CMPUT463/563
Probabilistic Graphical Models

Representation: Markov Network

Lili Mou

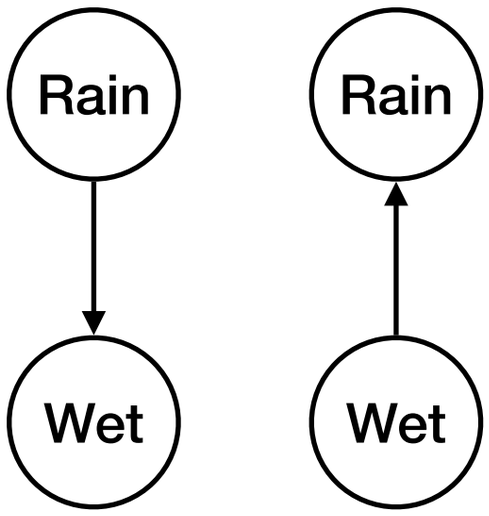
Dept. Computing Science, University of Alberta

lmou@ualberta.ca

Recap: Bayesian Network

BN models the relationship of (random variables) RVs in a “causal” manner

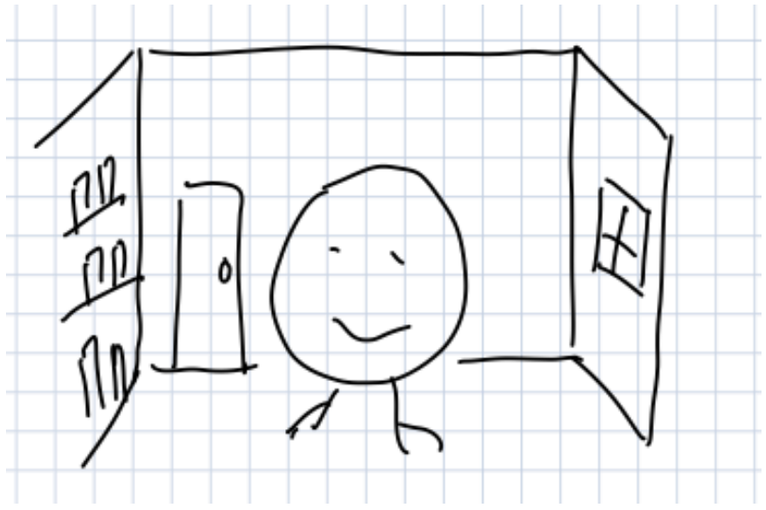
The “cause-and-effect” here is better interpreted as reasoning patterns, rather than physical causality

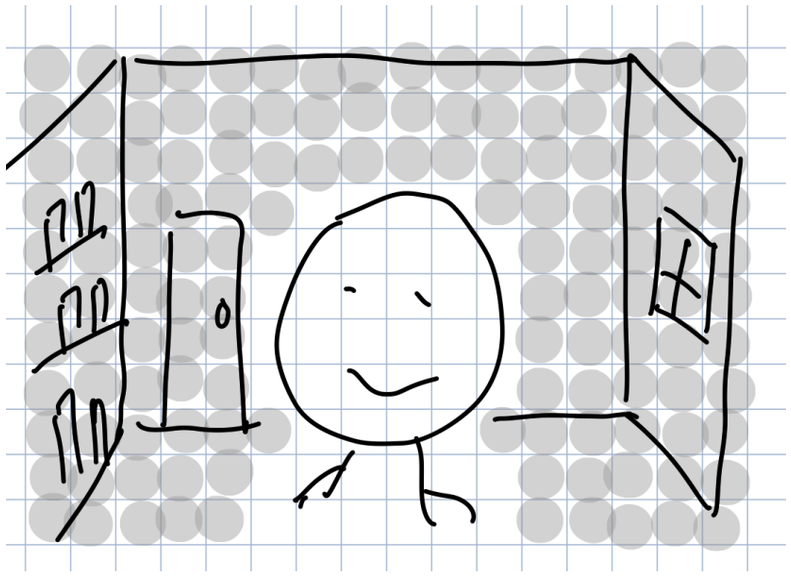


Beyond BN

In reality, the relationship between RVs may be vague and non-causal.

Example:

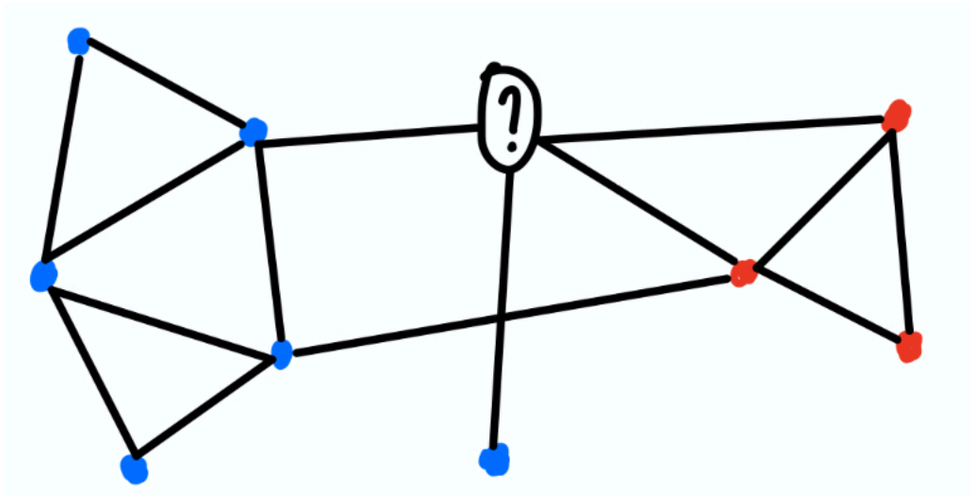




Pixels are related

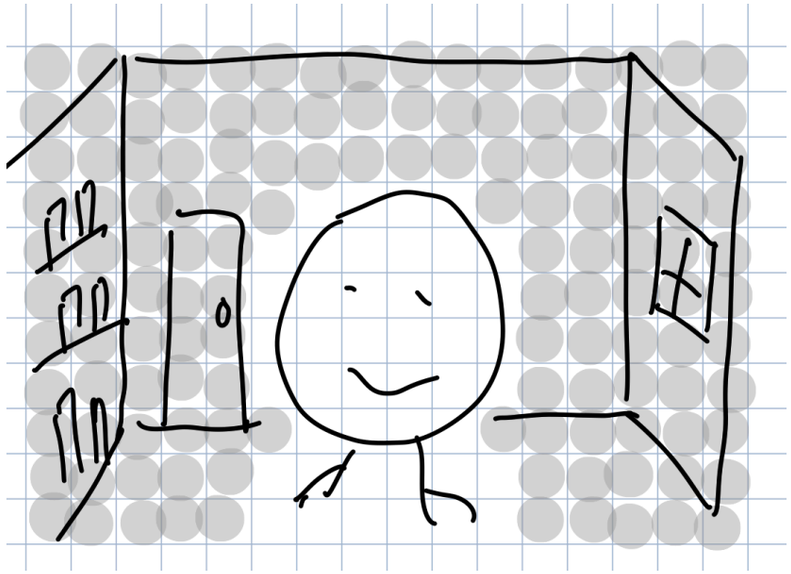
Humans are related

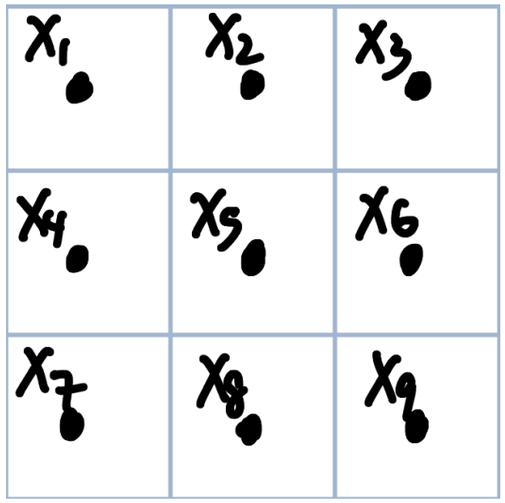
, but they are not causal

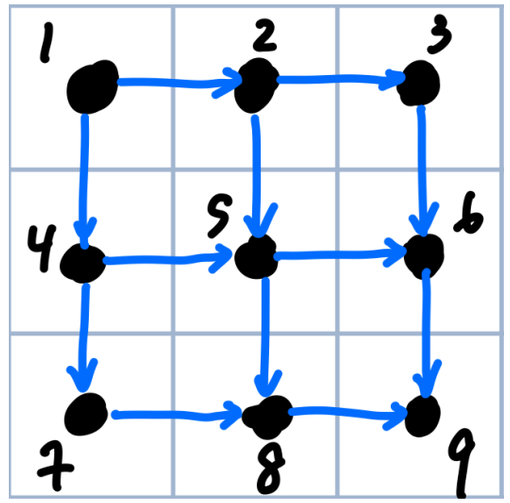


Modeling Such Relations by BN

Suppose the true dependencies are such that every pixel is immediately affected by its four neighbors



