PROBABILISTIC GRAPHICAL MODELS

Justin Domke
National ICT Australia

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

Ground Rules

- *Few proofs*
- *Too much material*

Ground Rules

- *Please interrupt and ask questions.*
- *Including questions about notation.*

What is a graphical model?

- Informal definition:
  - A graphical model is a probability distribution written in a factorized form.
  - For example:

$$p(x) \propto \psi(x_1, x_2)\psi(x_2, x_3)\psi(x_3, x_4)$$

Why probabilistic graphical models?

- Better question: why model-based machine learning?
- Image recognition. E.g. given $x \in \mathbb{R}^{3,000,000}$, want to predict $y \in \{\text{cat}, \text{dog}, \text{frisbee}\}$.

  - Neural networks: Huge, generic nonlinear function f from x to y.
  - Model-based learning. Use domain knowledge to specify f.
  - Probabilistic learning. Use domain knowledge to specify probability distribution $p(y|x)$ or $p(y, x)$.

Why probabilistic graphical models?

- Neural Network
  - Pro: Very powerful class of functions.
  - Con: May need a lot of data to learn.

- Model-Based approach
  - Pro: Fewer parameters– can learn with fewer data.
  - Con: Results only as good as the model you start with.

- Probabilistic model approach.
  - Pro: Can answer general questions. e.g. $P(y = \text{cat} | x, y \neq \text{dog})$ or sample from $p(x|y = \text{dog})$
- In reality, cannot neatly separate these approaches.

The curse of dimensionality

- Modern machine learning is usually concerned with high-dimensional objects.
- Consider learning a distribution over $x \in \{0, 1\}^N$.
- If N=100, $p(x)$ has 1267650600228229401496703205375 free parameters.
- If we are to have any hope, we must assume some kind of special structure for $p(x)$.

Conditional independence

- The special structure graphical models assume is conditional independence:

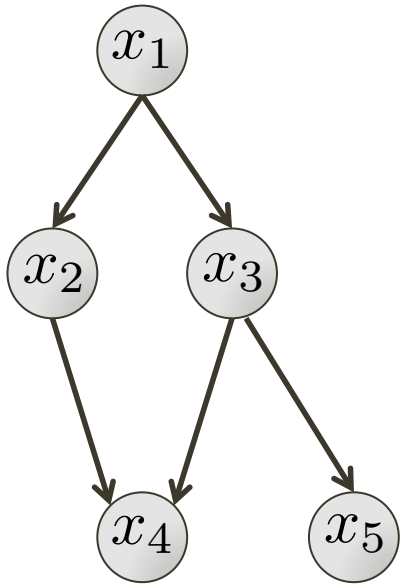
- If you want to guess the value of some variable x_i, then once you know the values of some "neighboring" variables $x_{\mathcal{N}(i)}$, then you get no additional benefit from knowing all other variables.

- Turns out, this leads to factorized distributions.

Directed graphical models

- Represent joint distribution as a product of local conditionals. e.g.

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3)$$



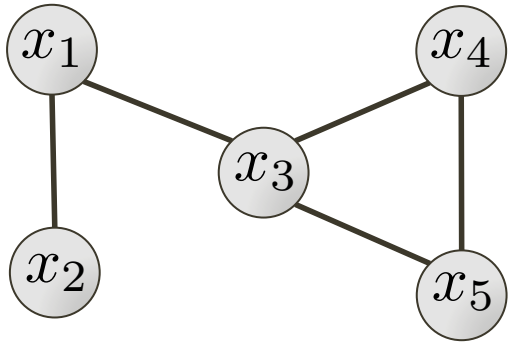
- Also known as “Bayes net”.

Undirected graphical models

- Represent graphical models as a product of functions on cliques. (AKA “Markov Random Field”

$$p(x) = \frac{1}{Z} f(x_1, x_2) f(x_1, x_3) f(x_3, x_4, x_5)$$

$$Z = \sum_{x_1, \ldots, x_5} f(x_1, x_2) f(x_1, x_3) f(x_3, x_4, x_5)$$

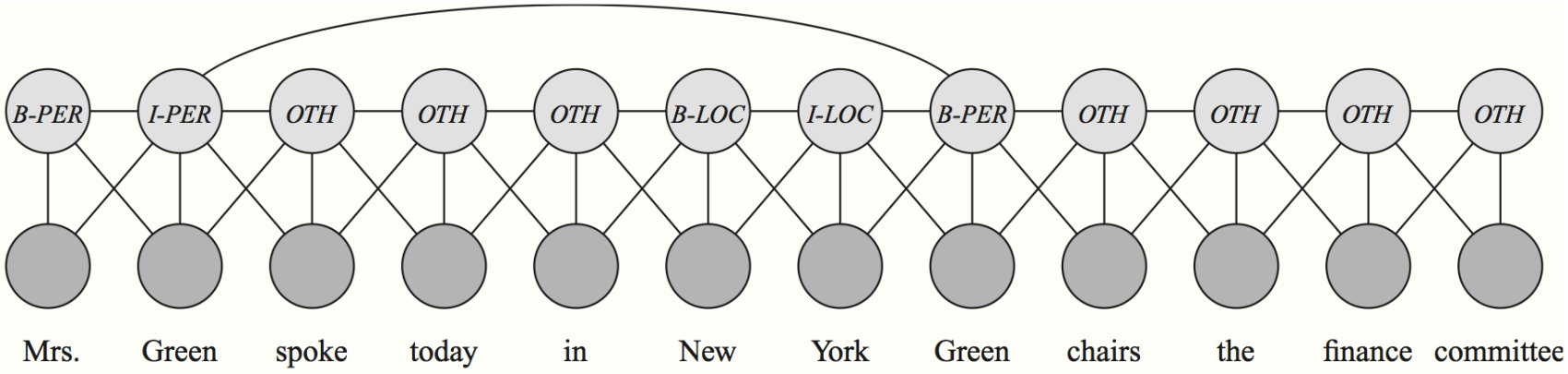


- Local functions have no direct probabilistic interpretation.

- Z ensures normalization, and has many interesting and sometimes troublesome properties.

Examples of Graphical Models

- Named-entity recognition.



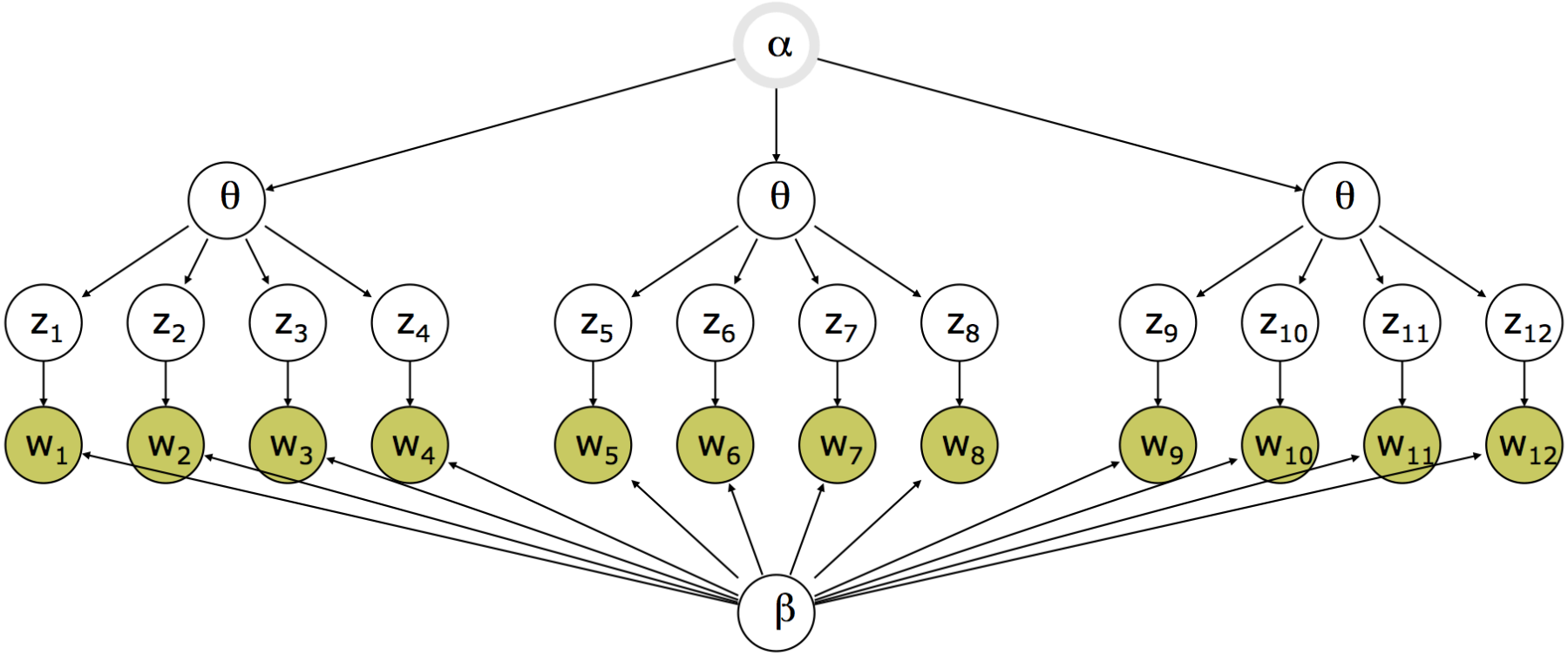
KEY

B-PER	Begin person name	*I-LOC*	Within location name
I-PER	Within person name	*OTH*	Not an entitiy
B-LOC	Begin location name		

Credit: Koller and Friedman “Probabilistic Graphical Models”

Examples of Graphical Models

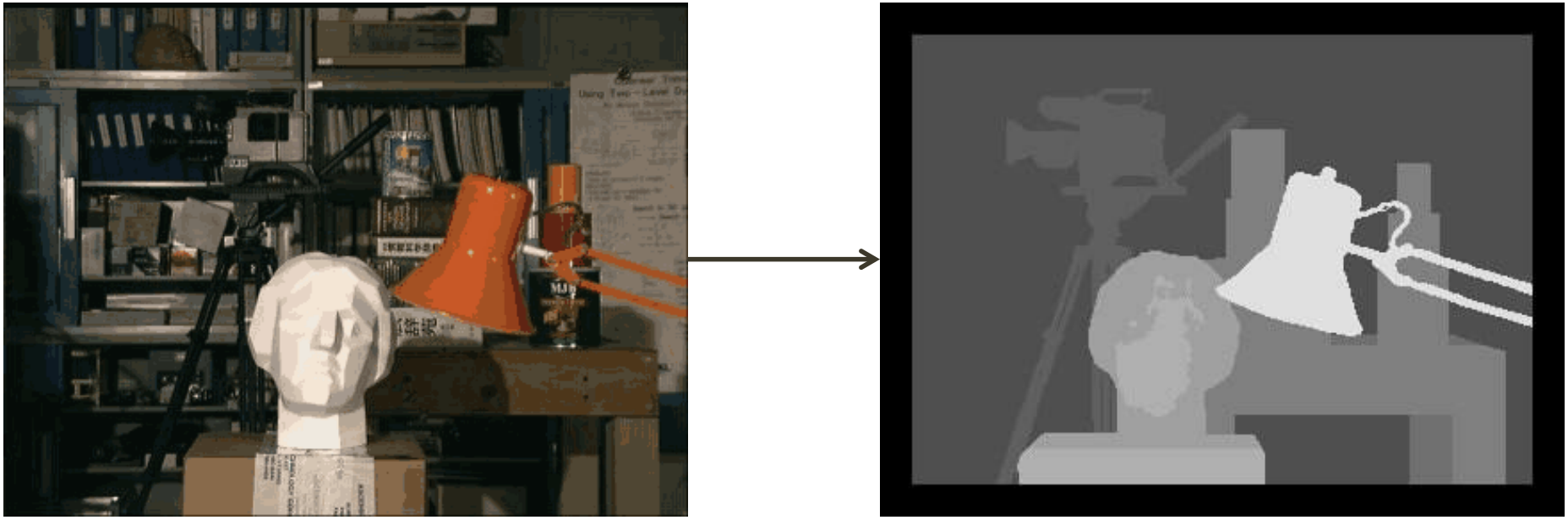
- Topic models



Credit: Andrew McCallum

Examples of Graphical Models

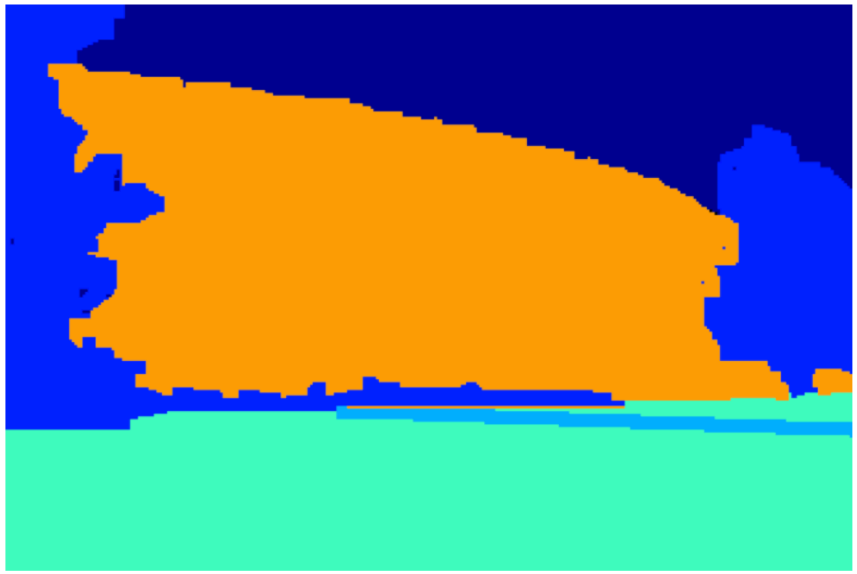
- Stereo Vision



Credit: Middlebury Computer Vision

Examples of Graphical Models

- Semantic Segmentation



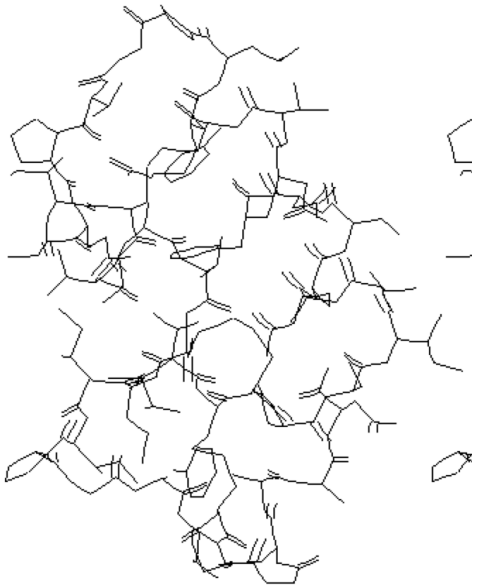
Examples of Graphical Models

- Protein side-chain prediction:

- Amino acids x, discrete sequence of angles y.

$$p(y|x) \propto \exp(\frac{1}{T}E(x,y,\theta)) \qquad E(x,y,\theta) = \sum_{j=1}^{D} \theta_j E_j(x,y)$$

- Different E_j model different:
  - Electrostatic charges
  - Hydrogen bonding potentials
  - Etc.



Credit: Steffen Schulze-Kremer

Summary

- There are tradeoffs between model-based and non model-based learning.
- It is not practical to learn distributions over high-dimensional spaces without simplifying assumptions.
- Conditional independence assumptions lead to factorized graphical models.
- There are various interesting applications.

