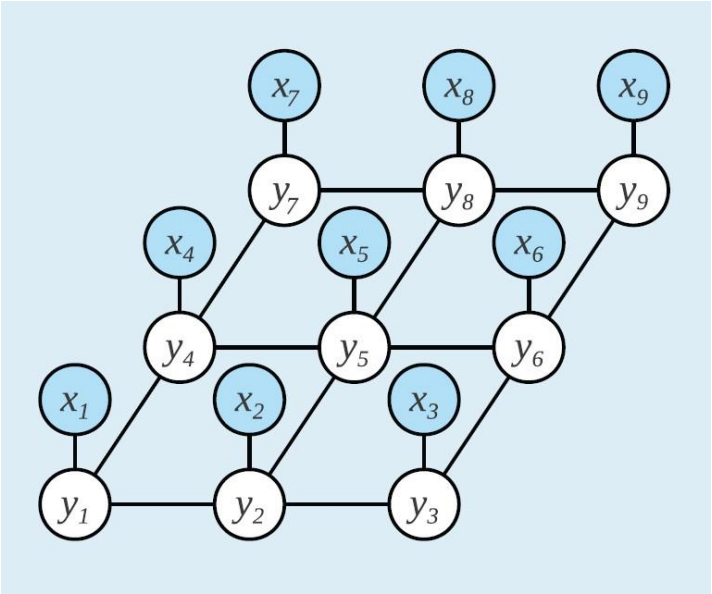
Probabilistic Graphical Models in Bioinformatics

Lecture 1: Introduction



Hesam Montazeri
Department of Bioinformatics, IBB, University of Tehran

Bahman 16, 1397
(Feb 5, 2019)

Overview

- **Course website:**
  https://github.com/hesmon/pgm

- **Teaching assistants:**
  - Naser Elmi and Samaneh Maleknia

- **prerequisites:** Statistics, basic probability theory, programming

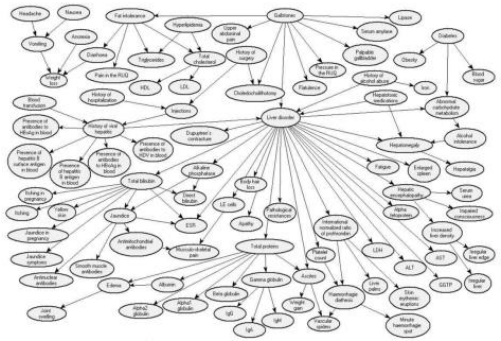
- A note to students with background in biology.

- **My Office hours**
  - Sunday 9:00-12:00

Probabilistic graphical model (PGM)

- *"Probabilistic graphical models are a marriage between graph theory and probability."*
- Graphical representations of probability distributions.