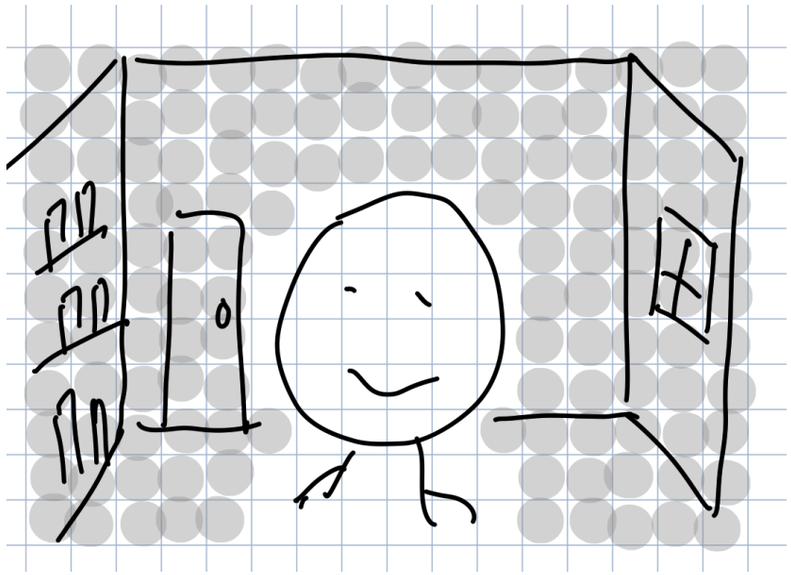


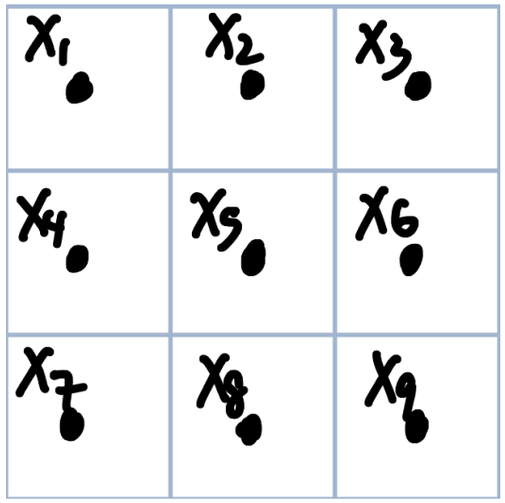


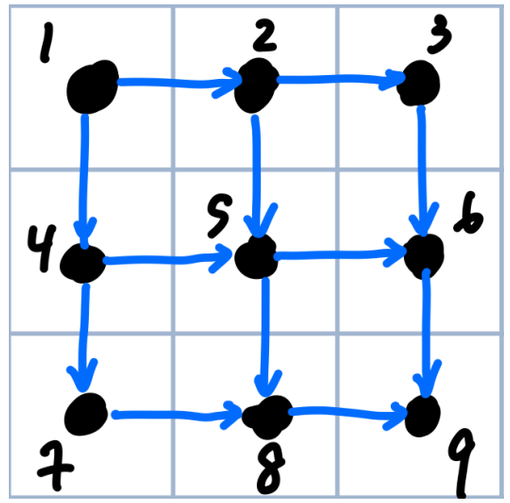
A plausible BN, but is it an I-map (capturing all dependencies)?

Modeling Such Relations by BN

Suppose the true dependencies are such that every pixel is immediately affected by its four neighbors







A plausible BN, but is it an I-map (capturing all dependencies)?

In this BN, $X_1 \perp X_3 \,|\, X_2$

But our assumption allows information flow $X_1 - X_4 - X_5 - X_6 - X_3$, meaning that $X_1 \perp X_3$ may not hold in general

To allow such dependencies, we need more edges.
But, very soon, we will get a fully connected BN

We need a new type of factorization

- $P?(X_1, \cdots, X_9) = P?(X_1?X_2) \cdot P?(X_1?X_4) \cdots \quad P?(\quad X_i?X_j \quad)$
- Here, we want to keep the factorization of BN.
- But then, it becomes questionable whether they are probabilities and what kind of probabilities
- Apparently, they're not probabilities, because factorizing by conditional probabilities must have a root node, but we don't have it
- So I'll call them scores $s(X_i, X_j)$

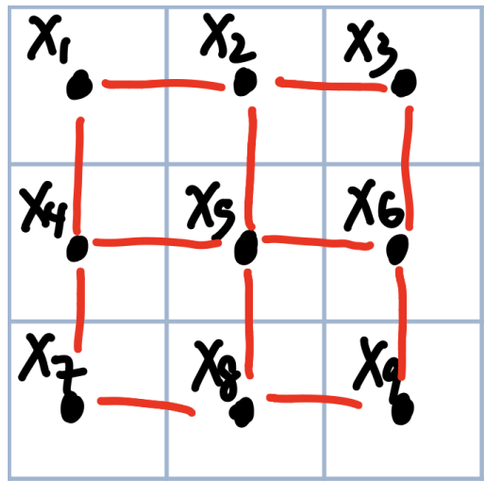
$$P?(X_1, \cdots, X_9) = s(X_1?X_2) \cdot s(X_1?X_4) \cdots$$

- Since the scores are multiplied together, they only make sense if non-negative, i.e., $s(X_i, X_j) \geq 0$
- Is the resulting number a probability? No, so I also call it a score.

$$s(X_1, \cdots, X_9) = s(X_1?X_2) \cdot s(X_1?X_4) \cdots$$

- How can we get a probability? Normalize it.

$$P(X_1, \cdots, X_9) = \frac{s(X_1, \cdots, X_9)}{\sum_{x1,\cdots,x_9} s(x_1, \cdots, x_9)}$$



Markov Random Field

Consider random variables $X_1, \cdots, X_n$

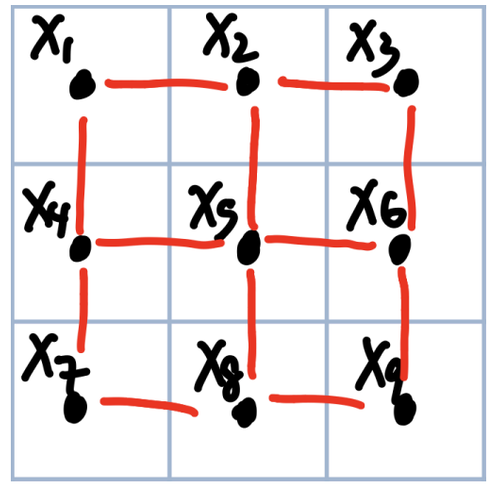
Def (scope). A scope D is a subset of $X_1, \cdots, X_n$.

Def (factor/potential). A factor ϕ is mapping the value assignment of its scope to a nonnegative number

A factor is similar to a conditional probability table, except that it does not normalize.

However, it's not a probability by itself. It reflects local "happiness" of "goodness" for this value assignment.