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

Conditional Independence

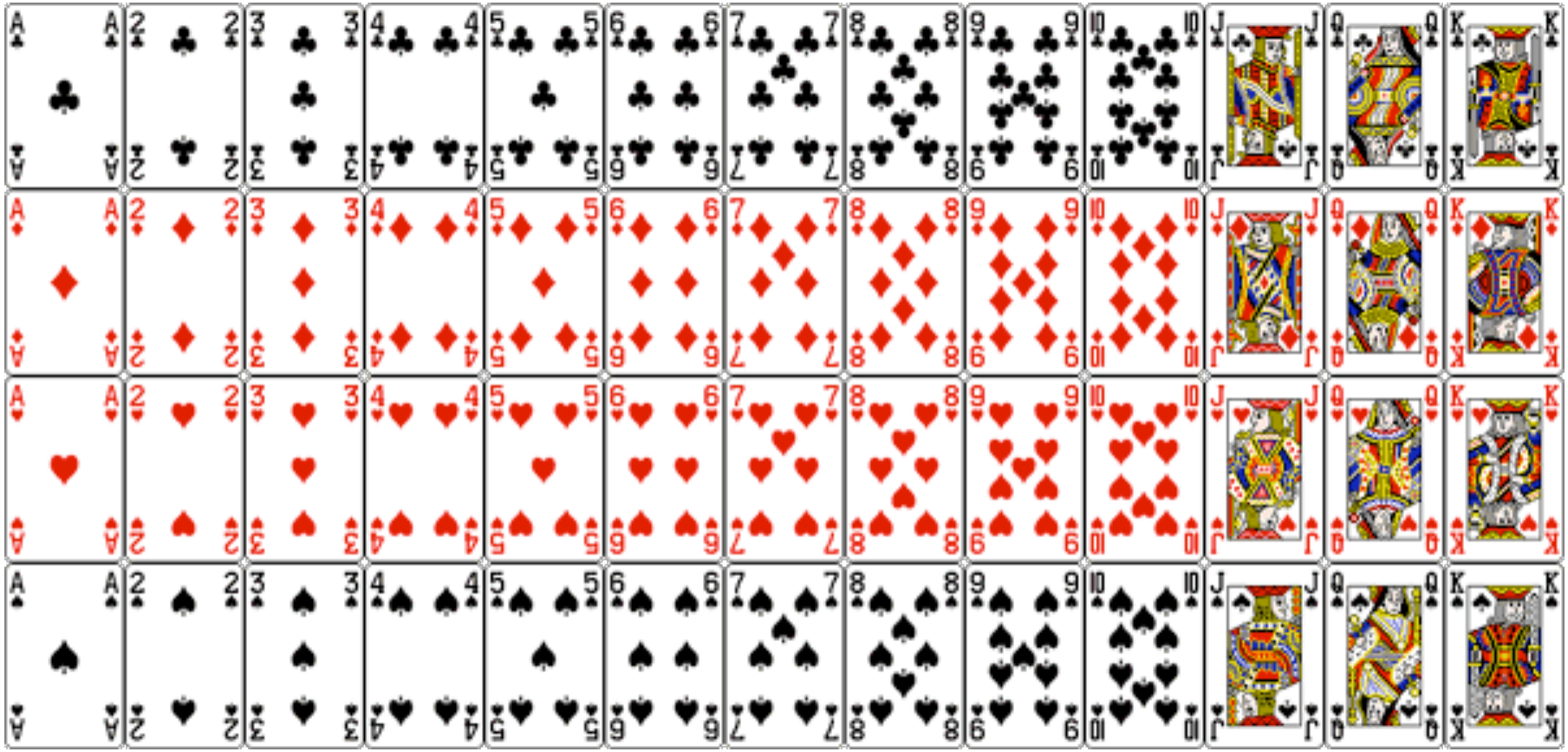
- X is independent of Y if "knowing Y doesn't help you to guess X".

$$X \perp Y \leftrightarrow \mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$$

- X is conditionally independent of Y given Z if "once you know Z, knowing Y doesn't help you to guess X".

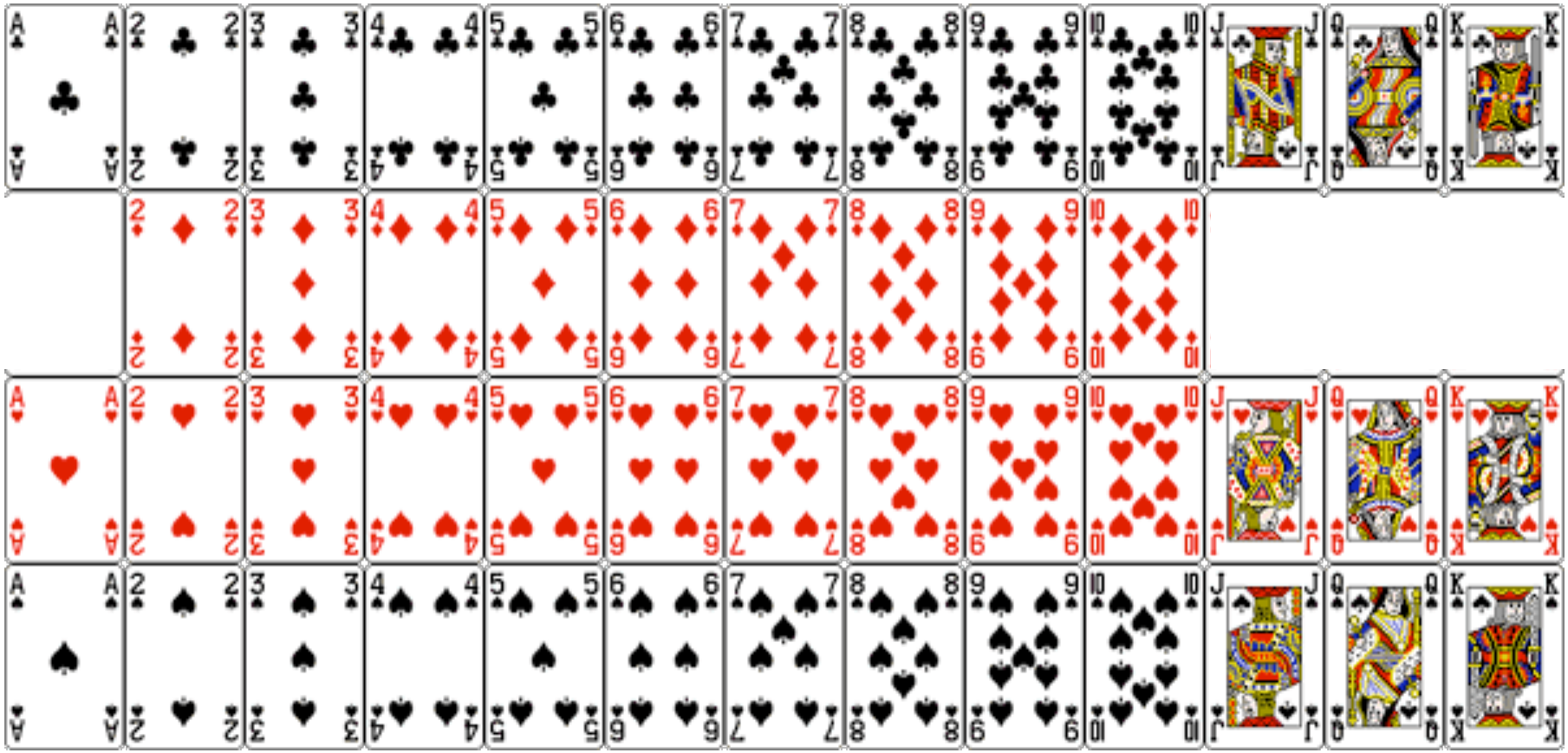
$$X \perp Y|Z \leftrightarrow \mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)$$

Conditional Independence



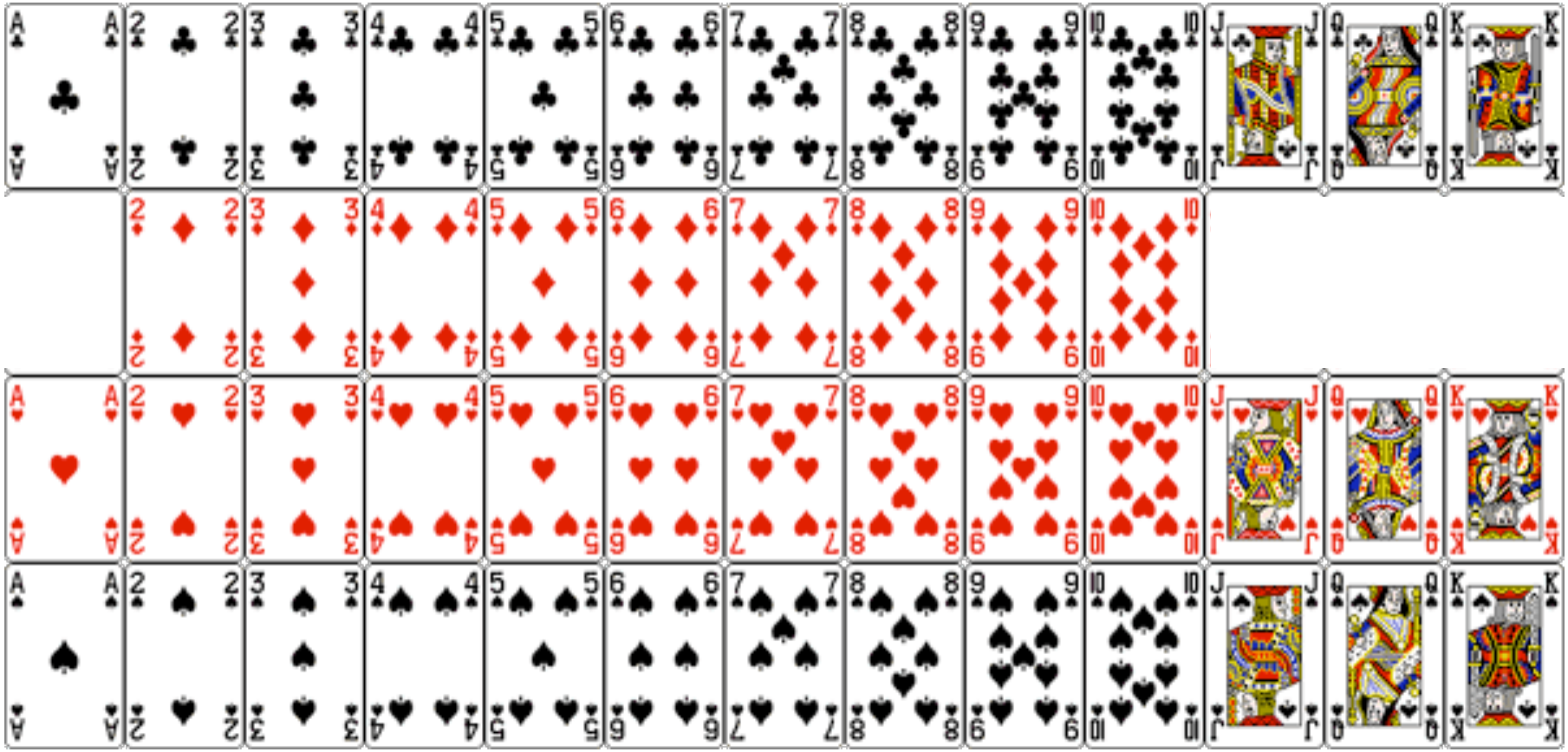
Value $\perp$ Color

Conditional Independence



Value $\not\perp$ Color

Conditional Independence

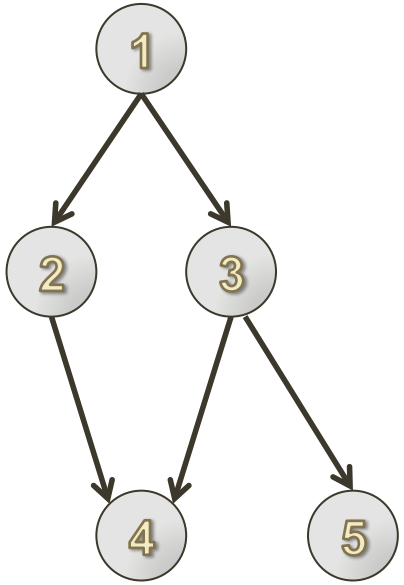


$$\text{Value} \perp \text{Color} \mid \text{Facecard}$$

Directed Acyclic Graph

- A DAG is determined by a set of "parent" nodes $\pi(i)$ for each node i. W.O.L.O.G. assume that $\forall j \in \pi(i), j < i$.

$$\pi(1) = \{\}$$
$$\pi(2) = \{1\}$$
$$\pi(3) = \{1\}$$
$$\pi(4) = \{2, 3\}$$
$$\pi(5) = \{3\}$$



Directed Model

- In a directed model, we assume:

$$X_i \perp X_1, ..., X_{i-1} | X_{\pi(i)}$$

- Or, equivalently

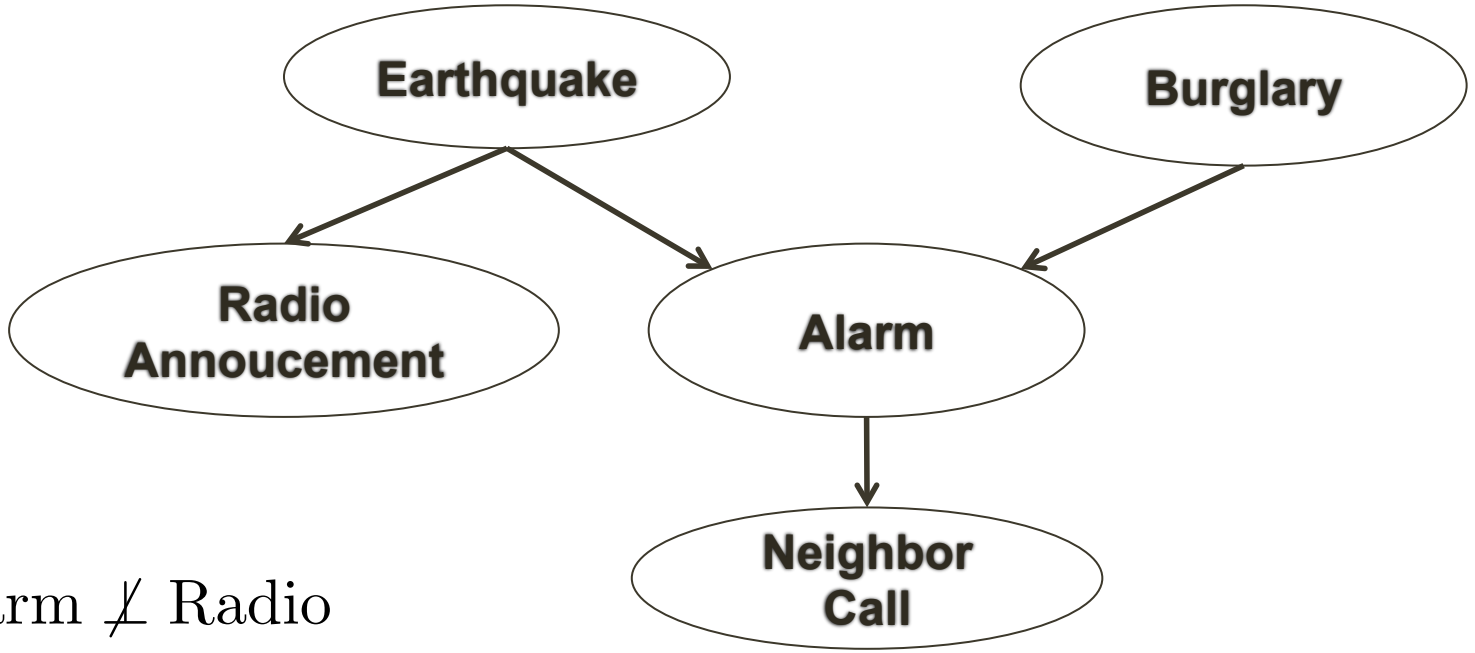
$$P(x_i | x_1, ..., x_{i-1}) = P(x_i | x_{\pi(i)})$$

- Then, we can write that

$$\begin{aligned} p(x) &= p(x_1)p(x_2|x_1)...p(x_n|x_1, ..., x_{n-1}) \\ &= \prod_i p(x_i|x_1, ..., x_{i-1}) \\ &= \prod_i p(x_i|x_{\pi(i)}) \end{aligned}$$

- So, this conditional independence assumption leads to this factorized representation.

