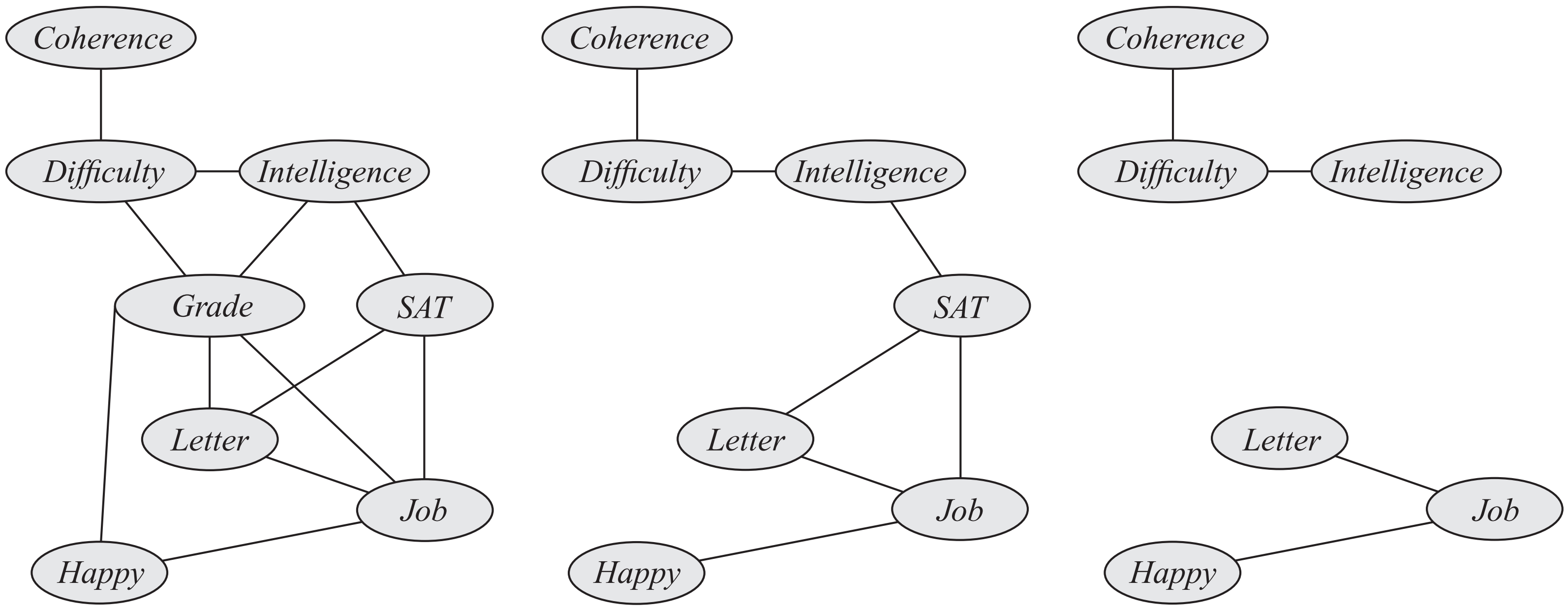
a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

Conditioning distribution on some assignment $\boldsymbol{u}$ to set of variables $\boldsymbol{U}$

Conditioning a distribution is to eliminate all entries in the joint distribution that are inconsistent with a subset of variables $\boldsymbol{U}=\boldsymbol{u}$

and renormalizing remaining entries to sum to 1

Reduced Markov Network Example



(a) (b) (c)

Initial set of factors

Reduced to context $G{=}g$

Reduced to context $G{=}g,\ S{=}s$