X_1	X_2	ϕ
0	0	2.0
0	1	0.1
1	0	0.5
1	1	10.0



Markov Random Field

We next define the joint probability

Def (Unnormalized measure) $$\tilde{p}(X_1, \cdots, X_n) = \prod_{k=1}^{K} \phi_k(X_1, \cdots, X_n)$$

ϕ_k only concerns those $X_i \in D_k$; a constant wrt other $X_i \notin D_k$

Same as BN factorization, except that it's not normalized

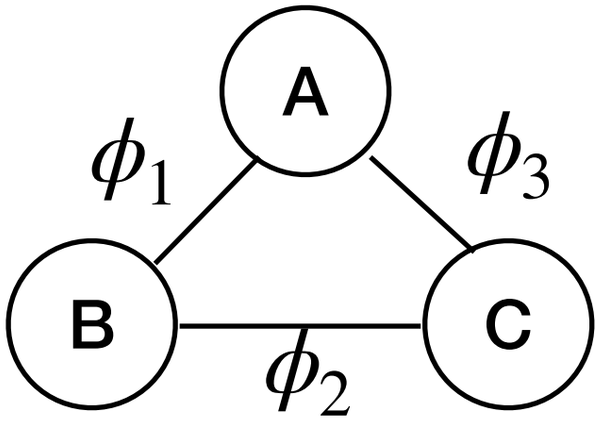
Def (Partition function)

$$Z = \sum_{x_1, \cdots x_n} \tilde{p}(x_1, \cdots, x_n) = \sum_{x_1, \cdots x_n} \prod_{k=1}^{K} \phi_k(x_1, \cdots, x_n)$$

Def (Probability)

$$P(X_1, \cdots, X_n) = \frac{1}{Z}\tilde{P}(X_1, \cdots, X_n) = \frac{\prod_{k=1}^{K} \phi_k(X_1, \cdots, X_n)}{\sum_{x'_1, \cdots, x'_n} \prod_{k=1}^{K} \phi_k(x'_1, \cdots, x'_n)}$$

This is known as a Boltzmann distribution or a Gibbs distribution



Example

A, B, C: three students, whether a student understands PGM correctly. 0: incorrect, 1: correct

A	B	ϕ_1
0	0	2
0	1	0.1
1	0	0.2
1	1	10

B	C	ϕ_2
0	0	0.2
0	1	5
1	0	3
1	1	0.3

A	C	ϕ_3
0	0	5
0	1	0.4
1	0	0.5
1	1	3

A	B	C	$\tilde{P}$	Z	P
0	0	0	2	34.532	0.0579172941
0	0	1	4		0.1158345882
0	1	0	1.5		0.04343797058
0	1	1	0.012		0.0003475037646
1	0	0	0.02		0.000579172941
1	0	1	3		0.08687594116
1	1	0	15		0.4343797058
1	1	1	9		0.2606278235

Quick observation: If a factor is multiplied by a positive constant, the joint distribution does not change

Markov Network

MRF and MN roughly refer to the same thing, but MRF emphasizes the probability, and MN emphasizes the graph.

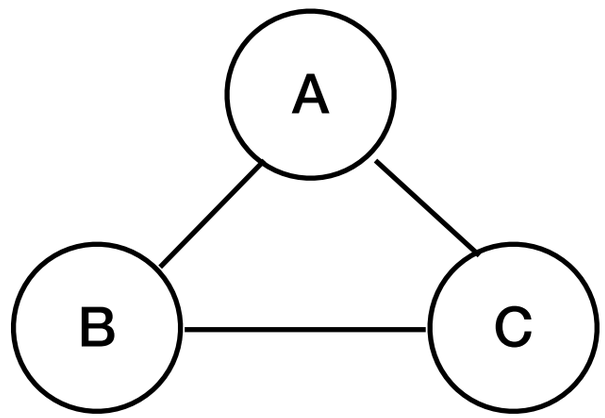
Given an MRF with scopes $\{D_k\}_{k=1}^{K}$. A Markov network can be induced by drawing an undirected edge $X_i - X_j$ iff X_i and X_j appear in (at least) one factor, i.e., there exists k such that $X_i \in D_k$ and $X_j \in D_k$.

Example 1

$$scope(\phi_1) = \{A, B\}$$
$$scope(\phi_2) = \{B, C\}$$
$$scope(\phi_3) = \{A, C\}$$

Example 2

$$scope(\phi_1) = \{A, B, C\}$$



Observation: factors-> graph is unique; graph -> factors not unique

Factorization over Markov Network

Def: A distribution $P(X_1, \cdots, X_n)$ factorizes over a Markov network H if $P = \frac{1}{Z}\prod_{k=1}^{K} \phi_k(D_k)$ such that every D_k is a complete subgraph of H.

The distributions in both Example 1 and Example 2 factorize over the MN.

Example 1

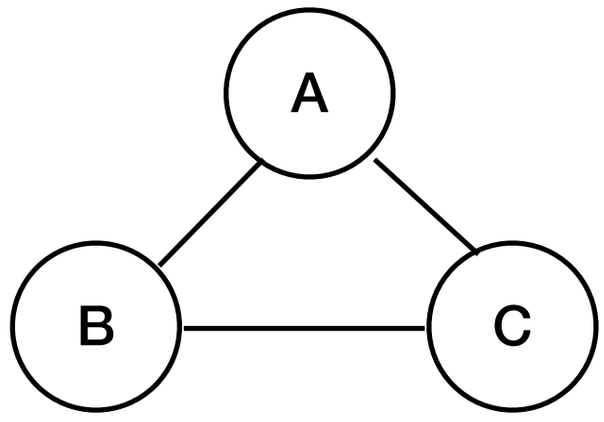
$scope(\phi_1) = \{A, B\}$

$scope(\phi_2) = \{B, C\}$

$scope(\phi_3) = \{A, C\}$

Example 2

$scope(\phi_1) = \{A, B, C\}$



Observation: factors-> graph is unique; graph -> factors not unique

Markov Network

Consider a complete graph (clique) of N variables each taking K values

- A giant factor of all variables: $(K^N - 1)$ free parameters, essentially modeling the joint table (except a normalizing factor)
- Pair-wise factor: $(K^2 - 1)N(N-1)/2$ free parameters

Conclusion: Giant factors are more powerful; small factors are more restricted, although the induced MN may be the same

Again, important difference between MN and BN

- BN: you can read out the factorization
- MN: you cannot

Modeling Power of MN

Any distribution defined on a finite set of finite-value discrete variables can be modeled by a MN.

The easy hack: a factor involving all variables

But no advantage is gained compared with a full probability table

Factor Graph

Factor graph is a bipartite graph

Consider an MRF with scopes $\{D_k\}_{k=1}^{K}$

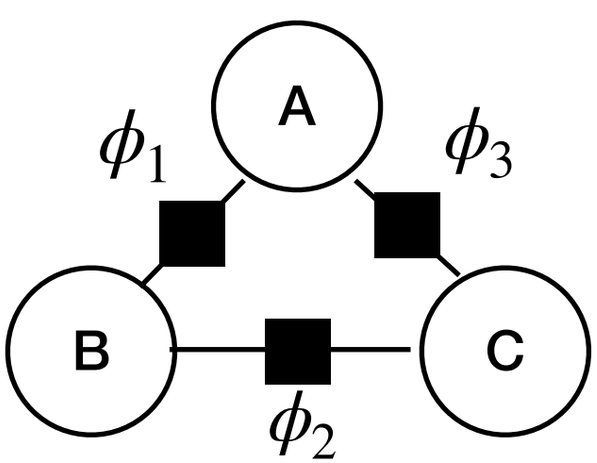
- Draw a square for each factor ϕ_k
- Draw a circle for each variable X_i
- Draw an undirected edge $X_i - D_k$ iff $X_i \in D_k$

Example 1

$scope(\phi_1) = \{A, B\}$

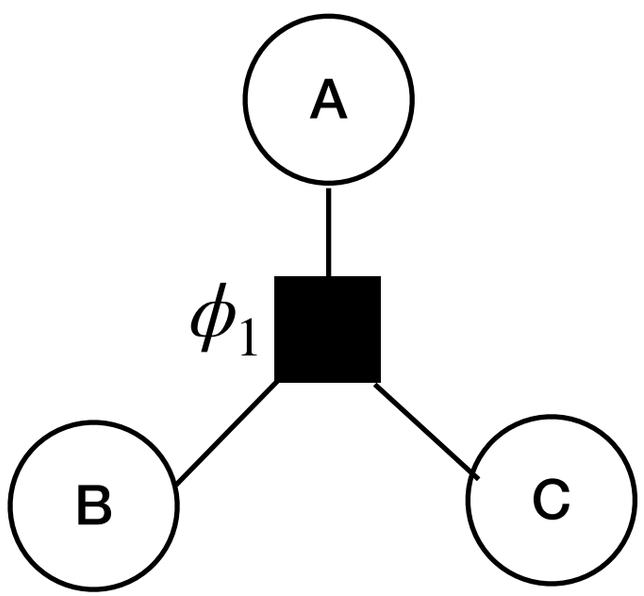
$scope(\phi_2) = \{B, C\}$

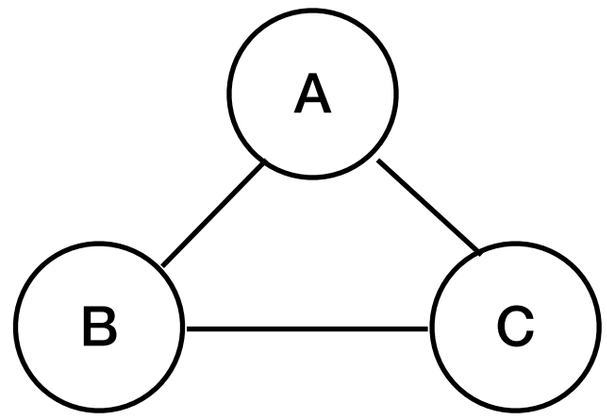
$scope(\phi_3) = \{A, C\}$



Example 2

$scope(\phi_1) = \{A, B, C\}$





Independencies in Markov Network

Consider a MN H with nodes $X_1, \cdots, X_n$.