Alarm Network



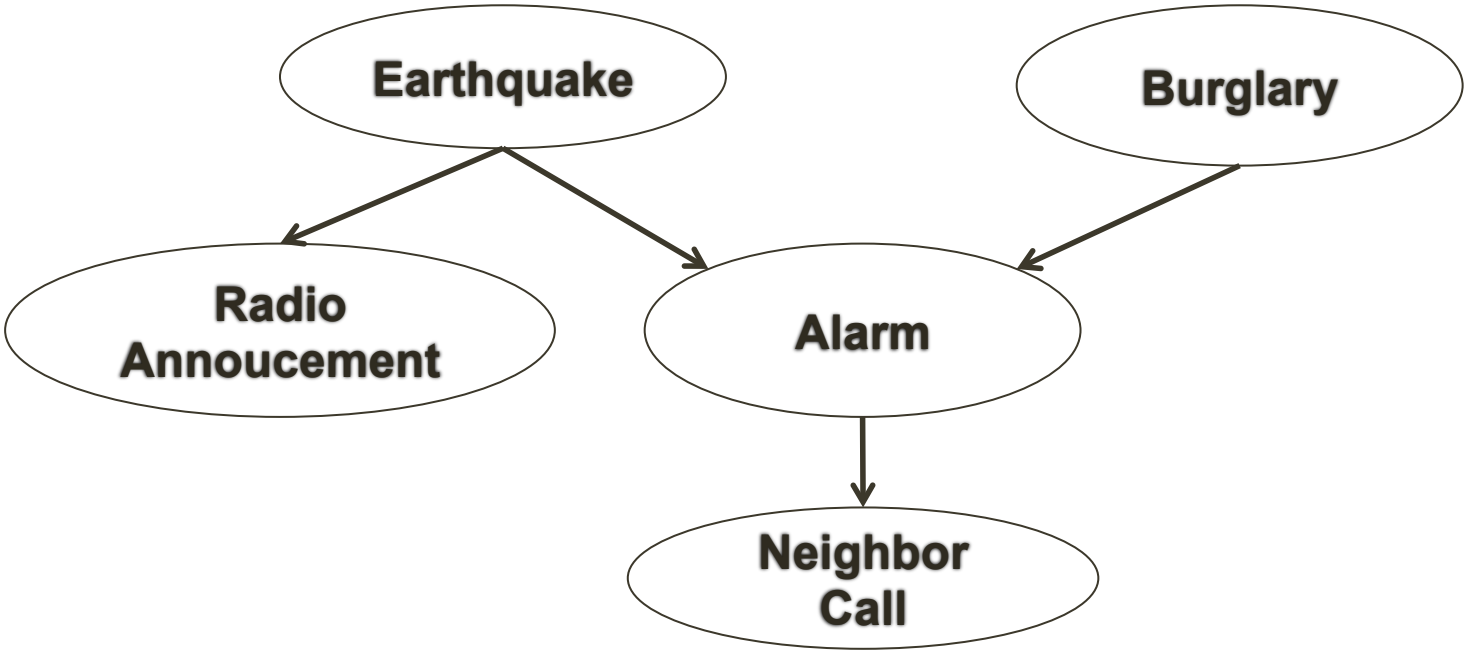
$\text{Alarm} \not\perp \text{Radio}$

$\text{Alarm} \perp \text{Radio} \mid \text{Earthquake}$

$\text{Earthquake} \perp \text{Burglary}$

$\text{Earthquake} \not\perp \text{Burglary} \mid \text{Alarm}$

Alarm Network



Markov Blanket for Radio? $\{\text{Earthquake}\}$

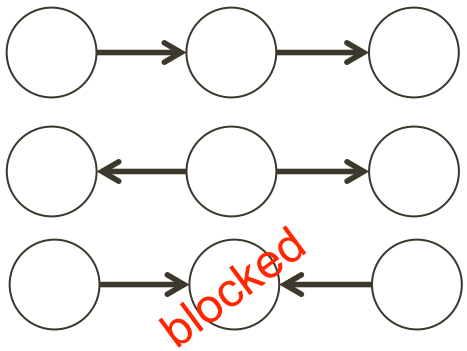
For Neighbor? $\{\text{Alarm}\}$

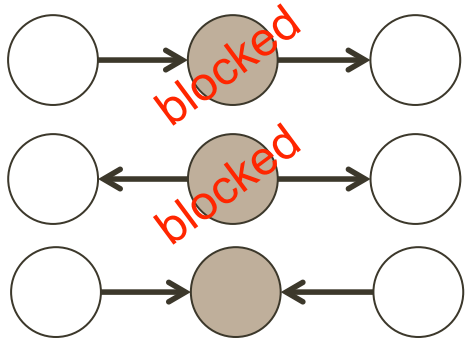
For Burglary? $\{\text{Alarm}, \text{Earthquake}\}$

For Alarm ? $\{\text{Earthquake}, \text{Burglary}, \text{Neighbor}\}$

D-separation

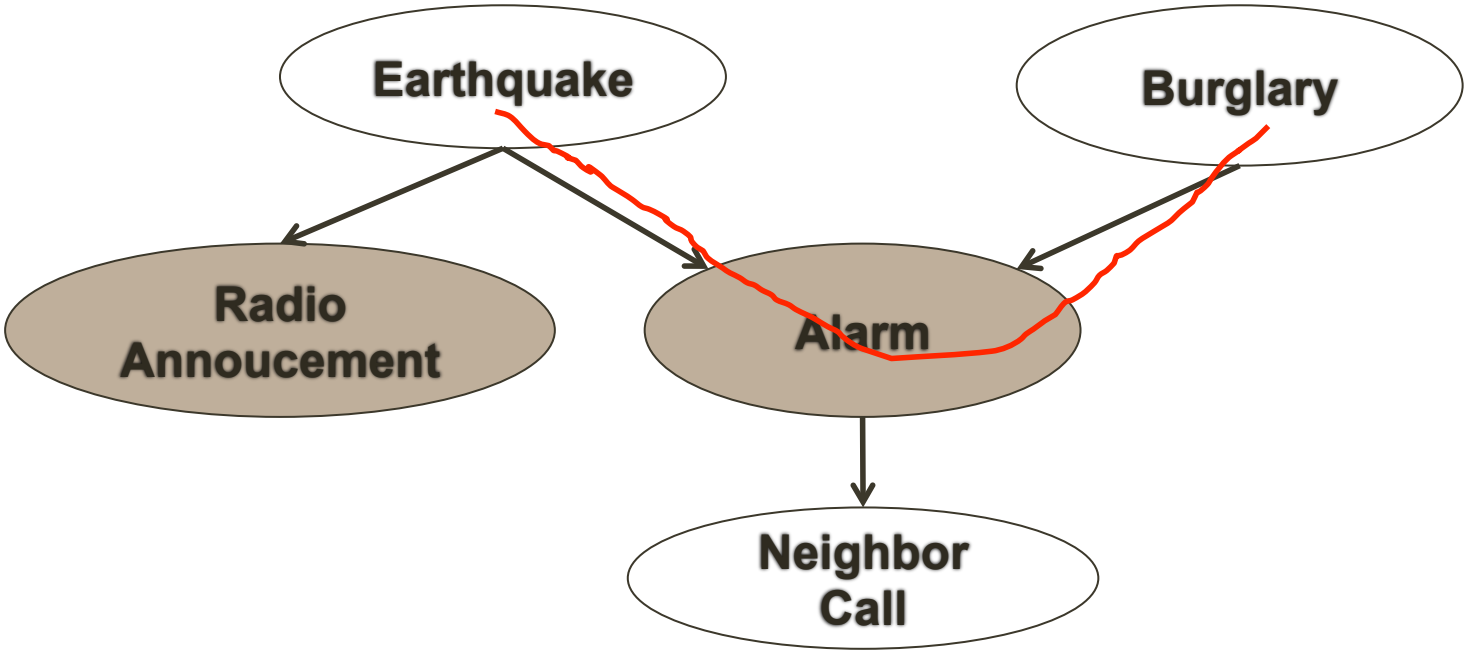
- Given a directed model, how to tell if some variable X_i is conditionally independent of some variable X_j given X_A?
- Take the graph, color all nodes A, start a ball at i.
- Bounce the ball around the graph with the following rules:



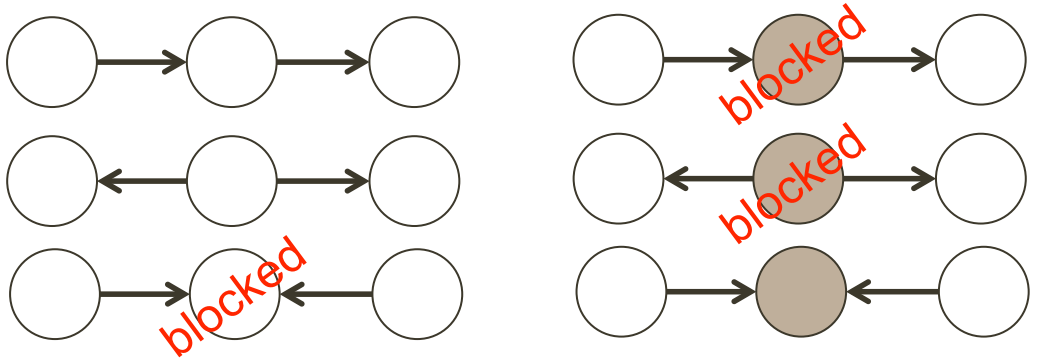


- If you can hit j, they are not conditionally independent.
- If it is impossible to hit j, they are conditionally independent.

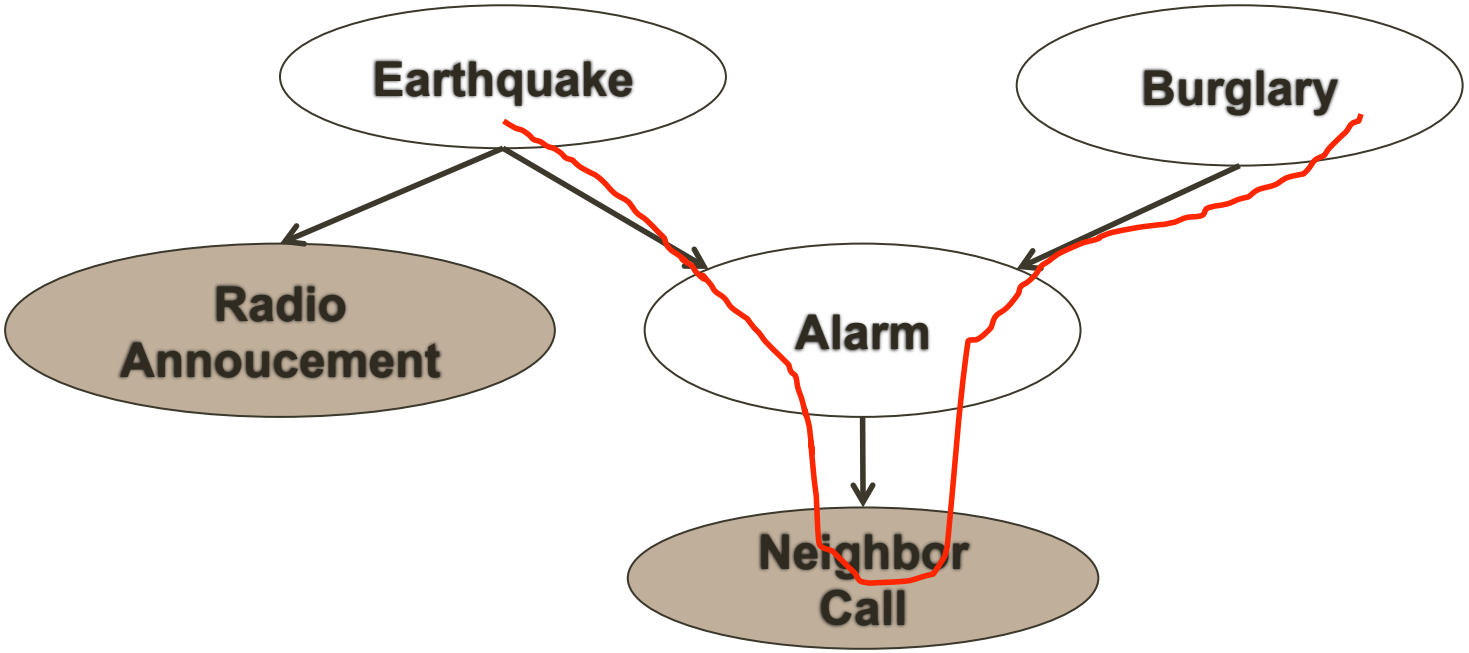
Alarm Network



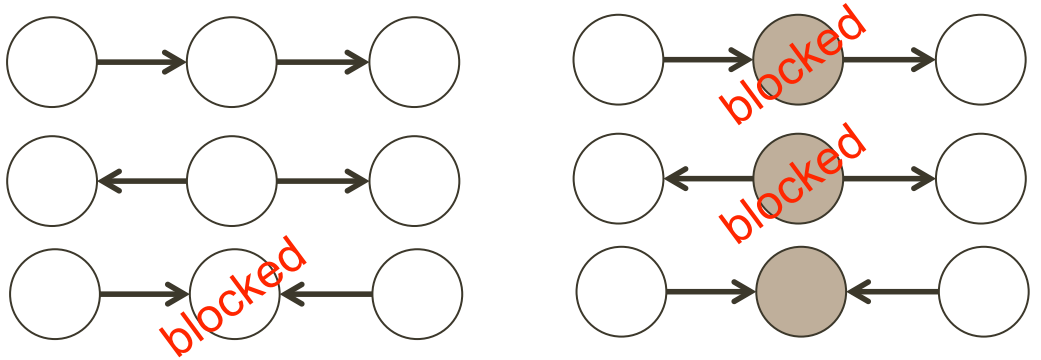
$\text{Earthquake} \perp \text{Burglary} \mid \text{Radio}, \text{ Alarm}?$ No!



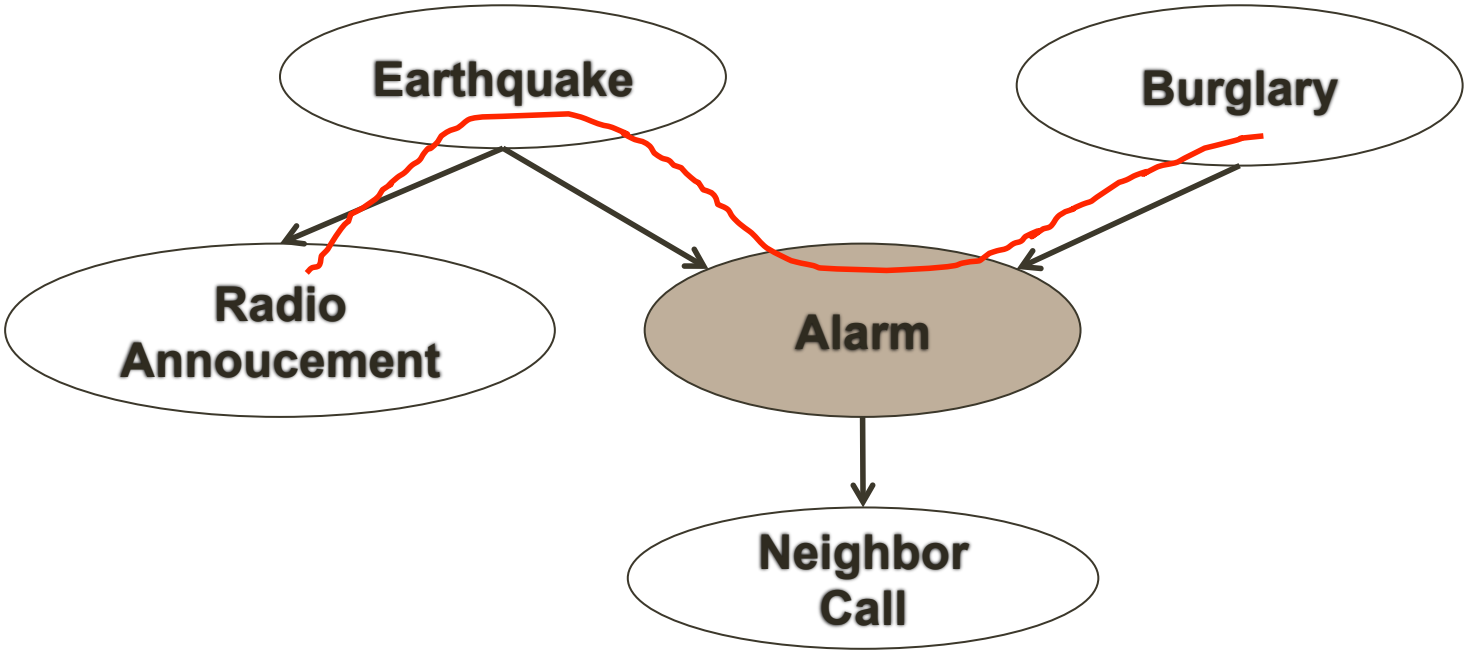
Alarm Network



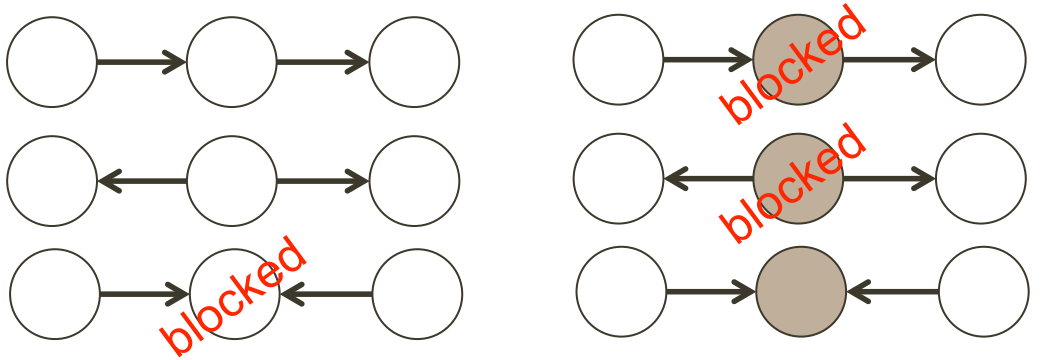
$\text{Earthquake} \perp \text{Burglary} \mid \text{Radio}, \text{ Neighbor}?$ No!



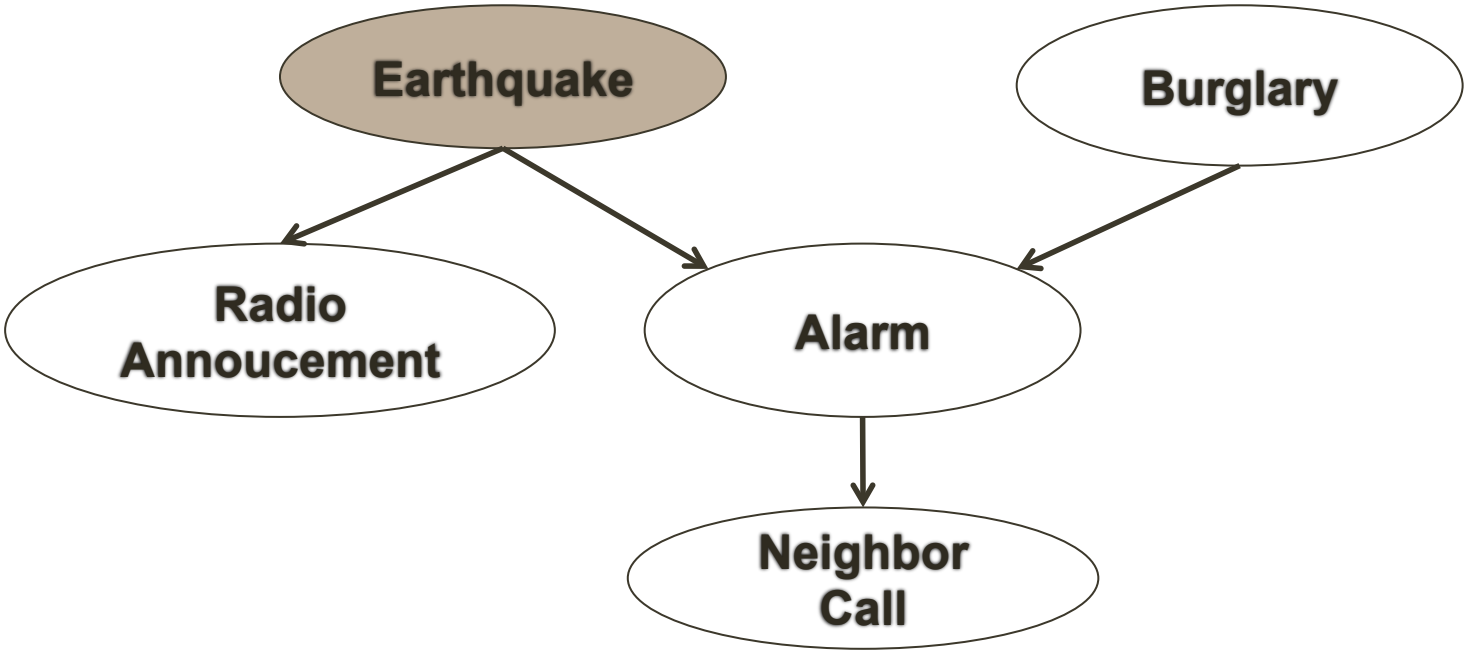
Alarm Network



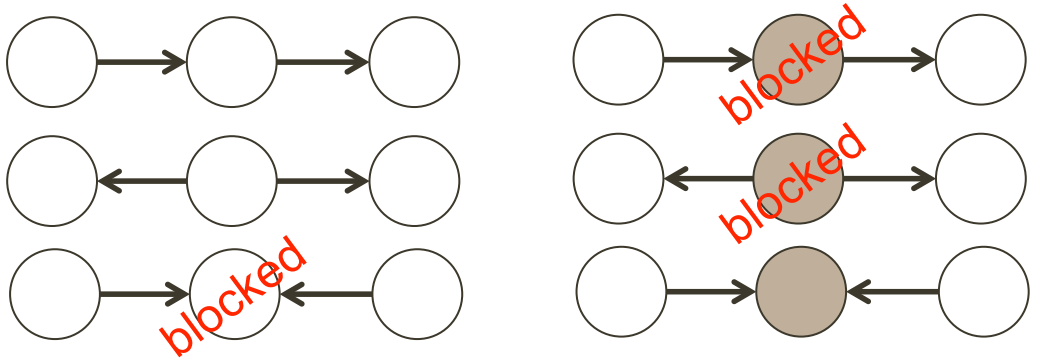
$\text{Radio} \perp \text{Burglary} \mid \text{Alarm}?$ No!



Alarm Network

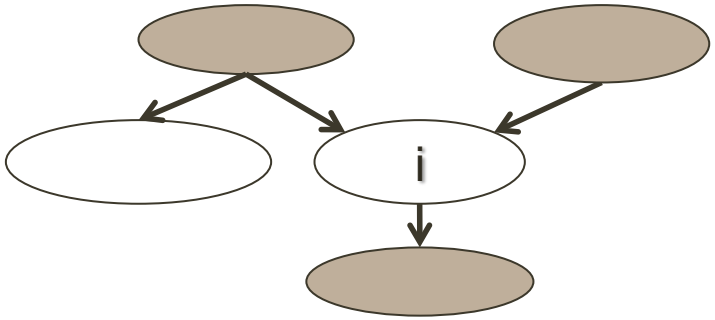


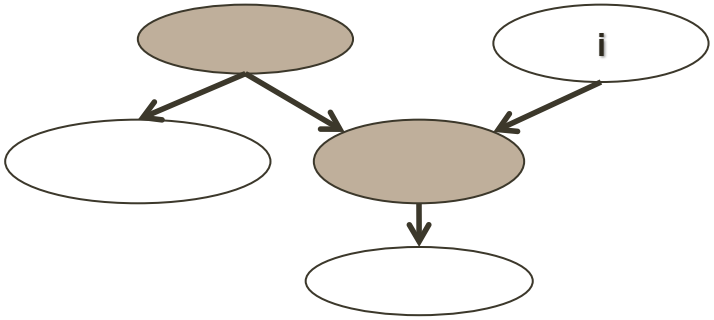
$\text{Radio} \perp \text{Alarm} \mid \text{Earthquake}?$ Yes!

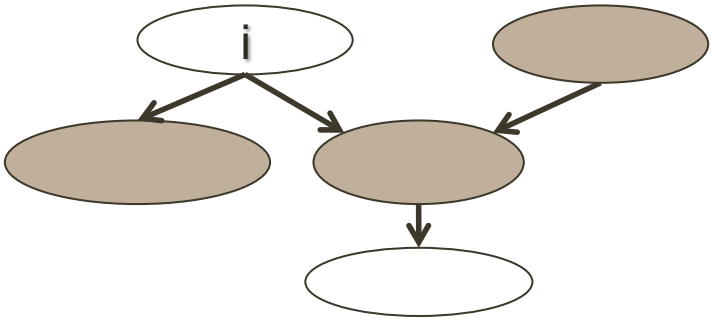


Markov Blankets

- In general, what is the Markov blanket of a given node i?
  - Parents of i.
  - Children of i.
  - Parents of children of i.

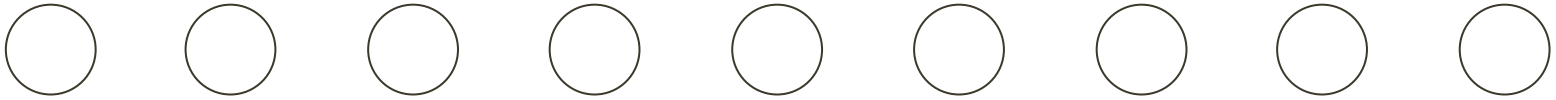




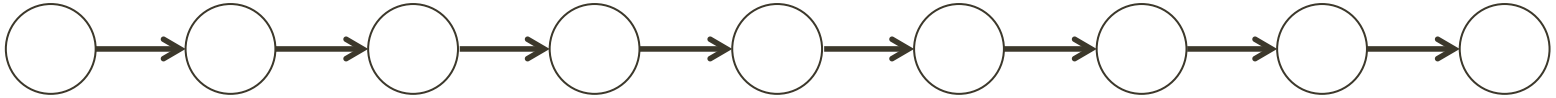


Chain Model

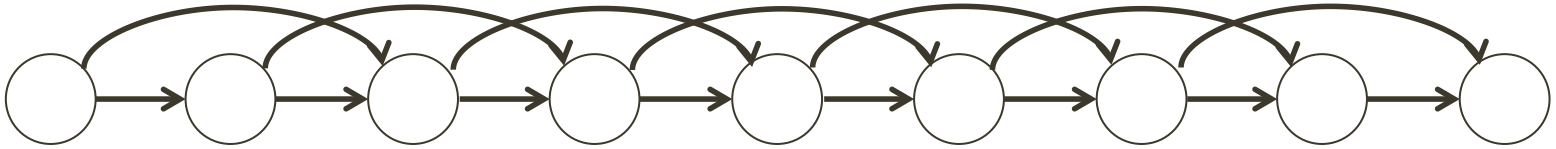
- Shannon's 1948 paper "A Mathematical Theory of Communication" proposed model for language. (letters)



OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.



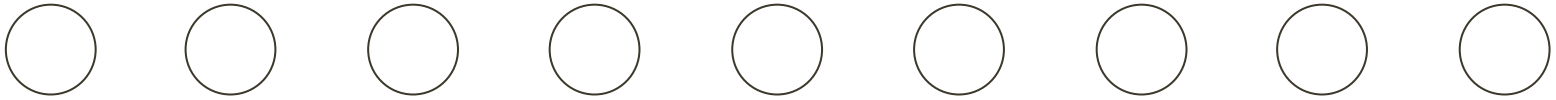
ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.



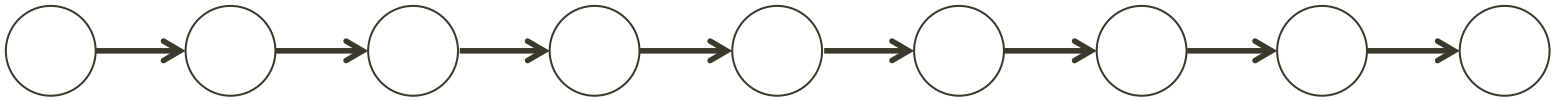
IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

Chain Model

- Shannon's 1948 paper "A Mathematical Theory of Communication" proposed model for language. (words)



REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

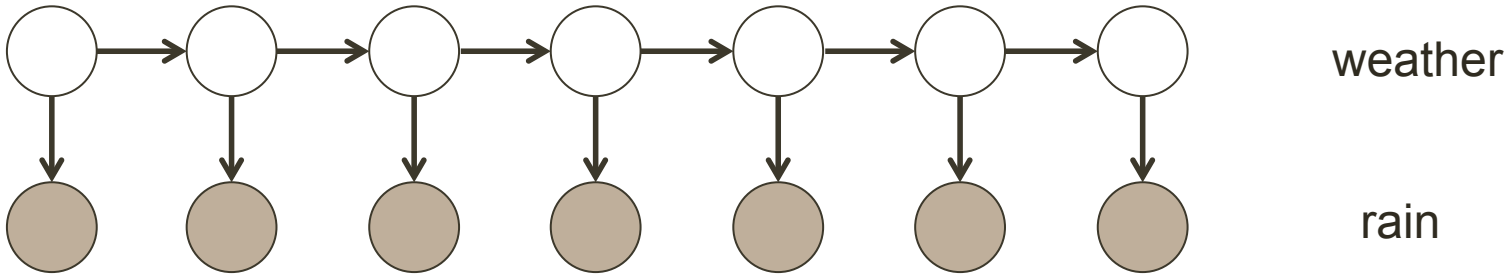


THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Hidden Markov Model

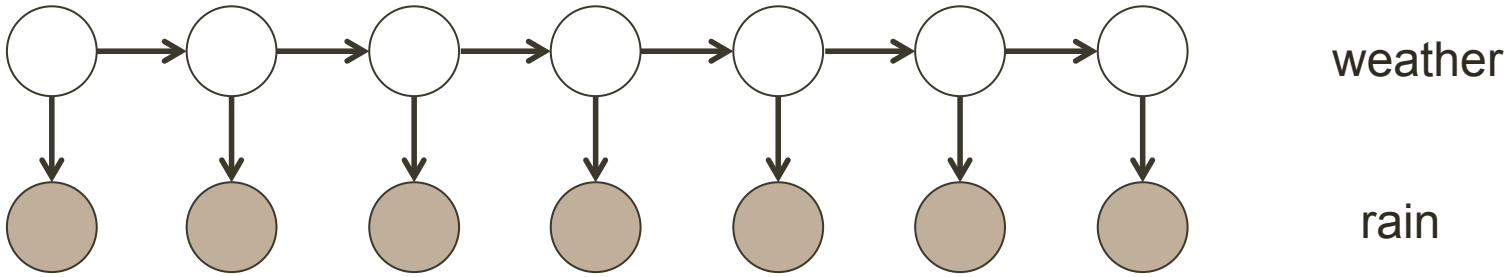
- You are a PhD student, working very (very!) hard on your thesis. You have no time to look at the sky, but only hear if it rains or not.

- If it is cloudy, it is more likely to rain.
  - If it is sunny, it is less likely to rain.
- If it is cloudy, it is more likely to be cloudy the next day.
  - If it is sunny, it is more likely to be sunny the next day.