Def (active trail). $X_1 - \cdots - X_k$ is an active trail given $\boldsymbol{Z}$ if none of $X_1, \cdots, X_k$ is in $\boldsymbol{Z}$ (i.e., none of the variables are observed)

Def (separation): $\boldsymbol{X}$ and $\boldsymbol{Y}$ are separated given $\boldsymbol{Z}$ in H if there does not exist an active path from $\boldsymbol{X}$ to $\boldsymbol{Y}$

Def (independencies captured by H).
$I(H) = \{(\boldsymbol{X} \perp \boldsymbol{Y} | \boldsymbol{Z}) : sep_H(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z})\}$

Independencies in Markov Network

Soundness. Suppose P factorizes over H. If $sep_H(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$, then $X \perp Y \mid \boldsymbol{Z}$ in P.

Completeness. If not $sep_H(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$, then $\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$ does not hold in some P factorizing over H.

Hammersley—Clifford Theorem: Let P be a positive distribution and H be a MN. If H is an I-map for P, i.e., $I(H) \subseteq I(P)$, then P factorizes over H.

Proof of soundness (Suppose P factorizes over H. If $sep_H(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$, then $X \perp Y \mid \boldsymbol{Z}$ in P**)**

Without loss of generality, we assume H only contains $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$.

Otherwise, $\boldsymbol{Z}$ still separates H into two parts. Just include more variables

Since there's no connections between $\boldsymbol{X}$ and $\boldsymbol{Y}$, we can group factors into two subsets

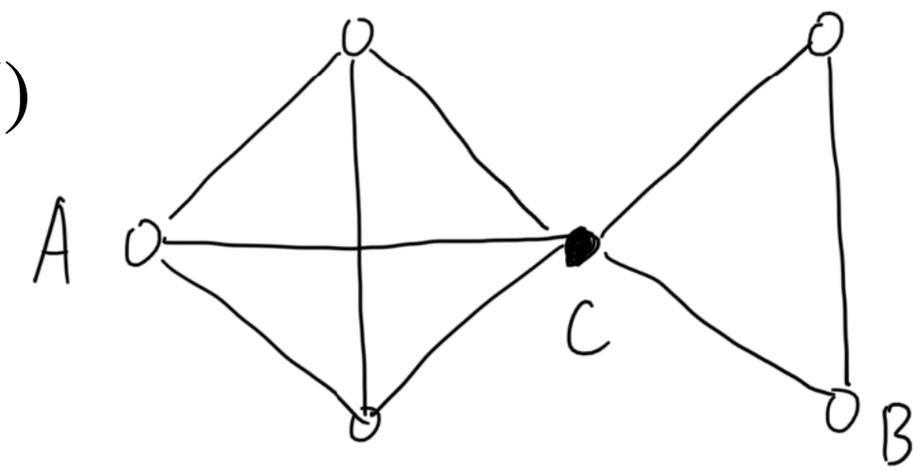
- I_X: the factors whose scopes are contained in $\boldsymbol{X} \cup \boldsymbol{Z}$
- I_Y: other factors

Then, $P(X_1, \cdots, X_n) = \frac{1}{Z} \prod_{i \in I_X} \phi_i(D_i) \prod_{i \in I_Y} \phi_i(D_i) = f(\boldsymbol{X}, \boldsymbol{Z}) g(\boldsymbol{Y}, \boldsymbol{Z})$ for some functions f, g

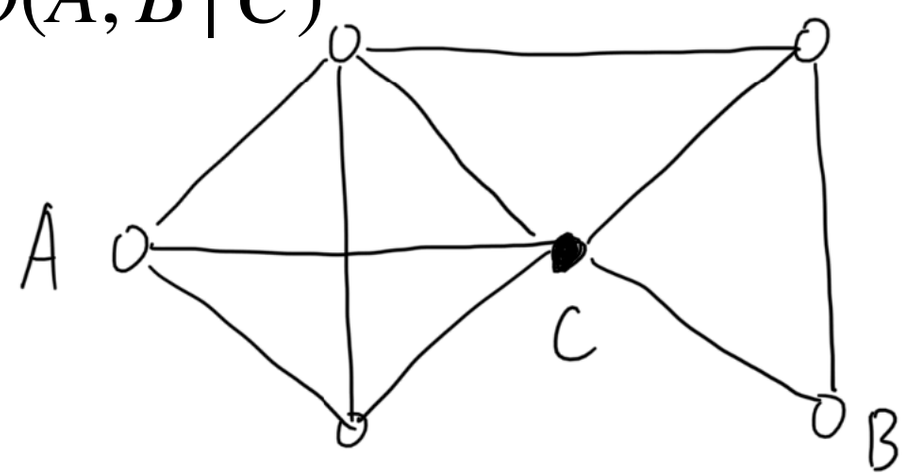
When $\boldsymbol{Z}$ is given, $P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = \tilde{f}(\boldsymbol{X}) \tilde{g}(\boldsymbol{Y})$ for some $\tilde{f}, \tilde{g}$ Next step = HW

Example

$sep(A, B \mid C)$

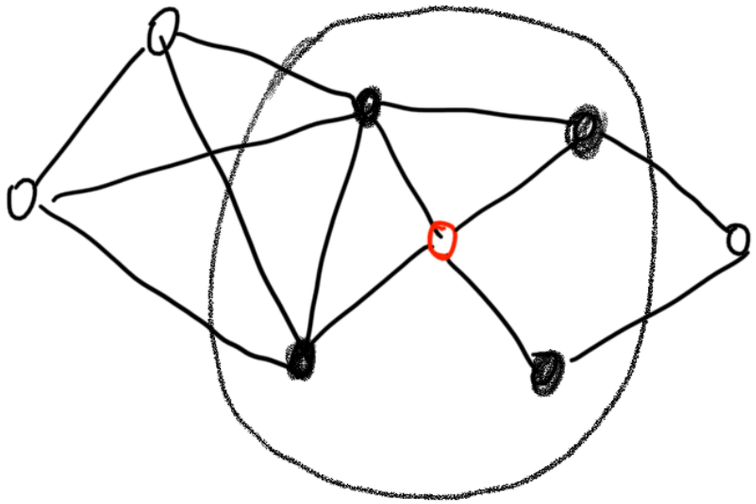


Not $sep(A, B \mid C)$

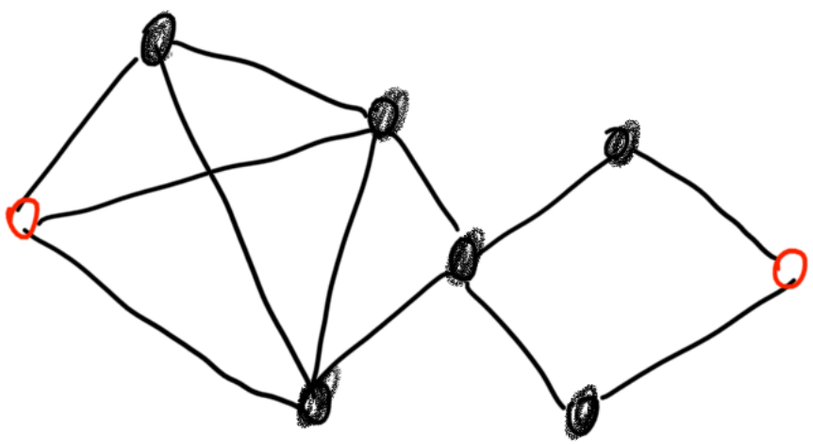


Independencies in Markov Network

Local Markov property: If a node's neighbors (called Markov blanket) are given, then it's independent of any other nodes.