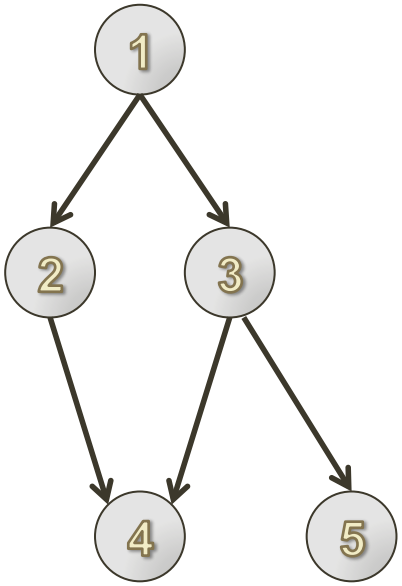


Hidden Markov Models (HMMs)

- Problems to solve:
  - Given a sequence of rain observations (e.g. <rain, rain, nope, rain, nope, nope>), what is the most likely weather sequence? (e.g. <cloudy, cloudy, cloudy, cloudy, sunny, sunny>).
  - Given same, what is probability it was sunny on day 3? (e.g 0.3)
  - Given a series of rain observations only, what are the transition probabilities?



Maximum Likelihood



- Given a fixed graph, how to do learning?
- Natural criterion:

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

- Solution is empirical conditionals!

$$p(X_i = x_i | X_{\pi(i)} = x_{\pi(i)}, \theta) = \frac{\#[X_i = x_i, X_{\pi(i)} = x_{\pi(i)}]}{\#[X_{\pi(i)} = x_{\pi(i)}]}$$

$$p(x_i| = x_{\pi(i)}; \theta) = \hat{p}(x_i|x_{\pi(i)})$$

- Note: Changes with regularization. Not so easy with hidden variables. And how to learn structure?

Summary

- If you assume $P(x_i|x_1, ..., x_{i-1}) = P(x_i|x_{\pi(i)})$, then $p(x) = \prod_i p(x_i|x_{\pi(i)})$. This is a "directed model".
- The Bayes Ball gives (complicated) rules for determining all conditional independencies implied by a DAG.
- With fully observed data, and fixed structure, ML solution is to match empirical conditionals.

- To come:
  - How to do inference
  - How to deal with hidden variables

- But first:
  - A different assumption, and a different type of model!

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

Asymmetry in Directed Models

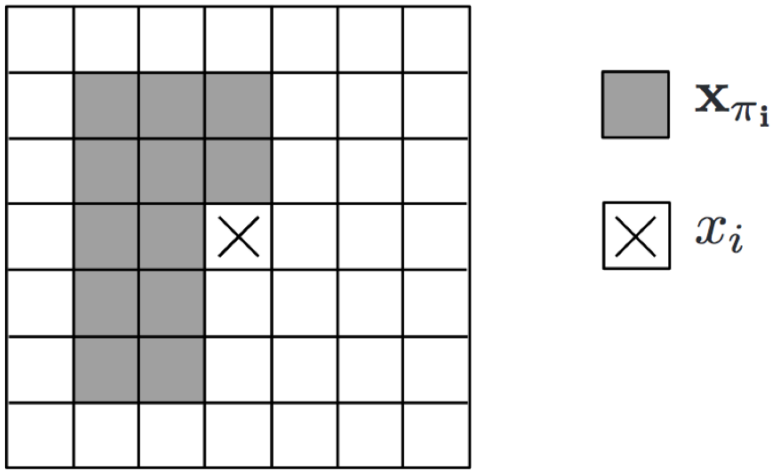
- Directed models assert that

$$X_i \perp X_1, ..., X_{i-1} | X_{\pi(i)}$$

- They do not assert that

$$X_i \perp X_1, ..., X_{i-1}, X_{i+1}, ..., X_n | X_{\pi(i)}$$
$$X_i \perp X_{-i} | X_{\pi(i)}$$

- In some applications, this asymmetry is awkward and/or a poor modeling choice.



Undirected Graphs

- Undirected models are based on a fundamentally different conditional independence assumption.

- Given an undirected graph, an undirected model (or Markov Random Field) is associated with probability distributions obeying:

$$X_A \perp X_B | X_C \text{ if and only if } C \text{ separates } A \text{ from } B$$

Which is true?

~~$A)\ x_1 \perp x_3 | x_2$~~

true $B)\ x_1 \perp x_3 | x_{2,4}$

true $C)\ x_1 \perp x_3 | x_{2,5}$

true $D)\ x_6 \perp x_1 | x_{2,3,4,5}$

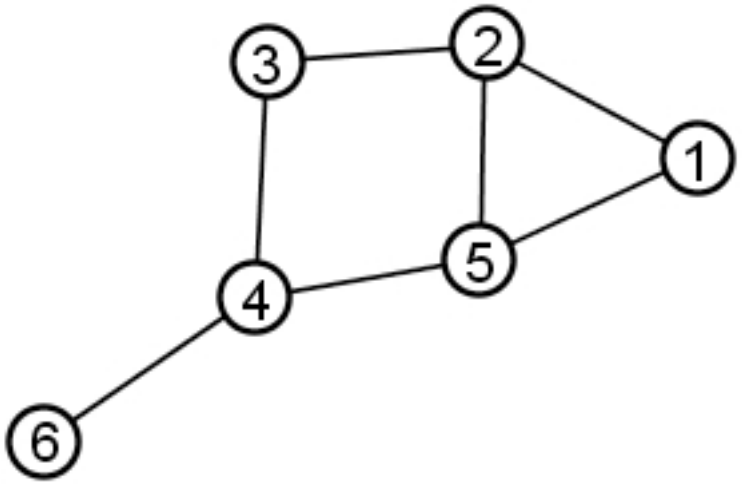
true $E)\ x_6 \perp x_1 | x_{2,4}$

true $F)\ x_6 \perp x_1 | x_4$

~~$G)\ x_6 \perp x_1 | x_2$~~

~~$H)\ x_{6,1} \perp x_{3,5} | x_{2,4}$~~

~~$I)\ x_{6,1} \perp x_{3,5} | x_4$~~

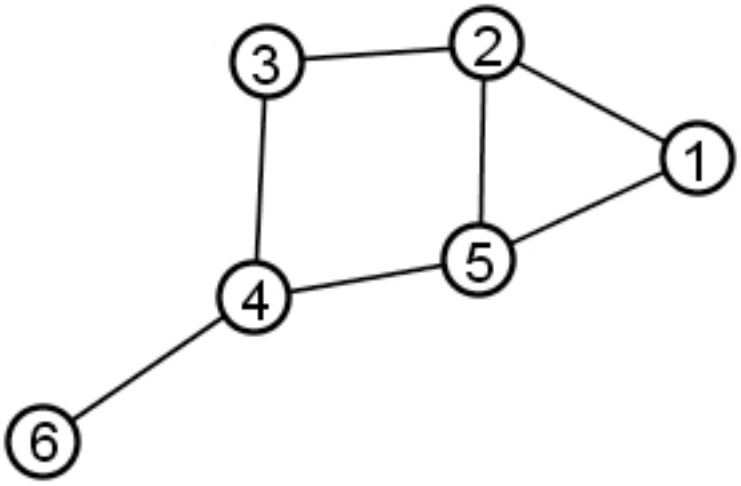


Undirected Graphs

- Equivalently (when $p(x) > 0$), a graph asserts that

$$p(x_i|x_{-i}) = p(x_i|x_{N(i)})$$

$$\begin{aligned}
p(x_1|x_{-1}) &= p(x_1|x_2, x_5) \\
p(x_2|x_{-2}) &= p(x_2|x_1, x_3, x_5) \\
p(x_3|x_{-3}) &= p(x_3|x_2, x_4) \\
p(x_4|x_{-4}) &= p(x_4|x_3, x_5) \\
p(x_5|x_{-5}) &= p(x_5|x_1, x_2, x_4) \\
p(x_6|x_{-6}) &= p(x_6|x_4)
\end{aligned}$$



Undirected Graphs

- Equivalently (when $p(x) > 0$), a graph asserts that

$$p(x_i|x_{-i}) = p(x_i|x_{N(i)})$$

- Great, but what's the formula for $p(x)$?

- Hammersley-Clifford Theorem.
  - Hammersley and Clifford found answer in 1971 via "blackening algebra"
  - H&C didn't like the positivity condition. Delayed publishing.
  - Besag proved/published theorem in 1974.
    - Grimmett, Preston and Sherman, same year.
  - Moussouris in 1974 gave an example with 4 nodes that needs positivity.

Hammersley-Clifford Theorem

- A positive distribution $p(x) > 0$ obeys the conditional independencies of a graph G when $p(x)$ can be represented as

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

where $\mathcal{C}$ is the set of all cliques, and

$$Z = \sum_x \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

is the "partition function". (Who we will meet later!)

Notes:

- This isn't obvious!
- No direct probabilistic interpretation for ψ .

Hammersley-Clifford Theorem

Proof sketch:

- It's easy to show that $p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$ obeys this conditional independence assumptions of a graph.

- Instead, we start with a arbitrary distribution that obeys the conditional independence assumptions, and must show that it can indeed be written like this.

Define $x^* = (0, 0, ..., 0)$ and $Q(x) := \ln(p(x)/p(x^*))$.

Step 1: Can write Q **uniquely** as:

$$Q(x) = \sum_i x_i G_i(x_i) + \sum_{i<j} x_i x_j G_{ij}(x_i, x_j) + \sum_{i<j<k} x_i x_j x_k G_{ijk}(x_i, x_j, x_k) + ... + x_1 x_2 ... x_n G_{12...n}(x_1, x_2, ..., x_n)$$

Step 2: Define $x^i = (x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_n)$. Then,

$$\exp(Q(x) - Q(x^i)) = \frac{p(x)}{p(x^i)} = \frac{p(x_i|x_{-i})}{p(0|x_{-i})}$$

Step 3: Pick node 1 W.O.L.O.G. Then, write

$$Q(x) - Q(x^1) = x_1(G_1(x_1) + \sum_{1<j} x_j G_{1j}(x_1, x_j) + \sum_{1<j<k} x_j x_k G_{1jk}(x_i, x_j, x_k) + ... + x_2 ... x_n G_{12...n}(x_1, x_2, ..., x_n))$$

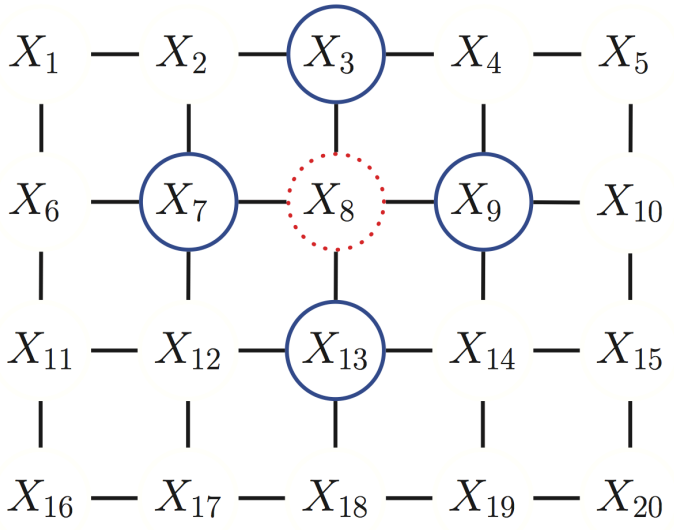
Step 4: Suppose t is not a neighbor of 1. All terms involving x_t must be zero.

Why? A) $Q(x) - Q(x^1)$ is independent of x_t since $p(x_i|x_{-i})$ is.
B) If we set $x_i = 0, i \notin \{1, t\}$ then G_{1t} must be zero.
C) Similarly for third/fourth/n-th order terms

Hammersley-Clifford Summary

- If $p(x) > 0$, and there is a graph such that $p(x_i|x_{-i}) = p(x_i|x_{N(i)})$, then

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$



What are the factors?

Only need neighboring pairs in the graph!

Exponential Family

An exponential family is a set of distributions

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp\left(\theta^T \phi(x)\right)$$
$$= \exp\left(\theta^T \phi(x) - A(\theta)\right)$$

parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$.

$Z(\theta) = \sum_x \exp(\theta^T \phi(x))$ if discrete.

$A(\theta) = \log Z(\theta)$ is the "log-partition function".

We care because:

Many interesting properties.

Undirected models are an exponential family.

Here, "zustandssumme" = "sum over states"

Exponential Family

- {Gaussian, Bernoulli, Binomial, Poisson, Exponential, Weibull, Laplace, gamma, beta, multinomial, Wishart} distributions are all exponential families.
  - All these share much common structure.

- We discuss only three facts:
  - The derivatives of the log-partition function are moments.
  - Maximum-likelihood learning is equivalent to moment matching.
  - Undirected models are exponential families.

Derivatives of A

$$p(x;\theta) = \exp\left(\theta^T \phi(x) - A(\theta)\right)$$

- A seemingly “magical” property of A is that it’s derivatives are the cumulants of the distribution.

$$\frac{dA}{d\theta} = \mathbb{E}_\theta[\phi(X)] \qquad \frac{dA}{d\theta d\theta^T} = \mathbb{V}_\theta[\phi(X)]$$

Proof that $\frac{dA}{d\theta} = \mathbb{E}_\theta[\phi(X)]$

$$
\begin{aligned}
\frac{dA}{d\theta} &= \frac{d}{d\theta} \log \sum_x \exp(\theta^T \phi(x)) \\
&= \frac{1}{\sum_x \exp(\theta^T \phi(X))} \frac{d}{d\theta} \sum_x \exp(\theta^T \phi(x)) \\
&= \frac{1}{Z(\theta)} \frac{d}{d\theta} \sum_x \exp(\theta^T \phi(x)) \\
&= \frac{1}{Z(\theta)} \sum_x \frac{d}{d\theta} \exp(\theta^T \phi(x)) \\
&= \frac{1}{Z(\theta)} \sum_x \exp(\theta^T \phi(x)) \frac{d}{d\theta} \theta^T \phi(x) \\
&= \frac{1}{Z(\theta)} \sum_x \exp(\theta^T \phi(x)) \phi(x) \\
&= \sum_x p(x;\theta) \phi(x)
\end{aligned}
$$

Proof that $\frac{dA}{d\theta d\theta^T} = \mathbb{V}_\theta[\phi(x)]$

$$
\begin{aligned}
\frac{dA}{d\theta d\theta^T} &= \frac{d}{d\theta^T}\frac{d}{d\theta}A(\theta) \\
&= \frac{d}{d\theta^T}\sum_x p(x;\theta)\phi(x) \\
&= \sum_x \frac{d}{d\theta^T}p(x;\theta)\phi(x) \\
&= \sum_x p(x;\theta)\phi(x)\frac{d}{d\theta^T}\left(\theta^T\phi(x) - A(\theta)\right) \\
&= \sum_x p(x;\theta)\phi(x)\left(\phi(x)^T - \mathbb{E}_\theta[\phi(X)]^T\right) \\
&= \sum_x p(x;\theta)\phi(x)\phi(x)^T - \sum_x p(x;\theta)\phi(x)\mathbb{E}_\theta[\phi(X)]^T \\
&= \mathbb{E}_\theta[\phi(X)\phi(X)^T] - \mathbb{E}_\theta[\phi(X)]\mathbb{E}_\theta[\phi(X)]^T
\end{aligned}
$$

Maximum Likelihood Learning

- Given $x^1, x^2, ..., x^D$, we want to solve

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D} \sum_{d=1}^{D} \log p(x^d; \theta)$$

- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

Maximum Likelihood Learning

$$\frac{dL}{d\theta} = \frac{1}{D}\sum_{d=1}^{D}\frac{d}{d\theta}\log p(x^d;\theta)$$

$$= \frac{1}{D}\sum_{d=1}^{D}\frac{d}{d\theta}(\theta^T\phi(x^d) - A(\theta))$$

$$= \frac{1}{D}\sum_{d=1}^{D}\phi(x^d) - \mathbb{E}_\theta[\phi(X)]$$

$$= \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_\theta[\phi(X)]$$

Maximum Likelihood Learning

- Given $x^1, x^2, ..., x^D$, we want to solve

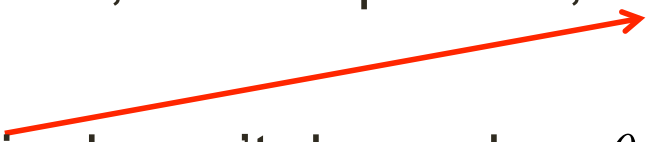
$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

$$\frac{dL}{d\theta} = \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_\theta[\phi(X)]$$

  - Notice, at the optimum, $\hat{\mathbb{E}}[\phi(X)] = \mathbb{E}_\theta[\phi(X)]$.