Pairwise Markov property: X_i and X_j are independent if they are not directly connected and all other variables are given.



Reparametrization

Parametrizing MRF by multiplicative factors is analogous to BNs, but is inconvenient in future analysis.

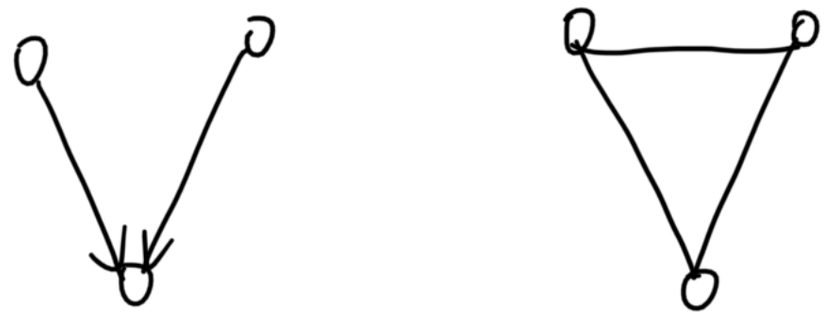
We may represent parametrize a factor by its log value. Then, multiplicative factors become additive in the log space. To recover the unnormalized measure, we need to take the exponential

A	B	ϕ	$\theta = \log \phi$
0	0	2	log 2
0	1	0.1	log 0.1
1	0	0.2	log 0.2
1	1	10	log 10

Converting BN to MN

Many results are derived over MN. It's sometimes desired to convert a BN to an MN.

- By "converting" it means that it cannot drop dependencies
- Method: moralizing, i.e., marrying the parents of a v-structured node pairwise.



- Independencies in MN is monotonic, i.e., more observations ⇒ more independencies
- In MN, $X \perp Y$ but $X \not\perp Y \mid Z$ cannot be modeled in MN. Remedy: Drop the independencies of $X \perp Y$ (not a big issue)

Log-Linear Model

$$\tilde{p}(X_1, \cdots, X_n) = \exp\left\{ \sum_{k=1}^{K} \sum_{\boldsymbol{x}'|_{D_k}} \mathbf{1}\{\boldsymbol{X}|_{D_k} = \boldsymbol{x}'|_{D_k}\}\theta_{\boldsymbol{x}'|_{D_k}} \right\}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function, and
$|$ refers to a set of variables restricted to a scope

In general, each row of each factor can be thought of as a feature. Suppose, we have j features in total.

$$\tilde{p}(X_1, \cdots, X_n) = \exp\left\{ \sum_{i=1}^{j} f_i(X_1, \cdots, X_n)\theta_i \right\}$$

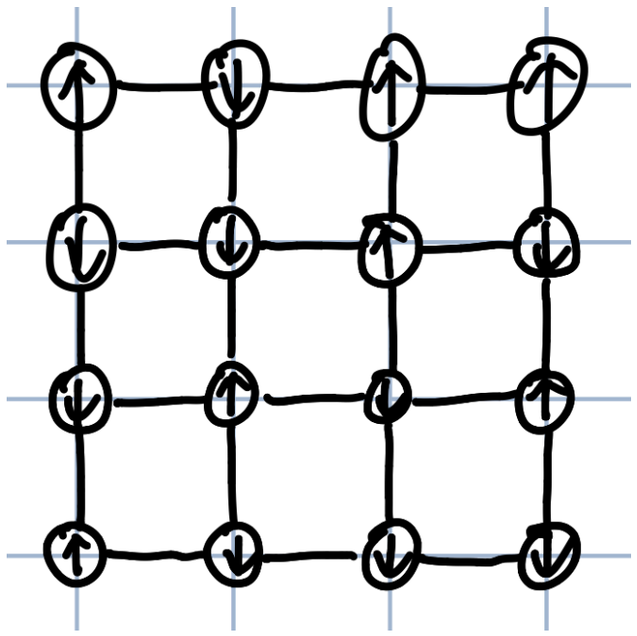
Called a score or negative energy

A	B	ϕ	$\theta = \log \phi$
0	0	2	log 2
0	1	0.1	log 0.1
1	0	0.2	log 0.2
1	1	10	log 10

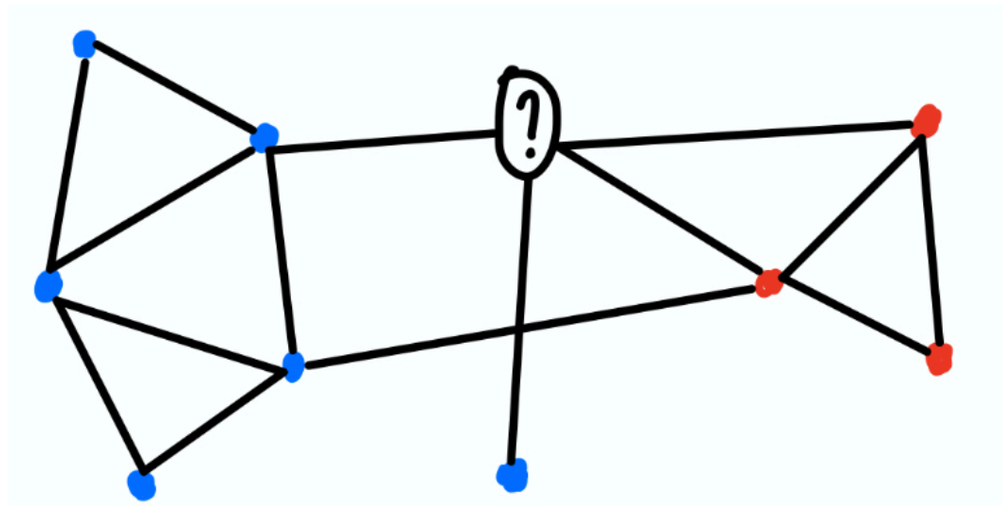
The features can be more general than a table lookup factor. Thus, the log-linear representation is more general.

Applications

Ising models



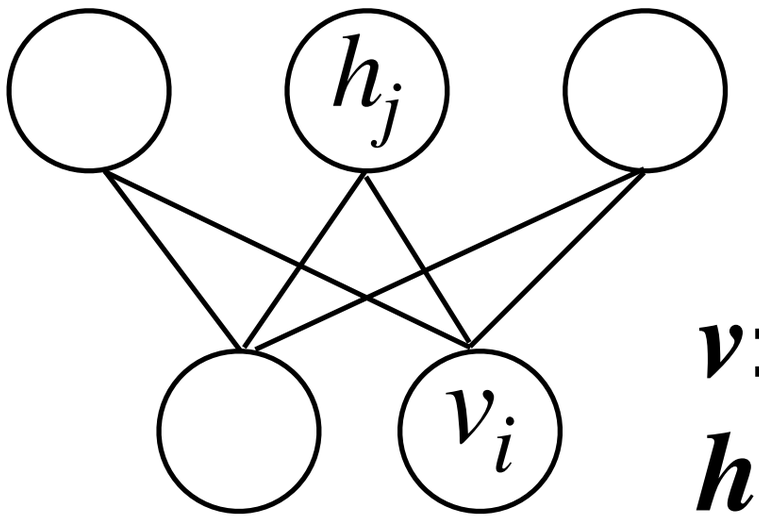
Social Network Analysis



Natural Language Processing: POS Tagging

Pron	Verb	DT	—	Noun / Verb	?
This	is	a		book	

Restricted Boltzmann Machine



$\boldsymbol{v}$: visible nodes (e.g., images)

$\boldsymbol{h}$: hidden nodes (latent features)