  - This doesn't depend on θ. Called a sufficient statistic.

What you know so far

- Definition: $p(x;\theta) = \exp\left(\theta^T \phi(x) - A(\theta)\right)$
- Derivatives: $\frac{dA}{d\theta} = \mathbb{E}_\theta[\phi(X)]$
- Maximum likelihood and moment matching

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d|\theta)$$

$$\hat{\mathbb{E}}[\phi(X)] = \mathbb{E}_\theta[\phi(X)]$$

- Now: What does this tell us about undirected models?

Undirected Models

- Typically written as

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

- Re-write as

$$p(x; \theta) = \exp(\theta^T \phi(x) - A(\theta))$$

$$\phi(x) = \{\mathbb{I}(x_c = x_c^*) | c \in \mathcal{C}, \text{all possible } x_c^*\}$$

Four node

$$\frac{dA}{d\theta} = \mathbb{E}_\theta[\phi(X)] = [p(x_1=0,x_2=0;\theta), p(x_1=0,x_2=1;\theta),$$
$$p(x_1=1,x_2=0;\theta), p(x_1=1,x_2=1;\theta),$$
$$p(x_2=0,x_3=0;\theta), p(x_2=0,x_3=1;\theta),$$
$$\ldots,$$
$$p(x_3=1,x_4=0;\theta), p(x_3=1,x_4=1;\theta)]$$

- Assume x is b

$$p(x) = \frac{1}{Z}\psi_{12}(x_1,x_2)\psi_{23}(x_2,x_3)\psi_{34}(x_3,x_4)$$

- Equivalent to $p(x;\theta) = \exp(\theta^T\phi(x) - A(\theta))$ with

$$\theta = [\theta(x_1=0,x_2=0), \theta(x_1=0,x_2=1),$$
$$\theta(x_1=1,x_2=0), \theta(x_1=1,x_2=1),$$
$$\theta(x_2=0,x_3=0), \theta(x_2=0,x_3=1),$$
$$\ldots,$$
$$\theta(x_3=1,x_4=0), \theta(x_3=1,x_4=1)]$$

$$\phi(x) = [\mathbb{I}(x_1=0,x_2=0), \mathbb{I}(x_1=0,x_2=1),$$
$$\mathbb{I}(x_1=1,x_2=0), \mathbb{I}(x_1=1,x_2=1),$$
$$\mathbb{I}(x_2=0,x_3=0), \mathbb{I}(x_2=0,x_3=1),$$
$$\ldots,$$
$$\mathbb{I}(x_3=1,x_4=0), \mathbb{I}(x_3=1,x_4=1)]$$

Undirected Models = Exponential Families

An undirected model is an E.F. where $\phi(x)$ has indicator functions for every configuration of every clique.

So, $\mathbb{E}_\theta[\phi(x)]$ is the marginals of the distribution, $p(x_c;\theta)$.

But, recall that $\frac{dA}{d\theta} = \mathbb{E}_\theta[\phi(X)]$. So, the gradient of the log-partition function is a vector of all the marginals.

Recall also that at the maximum likelihood solution,

$$\hat{\mathbb{E}}[\phi(X)] = \mathbb{E}_\theta[\phi(X)]$$

Thus, at the maximum likelihood solution, $p(x_c;\theta) = \hat{p}(x_c)$.

Directed and Undirected Models

- Directed and undirected models stem from similar conditional independence assumptions:

Directed	Undirected
$p(x_i \mid x_1, ..., x_{i-1}) = p(x_i \mid x_{\pi(i)})$	$p(x_i \mid x_{-i}) = p(x_i \mid x_{N(i)})$

- These lead to quite different looking joint distributions:

$p(x) = \prod_i p(x_i \mid x_{\pi(i)})$	$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$

- Still, the ML solution has a similarity:

$p(x_i \mid = x_{\pi(i)}; \theta) = \hat{p}(x_i \mid x_{\pi(i)})$	$p(x_c; \theta) = \hat{p}(x_c)$

Summary

- Undirected models assume that $p(x_i|x_{-i}) = p(x_i|x_{N(i)})$.
- The Hammersley-Clifford Theorem says that such a distribution can be written as $p(x) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\psi_c(x_c)$.
- An EF is defined by $p(x;\theta) = \exp\left(\theta^T\phi(x) - A(\theta)\right)$.
- For a EF, $dA/d\theta = \mathbb{E}_\theta[\phi(X)]$.
- At the ML solution of an EF, $\hat{\mathbb{E}}[\phi(X)] = \mathbb{E}_\theta[\phi(X)]$.
- An undirected model is an EF where $\phi(x)$ is indicator functions for each configuration of each clique.
- For an undirected model, $\frac{dA}{d\theta}$ is all marginals.
- At the ML solution for an undirected model, $p(x_c;\theta) = \hat{p}(x_c)$

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

The (marginal) inference problem

- Take a graphical model $p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$

(If $\mathcal{C}$ is arbitrary, this could be directed or undirected.)

- Suppose you want to know $p(X_3 = x_3) = \sum_{x_1} \sum_{x_2} \sum_{x_4} \ldots \sum_{x_n} p(x)$
- This looks hard if n is large.
- Even computing Z looks hard if n is large!

Graph Structure

- It turns out that the difficulty of computing $p(x_3)$ (or Z) is related to structural properties of the graph.
- For trees, we can do everything efficiently.
- For graphs “close” to a tree, we can transform to a tree.
- For general graphs, inference is #P-hard.

$$p(x) = \frac{1}{Z}\psi(x_1, x_2)\psi(x_2, x_3)...\psi(x_{n-1}, x_n). \quad \text{Want: } p(x_i)$$

If we had access to the values

$$m_{i-1}(x_i) = \sum_{x_1} \sum_{x_2} ... \sum_{x_{i-1}} \psi(x_1, x_2)\psi(x_2, x_3)...\psi(x_{i-1}, x_i)$$
$$m_{i+1}(x_i) = \sum_{x_{i+1}} ... \sum_{x_n} \psi(x_i, x_{i+1})\psi(x_{i+1}, x_{i+2})...\psi(x_{n-1}, x_n)$$

Then, we can calculate $p(x_i) \propto m_{i-1}(x_i)m_{i+1}(x_i).$

But: Note the recursions: $m_i(x_{i+1}) = \sum_{x_i} \psi(x_i, x_{i+1})m_{i-1}(x_i)$

$$m_i(x_{i-1}) = \sum_{x_i} \psi(x_{i-1}, x_i)m_{i+1}(x_i)$$

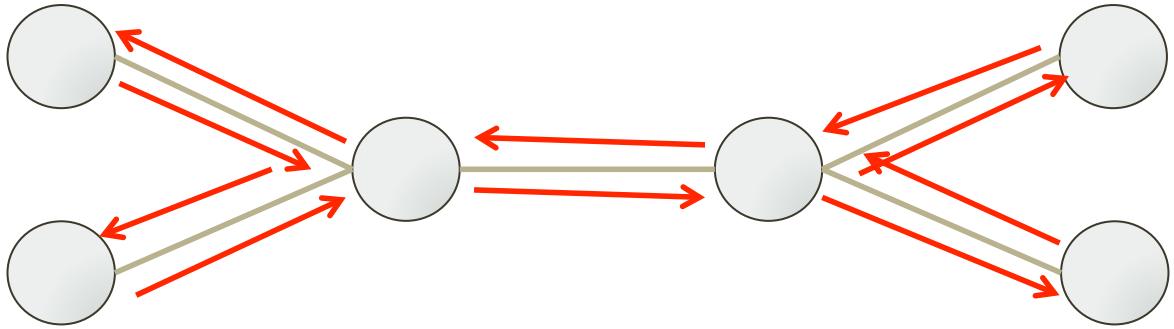
Belief Propagation:

- Initialize $m_i(x_j) = 1$
- For all pairs (i, j) set $m_i(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k \in N(i) \setminus j} m_k(x_i)$
- Recover $p(x_i) \propto m_{i-1}(x_i)m_{i+1}(x_i)$

Belief Propagation

- Initialize $m_i(x_j) = 1$
- For all pairs (i, j) set $m_i(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k \in N(i) \setminus k} m_k(x_i)$
- Recover $p(x_i) \propto m_{i-1}(x_i) m_{i+1}(x_i)$

- Also works on trees.



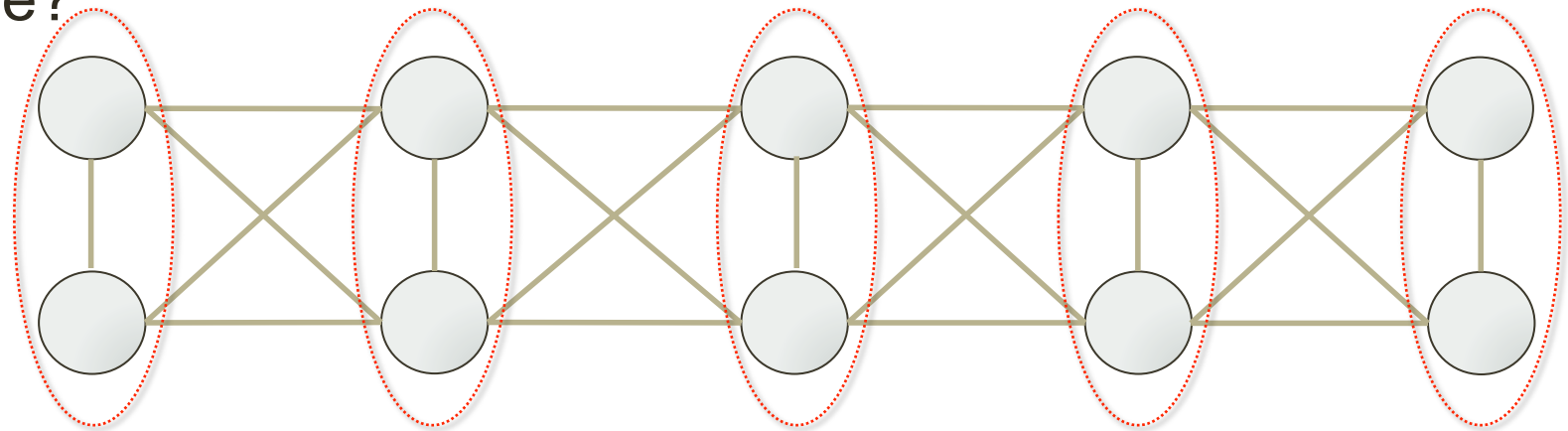
- Just need one pass over full graph in each direction.

Belief Propagation

- AKA “Sum-Product” algorithm. Very related to “alpha-beta”, and “Viterbi” algorithms.

Inference on “nearly trees”

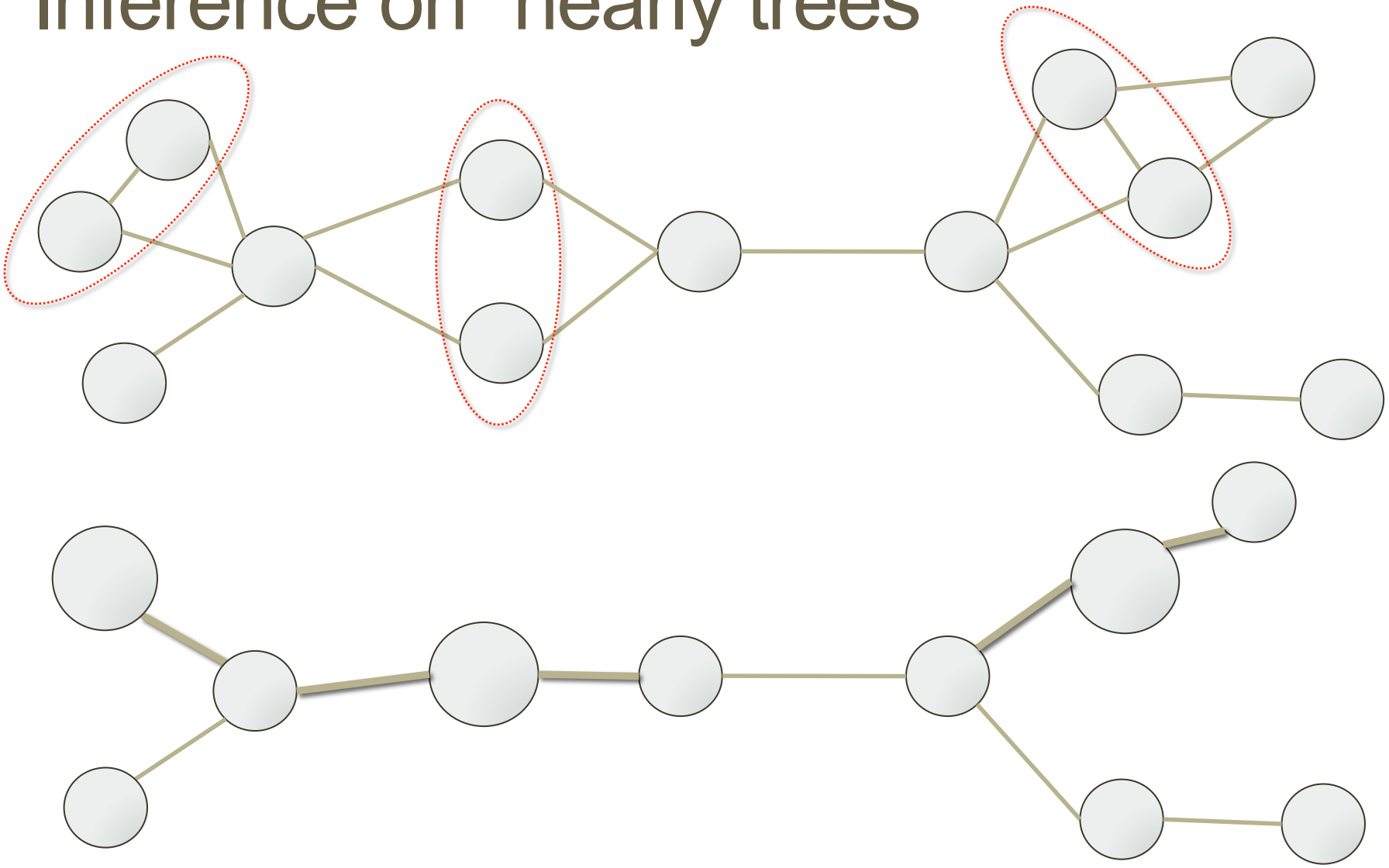
- But, what if we need to do inference on something that isn’t a tree?



- Idea: Cluster nodes:

- If original nodes have S states, “super” nodes have S^2 states

Inference on “nearly trees”

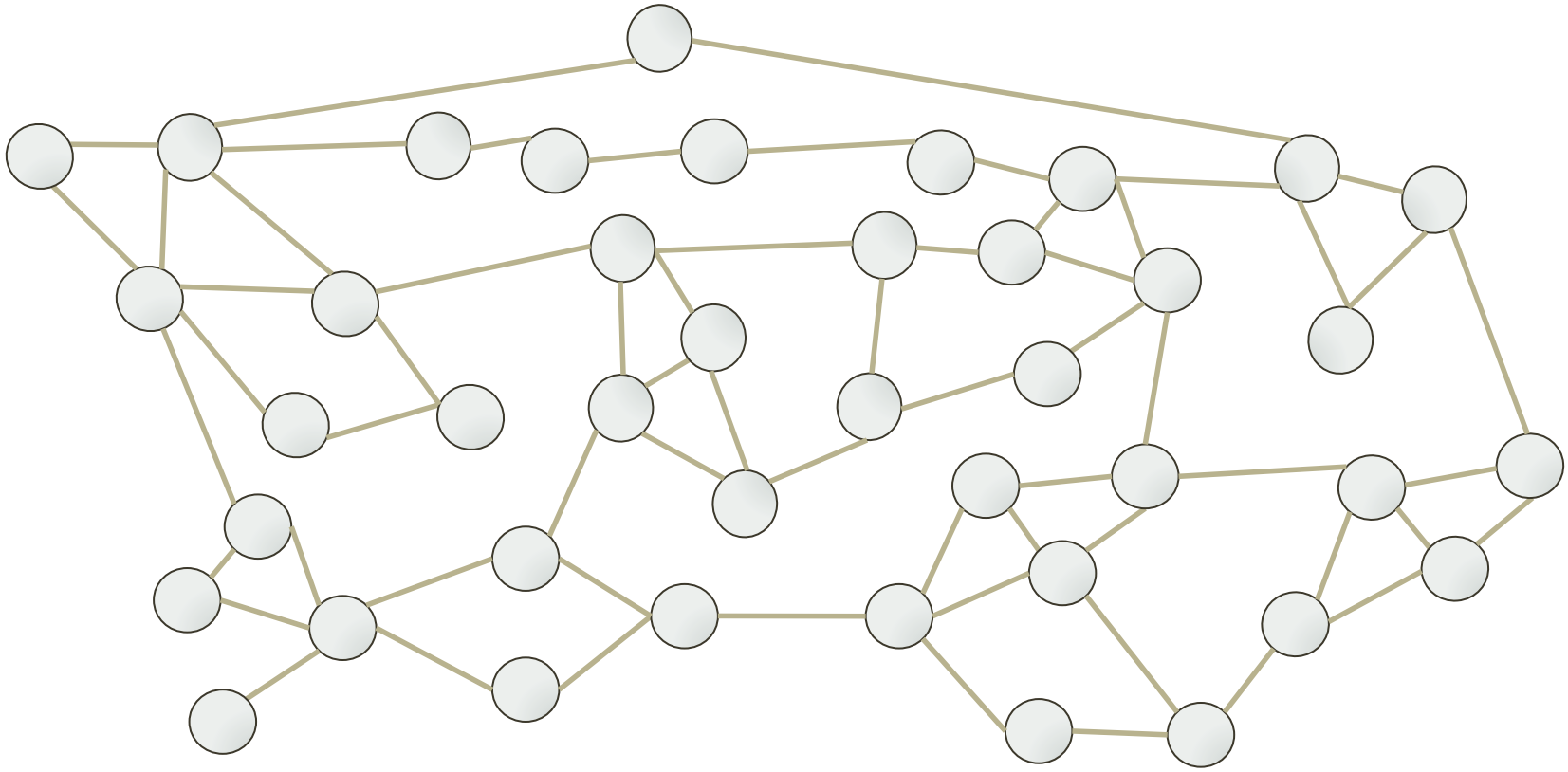


Inference on “nearly trees”

- Junction Tree Algorithm: Take an arbitrary graph, cluster nodes to obtain a tree, run sum-product on the tree.
- Complexity of inference on final graph exponential in size of cluster.
- Multiple ways to cluster nodes. (NP-hard to find best.)
- The smallest possible clustering is the “treewidth” of the graph. (NP-hard to compute.)

Inference on general graphs

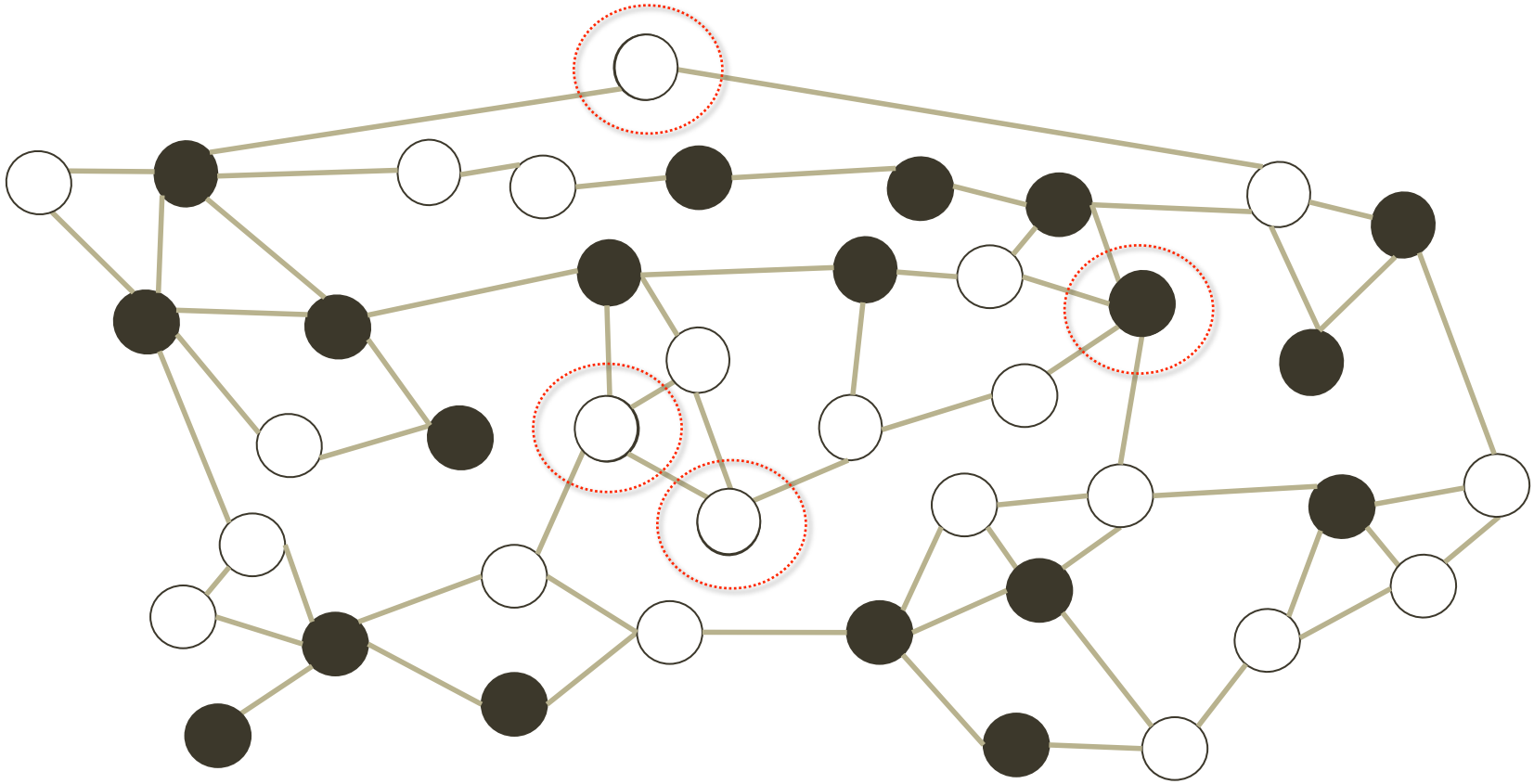
- What if the graph isn't even close to a tree?



- One option: Run BP, hope for the best. "Loopy Belief Propagation"

Inference on general graphs

- Another option: Markov chain Monte Carlo



- Gibbs sampling: Repeatedly pick i randomly and sample

$$x_i \sim p(x_i | x_N(i))$$

Summary: How to calculate marginals?

- On small graphs:
  - Brute force summation:
- On trees:
  - Belief propagation
- On near trees:
  - Transform into a tree, run BP. (Junction-tree algorithm.)
- On general graphs:
  - Approximate inference.
  - Markov chain Monte Carlo.

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

How to do learning

- So far we have discussed:
  - How conditional independence assumptions lead to factorized distributions.
  - Given such a distribution, how to do inference.
- But this is MLSS, so let's talk about some ML.
- We (mostly) restrict ourselves to
  - Undirected models
  - Discrete variables

Maximum likelihood

- Given $x^1, x^2, ..., x^D$, we want to solve

$$\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(x^d; \theta)$$

- Assume structure known. Only need to estimate θ.
- Simple approach: Gradient descent. Repeatedly set

$$\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$$

- Recall that $\frac{dL}{d\theta} = \hat{\mathbb{E}}[\phi(X)] - \mathbb{E}_\theta[\phi(X)]$.

- Algorithm. Repeat:
  - Estimate current marginals $p(x_c|\theta)$ for all c and x_c.
  - Calculate $dL/d\theta$ from all the differences $\hat{p}(x_c) - p(x_c; \theta)$.
  - Take a small step, $\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$.

Maximum Likelihood

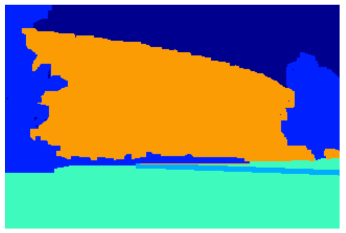
- Algorithm. Repeat:
  - Estimate current marginals $p(x_c|\theta)$ for all c and x_c.
  - Calculate $dL/d\theta$ from all the differences $\hat{p}(x_c) - p(x_c;\theta)$.
  - Take a small step, $\theta \leftarrow \theta + \lambda \frac{dL}{d\theta}$.

- Comments:
  - Need to be able to calculate marginals.
  - Better algorithms than simple gradient descent exist.
  - To calculate L along with it's gradient, you need to be able to calculate $A(\theta)$ as well as $dA/d\theta$.
  - Turns out that BP does return $A(\theta)$, as well as $dA/d\theta$ (as marginals). We skip the details here.

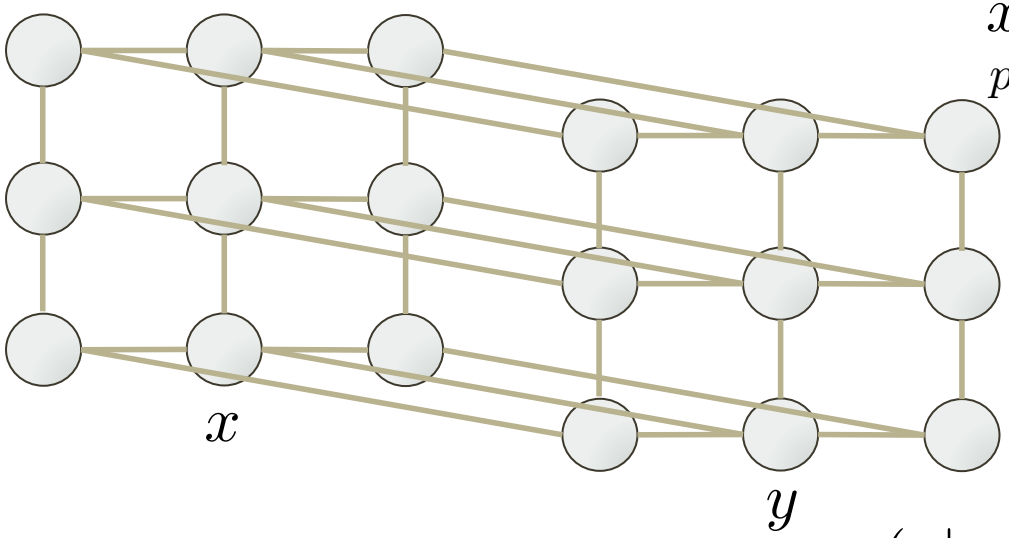
Conditional Likelihood

- Suppose that you want to build a model of (x=images) (y =segmentations)
- Can learn $p(x, y; \theta)$ to max.

$$\frac{1}{D} \sum_{d=1}^{D} \log p(x^d, y^d; \theta)$$



x y



$$p(x, y; \theta) = \prod_i \cancel{\psi(x_i; \theta)} \psi(y_i, x_i; \theta) \times \prod_{(i,j)} \cancel{\psi(x_i, x_j; \theta)} \psi(y_i, y_j; \theta)$$

- At test time, only make use of $p(y|x; \theta)$.
- Hmmmm….

Conditional Likelihood

- Idea: Since we only use $p(y|x;\theta)$ at test time, maybe we can directly train for this conditional distribution?

- Given $x^1, x^2, ..., x^D$, why not solve
  $\arg\max_\theta L(\theta), \quad L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(y^d|x^d;\theta)$?

- Suppose $p(x,y;\theta) = p(x;\theta^A)p(y|x;\theta^B), \theta =< \theta^A, \theta^B >$.
  - True for, e.g., directed models with only arrows from x to y.
  - L is independent of $p(x;\theta^A)$. Don't even need to specify it!
  - Exactly the same result for $p(y|x;\theta^B)$ as regular/"generative" likelihood.

- For undirected, no such separation between $p(x;\theta)$ and $p(y|x;\theta)$.
  - Conditional likelihood gives different result for $p(y|x;\theta)$.
  - Also, significant computational difference.

Conditional Likelihood (Undirected)

- If we want to optimize $L(\theta) = \frac{1}{D}\sum_{d=1}^{D} \log p(y^d|x^d;\theta)$ by gradient ascent, we need to be able to calculate $dL/d\theta$.

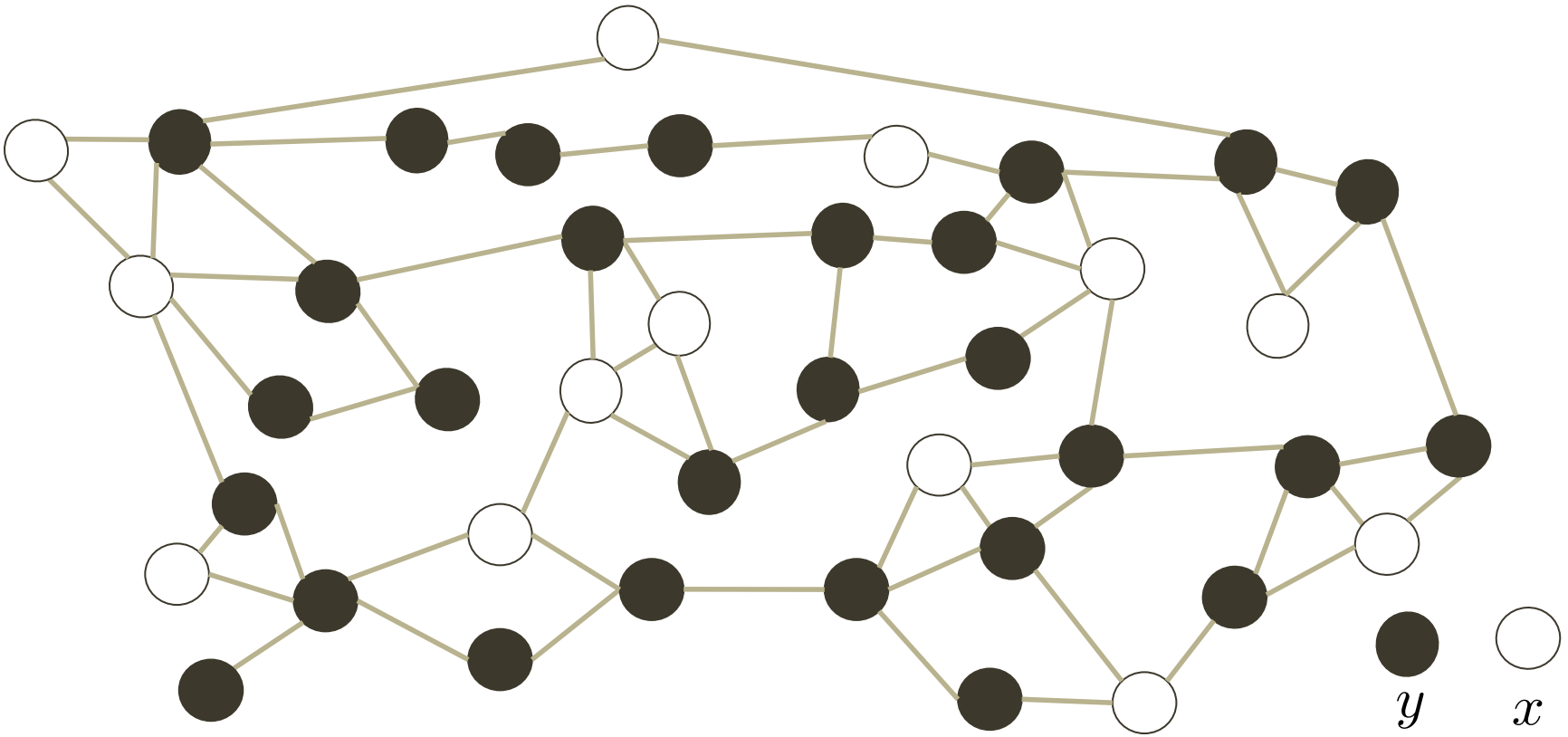
- Regular EF: $p(x,y;\theta) = \exp(\theta^T\phi(x,y) - A(\theta))$
  $A(\theta) = \log\sum_{x,y}\exp(\theta^T\phi(x,y))$
- Conditional EF: $p(y|x;\theta) = \exp(\theta^T\phi(x,y) - A(x;\theta))$
  $A(x;\theta) = \log\sum_{y}\exp(\theta^T\phi(x,y))$
  $\frac{dA(x;\theta)}{d\theta} = \mathbb{E}_\theta[\phi(X,Y)|X=x]$
- To calculate the loss:

$$L(\theta) = \frac{1}{D}\sum_{d=1}^{D}\left(\theta^T\phi(x^d,y^d) - A(x^d;\theta)\right)$$
$$\frac{dL}{d\theta} = \frac{1}{D}\sum_{d=1}^{D}\left(\phi(x^d,y^d) - \mathbb{E}_\theta[\phi(X,Y)|X=x^d]\right)$$