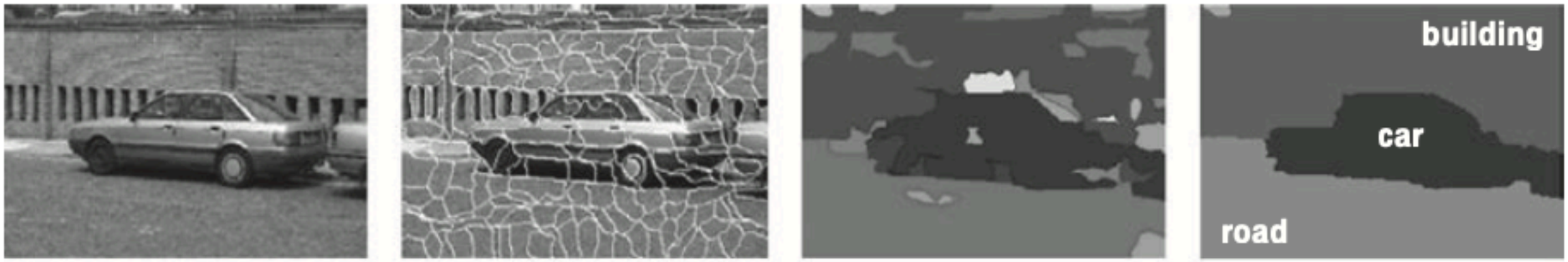
- A Boltzmann distribution is defined on $\boldsymbol{v}, \boldsymbol{h}$
- MLE training or its variants
- Used for pre-training neural networks

Ref: https://www.cs.toronto.edu/~hinton/csc2535/notes/lec4new.pdf

Applications

Computer Vision

[K&F, P.114]



MN is good, but joint distribution is too complicated

MN models $P(X, Y)$, but in applications, queries are all about $P(Y \mid X)$. Therefore, we need to develop a conditional version of MN.

Discriminative vs Generative Models

Consider a data sample, where variables are grouped into two sets $\boldsymbol{X}, \boldsymbol{Y}$

- Generative model: $P(\boldsymbol{X}, \boldsymbol{Y})$
  - It doesn't matter how we group $\boldsymbol{X}$ and $\boldsymbol{Y}$
- Conditional model: $P(\boldsymbol{Y} \mid \boldsymbol{X})$
  - $\boldsymbol{X}$ is given during both training and inference
  - $\boldsymbol{Y}$ is given during training, but is the prediction during inference
  - E.g., POS tagging where the sentence is given

    Image segmentation where the image is given

Conditional Random Fields

- Consider the factors $\phi_k(D_k)$. This time we can ignore all factors that only contains $\boldsymbol{X}$.
- Unnormalized measure $\tilde{P}(\boldsymbol{X}, \boldsymbol{Y}) = \prod_{k=1}^{K} \phi_k(\boldsymbol{X}, \boldsymbol{Y})$
- Partition function $\tilde{Z}_{\boldsymbol{X}} = \sum_{\boldsymbol{y}'} \prod_{k=1}^{K} \phi_k(\boldsymbol{X}, \boldsymbol{y}')$

  Here, the partition function is defined for each sample
- Conditional probability $P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z_{\boldsymbol{X}}} \tilde{P}(\boldsymbol{X}, \boldsymbol{Y})$

HW: Show that CRF is simply the conditional distribution induced by an MRF's joint distribution, where all the factors involving $\boldsymbol{Y}$ are exactly the same in CRF and MRF, but MRF may contain additional factors involving $\boldsymbol{X}$ only.

Applications of CRF

- Softmax regression

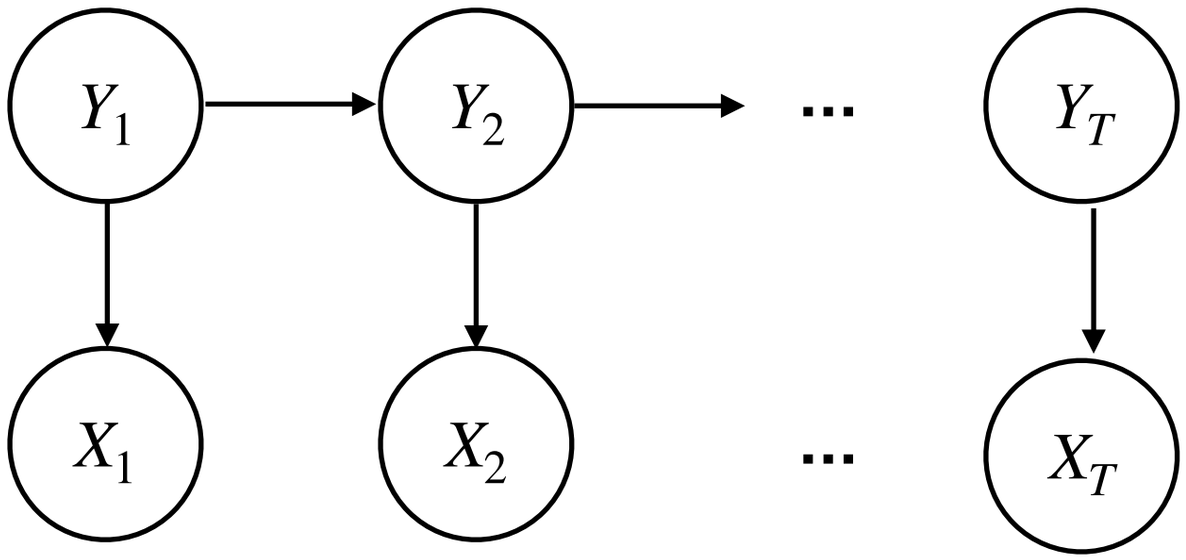
$$\Pr[Y = k | \boldsymbol{X} = \boldsymbol{x}] = \frac{\exp\{\boldsymbol{w}_k^\top \boldsymbol{x}_i\}}{\sum_{k'} \exp\{\boldsymbol{w}_{k'}^\top \boldsymbol{x}\}}$$

Identify the features and draw the factor graph with conditioned variables grayed out.

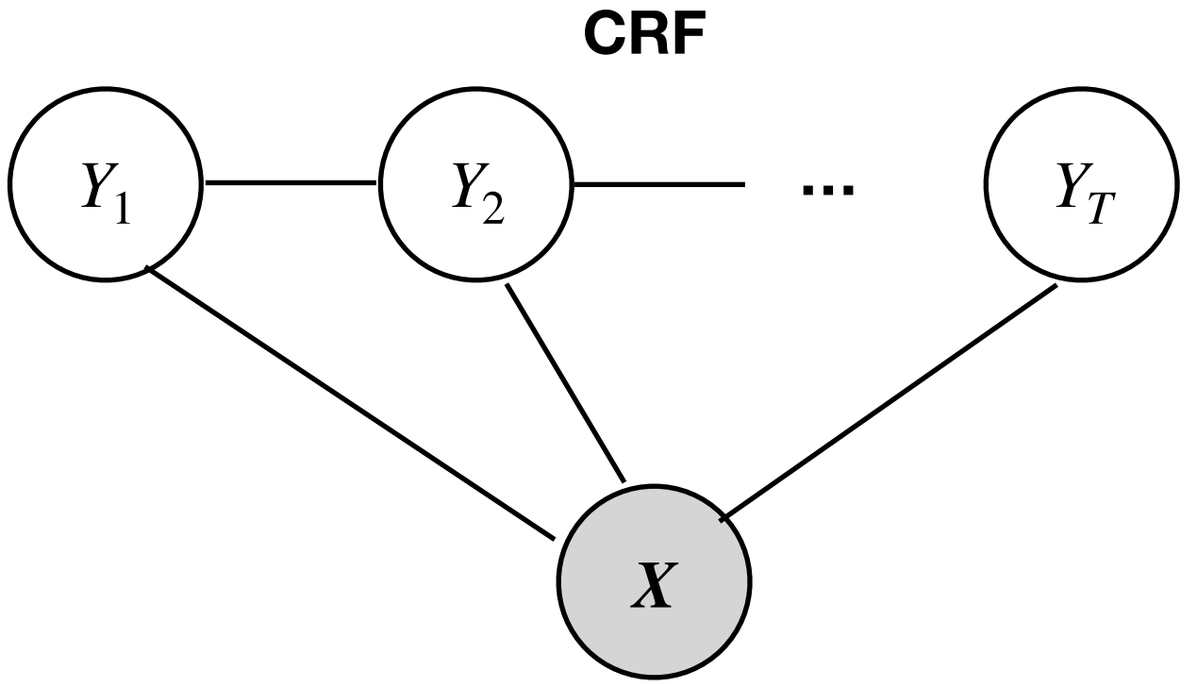
Applications of CRF

- POS tagging

HMM



CRF

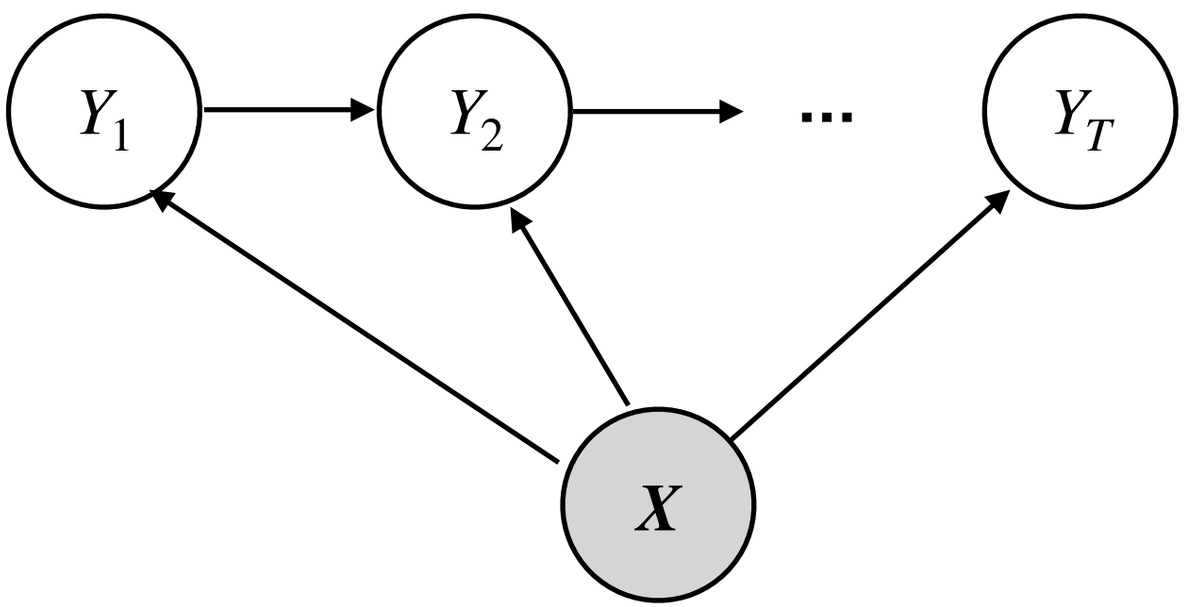


In CRF, the arrows from $\boldsymbol{X}$ to $\boldsymbol{Y}$ can be drawn either directed or undirected

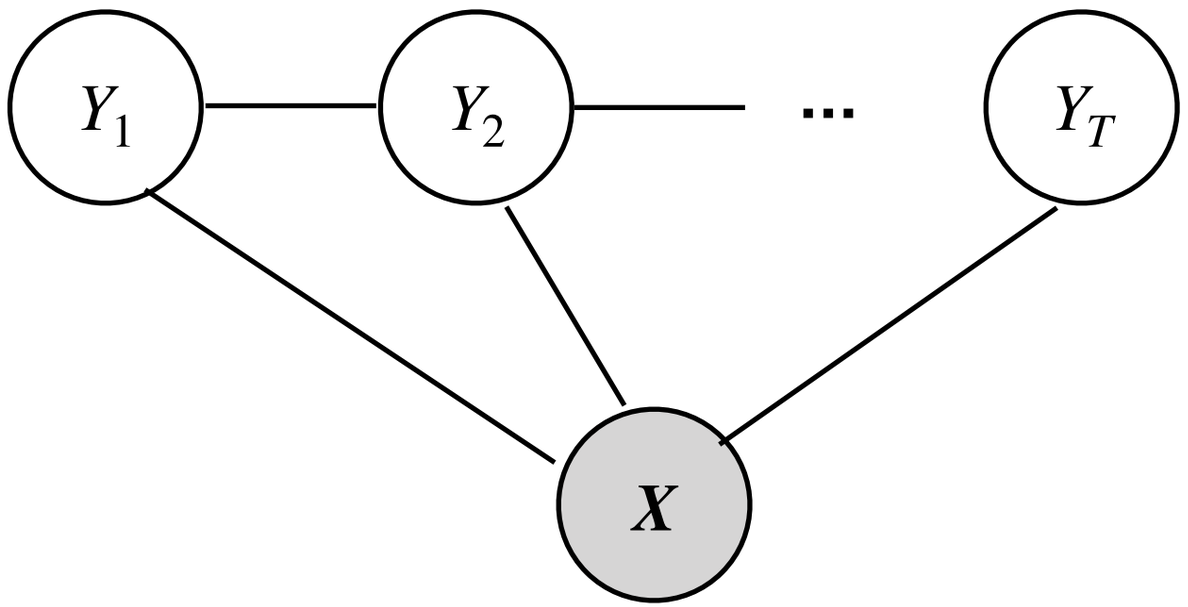
- HMM models the joint probability $P(\boldsymbol{X}, \boldsymbol{Y})$. However, the treatment of $\boldsymbol{X}$ is naïve, and it may hurt the performance. Recall naïve Bayes.
- CRF models the conditional probability $P(\boldsymbol{Y} \mid \boldsymbol{X})$. It does not model $P(\boldsymbol{X})$, which is irrelevant to $P(\boldsymbol{Y} \mid \boldsymbol{X})$. It also provides a flexible way of conditioning $\boldsymbol{Y}$ on $\boldsymbol{X}$ based on log-linear formulation.

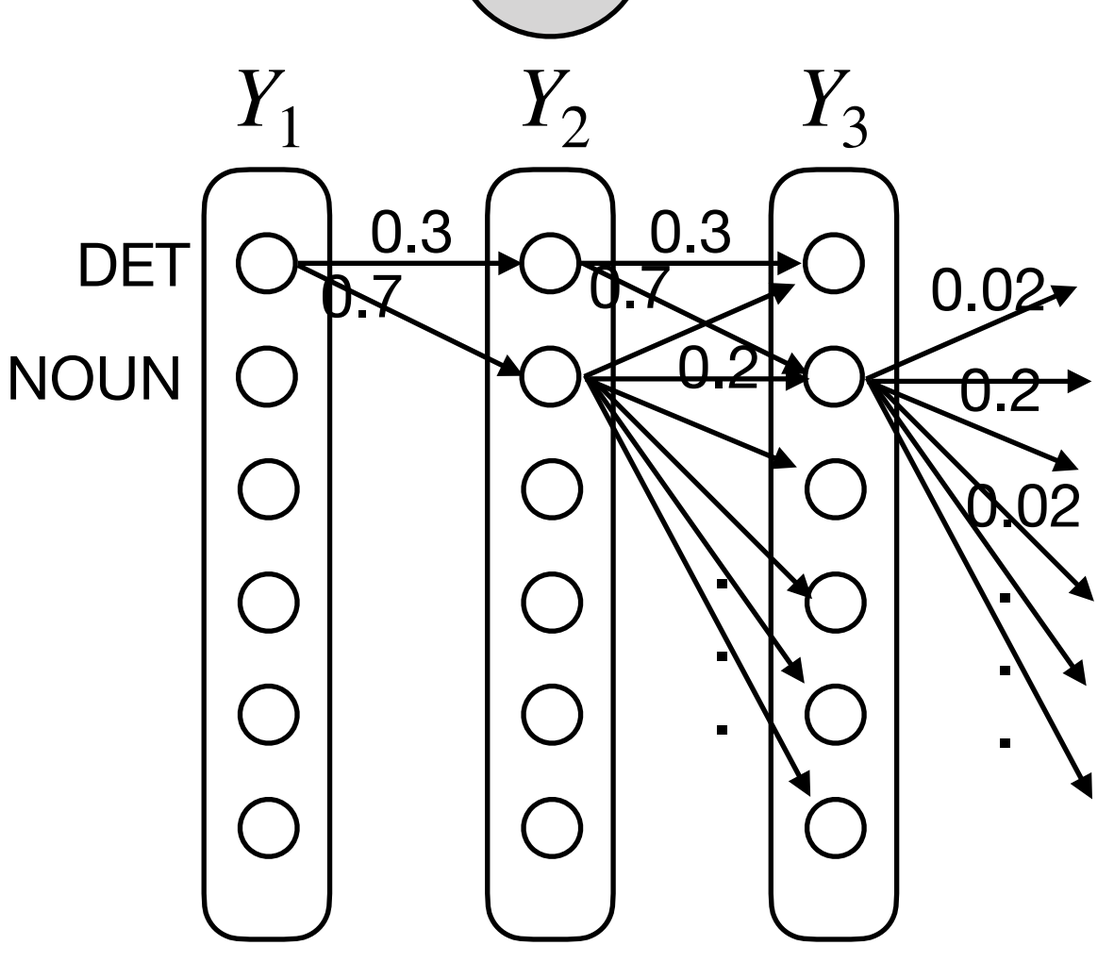
Label Bias

Max-Entropy MM (MEMM)