Conditional Likelihood (Undirected)

$$\frac{dL}{d\theta} = \frac{1}{D}\sum_{d=1}^{D}\left(\phi(x^d, y^d) - \mathbb{E}_\theta[\phi(X,Y)|X = x^d]\right)$$

- Calculating $\mathbb{E}_\theta[\phi(X,Y)|X = x]$ can be easier than $\mathbb{E}_\theta[\phi(X,Y)]$
- Need marginals $p(y_c|x^d;\theta)$ rather than $p(y_c, x_c;\theta)$.



Conditional Likelihood (Undirected)

$$\frac{dL}{d\theta} = \frac{1}{D}\sum_{d=1}^{D}\left(\phi(x^d, y^d) - \mathbb{E}_\theta[\phi(X, Y)|X = x^d]\right)$$

- On the other hand, need to do inference for each datum. With the generative likelihood, only need to calculate

$$\mathbb{E}_\theta[\phi(X, Y)]$$

once.

- So computationally, the conditional likelihood is:
  - Easier. ($p(y_c|x^d;\theta)$ instead of $p(y_c, x_c;\theta)$).
  - Harder. (Inference for each d).

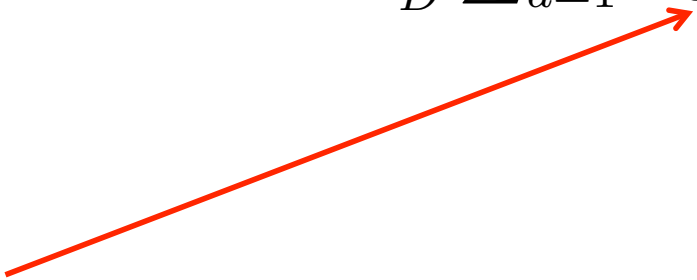
- What about the final θ? Is it better or worse?
  - Another tradeoff…

Conditional Models

$$\frac{1}{D}\sum_{d=1}^{D}\log p(y^d|x^d;\theta) \text{ vs. } \frac{1}{D}\sum_{d=1}^{D}\log p(x^d,y^d|\theta)$$

- We can always write that:

$$\frac{1}{D}\sum_{d=1}^{D}\log p(x^d,y^d|\theta) = \frac{1}{D}\sum_{d=1}^{D}\log p(y^d|x^d;\theta) + \frac{1}{D}\sum_{d=1}^{D}\log p(x^d|\theta)$$



- Switching from joint to conditional likelihood just means dropping this term.

Model Specification

$$\frac{1}{D}\sum_{d=1}^{D}\log p(x^d, y^d|\theta) = \frac{1}{D}\sum_{d=1}^{D}\log p(y^d|x^d;\theta) + \frac{1}{D}\sum_{d=1}^{D}\log p(x^d|\theta)$$

- If the model is not well-specified the conditional likelihood will tend to be better, given enough data.

- Joint likelihood converges to

$$\arg\min_\theta KL(p_0(x,y)||p(x,y|\theta)) = -\sum_{x,y} p_0(x,y)\log\frac{p_0(x,y)}{p(x,y|\theta)}$$

- Conditional likelihood converges to

$$\arg\min_\theta KL(p_0(y|x)\ ||\ p(y|x;\theta)) = -\sum_{x} p_0(x,y)\log\frac{p_0(y|x)}{p(y|x;\theta)}$$

Over-Fitting

- With a well-specified model, the joint likelihood will tend to over-fit less.

- Why? Both of these are minimized by true θ.

$$\frac{1}{D}\sum_{d=1}^{D}\log p(y^d|x^d;\theta)$$

$$\frac{1}{D}\sum_{d=1}^{D}\log p(x^d|\theta)$$

- With finite data, you face a trade-off.

Joint vs. Conditional Likelihood

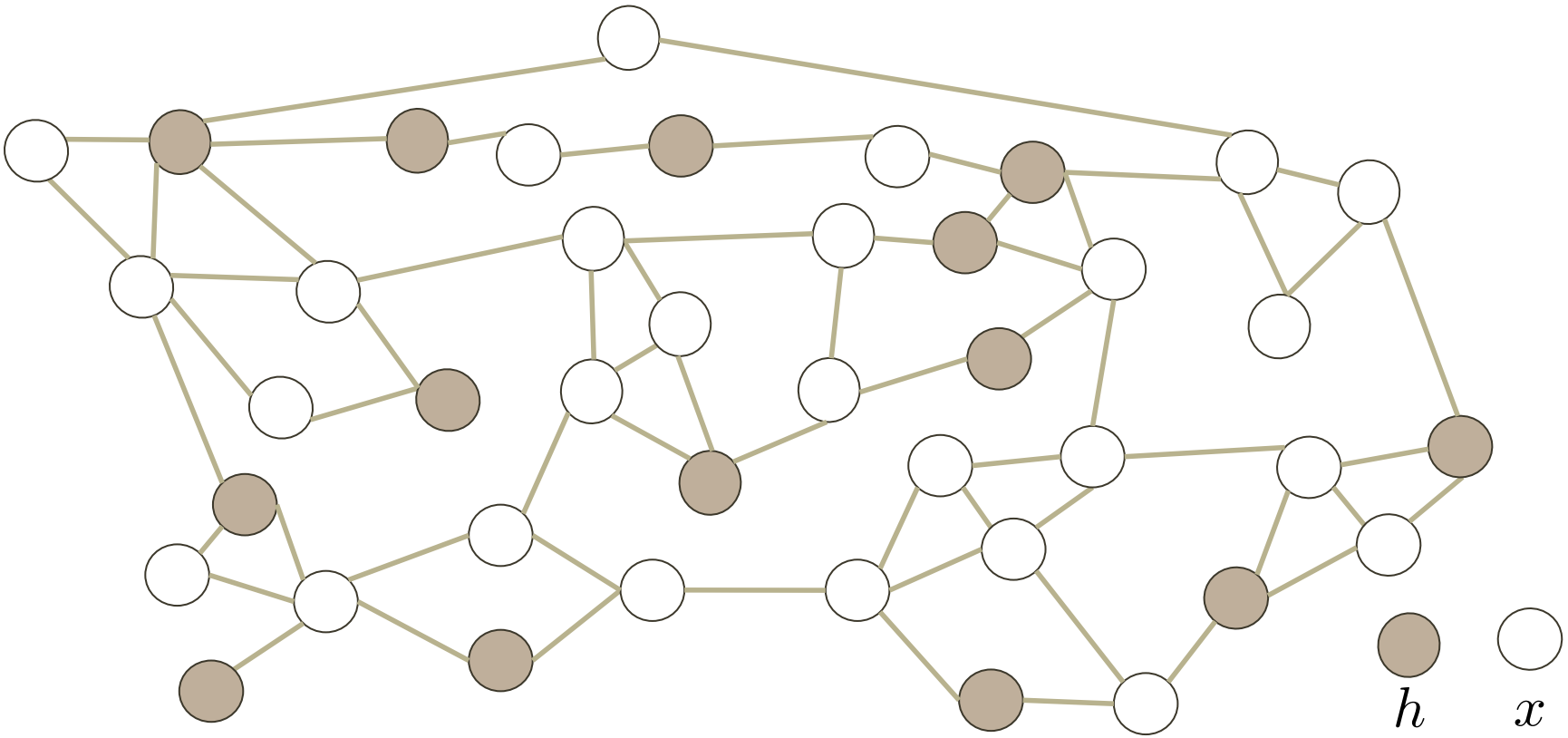
- Advantages of conditional likelihood:
  - With infinite data, at least as good. (Better if mis-specified)
  - Only need to compute $dA(x^d;\theta)/d\theta$ instead of $dA(\theta)/d\theta$.

- Advantage of joint likelihood:
  - Better generalization with finite data.
  - Only need to compute $dA(\theta)/d\theta$ once, rather than $dA(x^d;\theta)/d\theta$ for each datum.

- Alas, the real world tends to have finite data and mis-specified models.

Hidden Variables

- What if some of the data is hidden?
- Want a model of (x, h) together, but only have data for x.



Hidden Variables

- What if some of the data is hidden?

- Regular EF: $p(x, h; \theta) = \exp(\theta^T \phi(x, h) - A(\theta))$
$A(\theta) = \log \sum_{x,h} \exp(\theta^T \phi(x, h))$

- Conditional EF: $\log p(x; \theta) = A(x; \theta) - A(\theta)$
$A(x; \theta) = \log \sum_{h} \exp(\theta^T \phi(x, h))$
$\frac{dA(x;\theta)}{d\theta} = \mathbb{E}_\theta[\phi(X, H)|X = x]$

- To calculate the loss:

$$L(\theta) = \frac{1}{D} \sum_{d=1}^{D} \log p(x^d; \theta)$$
$$= \frac{1}{D} \sum_{d=1}^{D} A(x^d; \theta) - A(\theta)$$
$$\frac{dL}{d\theta} = \frac{1}{D} \sum_{d=1}^{D} \mathbb{E}_\theta[\phi(X, H)|X = x^d] - \mathbb{E}_\theta[\phi(X, H)]$$

  - Once "clamped" for data point, and once for all together.

Hidden Variables

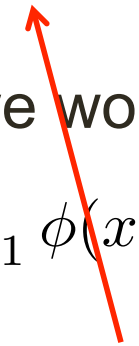
$$\frac{dL}{d\theta} = \frac{1}{D}\sum_{d=1}^{D} \mathbb{E}_\theta[\phi(X,H)|X = x^d] - \mathbb{E}_\theta[\phi(X,H)]$$

- At the optimum, it will be the case that

$$\frac{1}{D}\sum_{d=1}^{D} \mathbb{E}_\theta[\phi(X,H)|X = x^d] = \mathbb{E}_\theta[\phi(X,H)]$$

- If h was observed, we would have moment match

$$\frac{1}{D}\sum_{d=1}^{D} \phi(x^d, h^d) = \mathbb{E}_\theta[\phi(X,H)]$$



- Hidden loss finds a “hallucinated” moment match.

- Max likelihood with hidden data sometimes called “Expectatation Maximization”

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

Psudolikelihood

- The Likelihood is so hard to compute. Is there some other loss we might use? Pseudolikelihood:

$$L(\theta) = \sum_{d=1}^{D} \sum_i p(x_i^d | x_{N(i)}^D; \theta)$$

- Obvious variant for conditional setting.

- If model is well-specified, will converge to "true" θ , somewhat slower than likelihood.

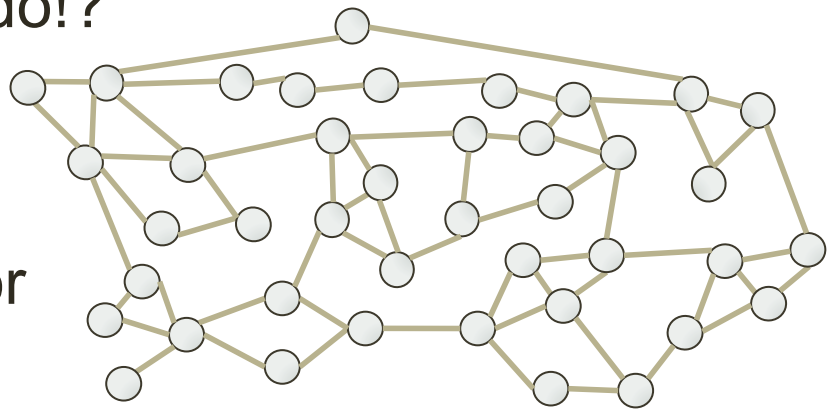
- If model is not well-specified, does not converge to the same answer as the likelihood. (Typically worse.)

Besag, "Statistical Analysis of Non-Lattice Data." 1975

Surrogate Likelihood

- Suppose our graph is a big loopy graph. Want to do Maximum likelihood. What to do!?
- Seeming heuristic option:

Use loopy belief propagation in place of true marginals, hope for the best.



- Actually, LBP can be understanding as (exactly) $d\tilde{A}/d\theta$ for an approximate $\tilde{A}(\theta)$.

- Regular likelihood: $\frac{1}{D}\sum_{d=1}^{D}\theta^T\phi(x^d;\theta) - A(\theta)$.
- Surrogate likelihood $\frac{1}{D}\sum_{d=1}^{D}\theta^T\phi(x^d;\theta) - \tilde{A}(\theta)$.

Wainwright, "Estimating the 'wrong' graphical model" 2006

Marginal-Based Learning

- Denote the approximate marginals produced by an approximate inference algorithm by $\mu(\theta)$.

- Define some kind of loss function

$$L(\theta) = \frac{1}{D} \sum_{d=1}^{D} Q(x^d; \mu(\theta))$$

$$L(\theta) = \frac{1}{D} \sum_{d=1}^{D} Q(y^d; \mu(x^d; \theta))$$

- $dQ/d\mu$ is easy to compute. $dQ/d\theta$ takes work.
- Q depends on application. Compensates for model defects and imperfect inference.

- Domke, ”Learning Graphical Model Parameters with Approximate Marginal Inference”, 2013
- Stoyanov and Eisner, “Minimum-risk training of approximate CRF-based NLP systems”, 2012

Outline

- Introduction
- Directed Models
- Undirected Models
- Inference
- Learning
- Life With Intractability
- Outro

What to read next

- Introductory:
  - Bishop, “Pattern Recognition and Machine Learning” Chapter 8 on graphical models freely available.
  - Murphy, “Machine Learning: A Probabilistic Perspective” (Especially chapters 17-27)
  - MacKay “Information Theory, Inference, and Learning Algorithms” Freely available.
- More advanced:
  - Wainwright and Jordan “Graphical models, exponential families, and variational inference”. Foundations and Trends in Machine Learning, 2008.
  - Koller and Friedman, “Probabilistic Graphical Models”, 2009.

Feedback

- I welcome your feedback!
  - Pacing: Too fast / too slow.
  - Coverage: More of this topic / less of that.
  - Topics that were particularly hard to understand.
  - Errors.

justin.domke@anu.edu.au