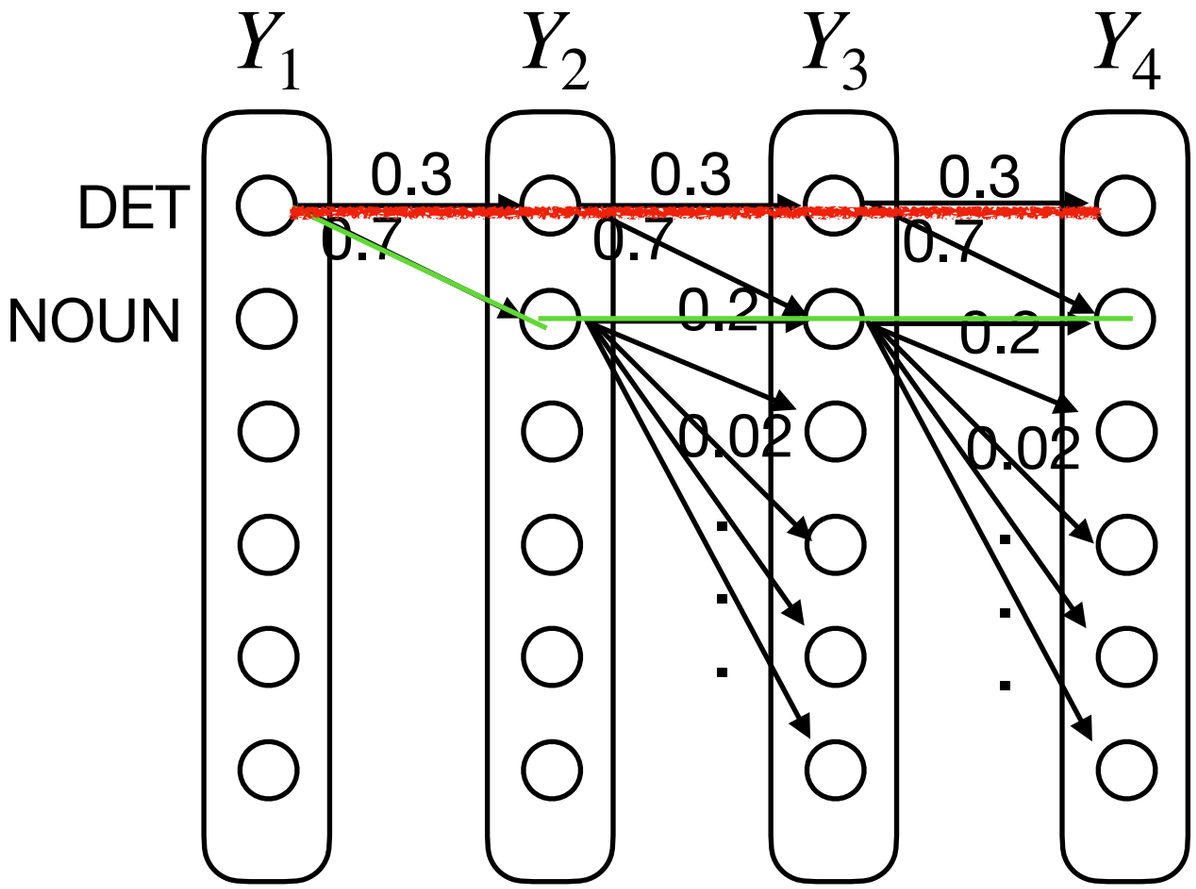
CRF





DET DET: the both

Label Bias

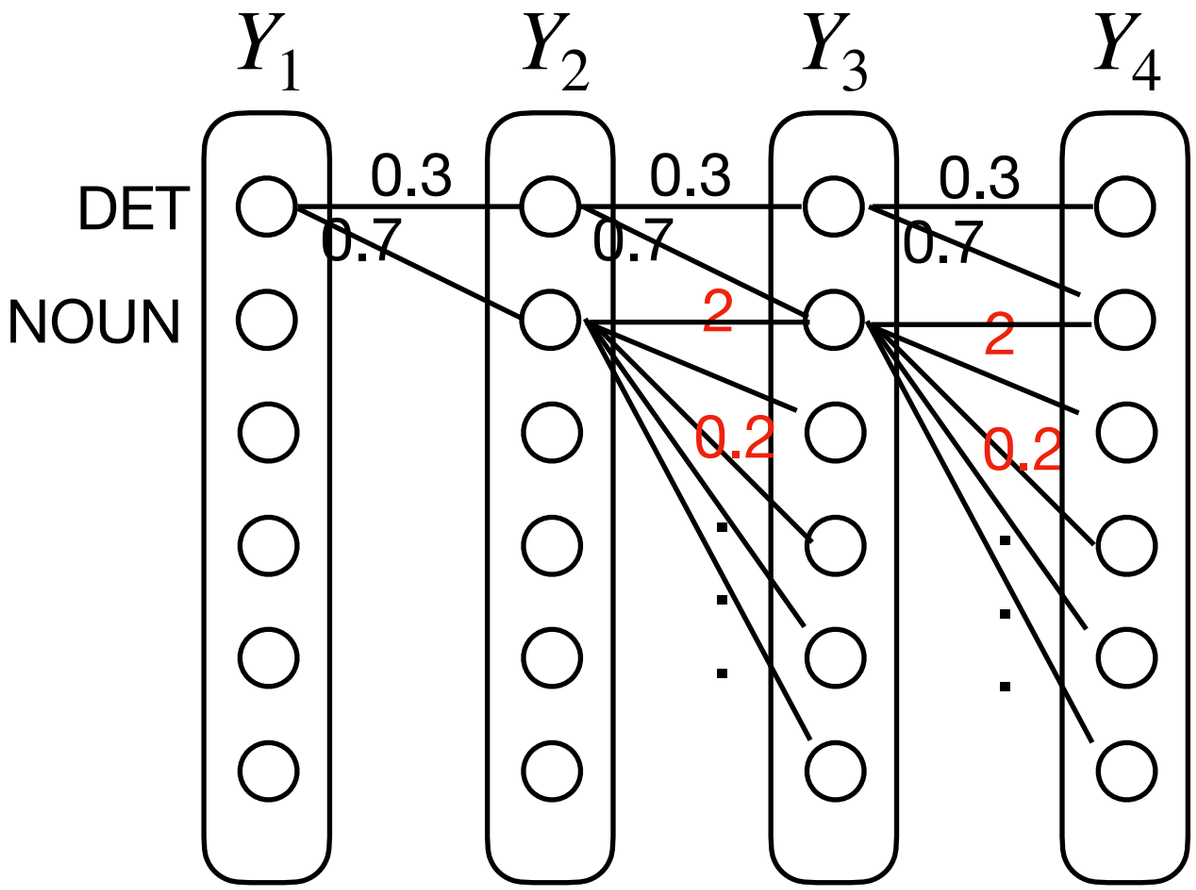


DET DET: the both

Locally speaking: DET NOUN NOUN (w.p. 0.0056) is more reasonable

Globally speaking: DET DET DET (w.p. 0.0081) is the global optimum

Label Bias



DET DET: the both

Locally speaking: DET NOUN NOUN (w.p. 0.0056) is more reasonable

Globally speaking: DET DET DET (w.p. 0.0081) is the global optimum

Failure to Correct Previous Predictions

- Consider your model is partially observable: ith step's prediction only based on previous predictions and observations

$$P(Y_i \mid \boldsymbol{Y}_{<i}, \boldsymbol{X}_{<i})$$

- Suppose the observation at X_t makes none of the labels for Y_t happy, BN cannot zero out the probability due to its local normalization, but MN can.

Summary

- Markov Random Field/Markov Network:
  - Local factors with certain scopes
  - Unnormalized measure, partition function, probability
- Conditional random field: A conditional version of MRF
- Label bias problem