Probabilistic graphical models



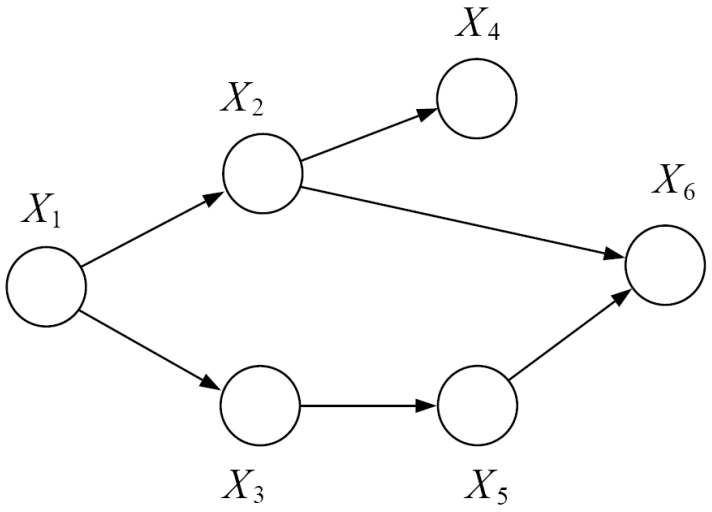
Data

159.6	103.3	82.7	88.9	91.3	553.1	123.9	134
463.8	119.8	63.6	75.1	160.3	97.5	117.2	78.1
42.1	116.9	72.9	257.6	243.3	28.9	344.9	164.6
732.4	768.6	411.3	829.6	608.4	543.4	388.9	488.5
4192.5	3933.5	4468.1	3479.2	4301.6	5582.6	3712.7	3351.4
8706.2	17183.3	10342.3	651.4	8546.2	13637.4	9616.9	8740
5402.7	5913.1	6058.4	4470.5	6159.1	7072.8	5219.3	5223.1
19502.4	16534.3	14855.6	11297.7	16252	19870.2	12580.9	12952.8
14081.4	13790.2	11699.8	8215	11276.3	16186.2	8365.4	10030.4
74377.5	67550.5	69036.1	49558.3	64603.6	73147.2	56352	52891.6
147249.2	140345	124028.6	91635.8	121466.1	148430.6	117098.4	122688.6
179460.8	186706.5	161774.2	125753.2	178065.3	187172.1	152429.9	159715.8
623	346.2	621.5	600.9	866.4	654.8	488	100.5
487.1	120.6	653.4	642.7	877.5	571.7	1020	438.3
263.1	100.2	190.7	108.9	315.2	163	478.5	174.9
126.3	152.2	73	115.5	104.8	286.3	152.1	110.3
38.8	67.9	58.4	20.5	101.2	45.2	56	53.6
1700.6	1859.7	2019.1	2016.9	2313.1	1640.6	2419.2	1968.1
783.8	1211	1775	1625.6	2043.3	1841.8	2491.5	1630.6
1008.1	970.5	793.2	676.2	1304.8	1348.2	1437.6	1020.2
2770.3	2605.9	2320.9	1912.8	3441.6	3029.3	2570	2470.1
223.9	313.1	63.5	178	243.6	323.7	54.9	50.6
318.5	361.3	153.6	271.8	391.3	243.4	308.6	93.7
259.6	269.7	33	158.4	242.5	52.3	71.1	138.9
165.6	264.1	39.8	26.5	166.3	52.5	28.3	135.9
112.1	281.6	280.7	57.4	139	361.6	70.2	338.2
106.4	55.4	12.1	31.6	33.5	45.1	161.5	282.1
33.6	45.6	20.7	21.3	44	27.2	48.1	50.6

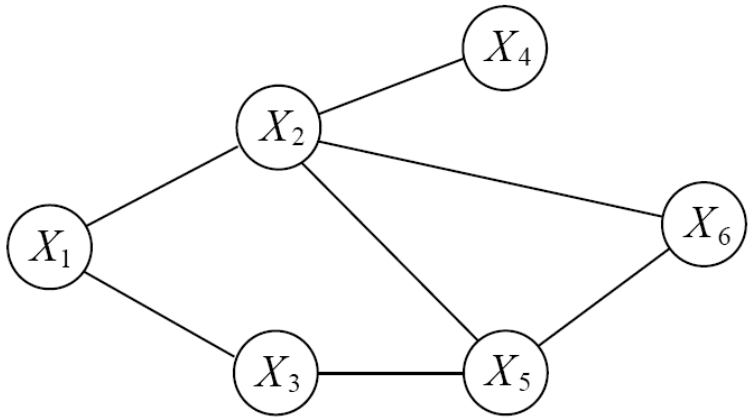
A useful tool for handling uncertainty and complexity

PGM

- Nodes represent random variables/states
- Edges represent probabilistic interaction between variables.
  - The missing arcs represent conditional independence assumptions
  - Makes it simpler to write down the **joint distribution** $P(X_1, X_2, X_3, X_4, X_5, X_6)$



Directed graphical models



undirected graphical models

Why joint distribution is important?

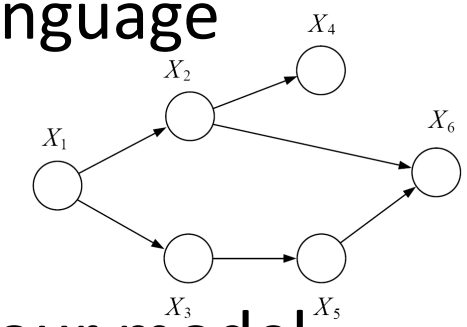
- I: intelligence, D: difficulty of the course,
  G: student's grade

- We can answer different queries using the joint distribution:
  - $P(G = g^1)$
  - $P(D = d_0 \mid I = i_1, G = g_1)$
  - $P(D = d_0 \mid I = i_1)$

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Syllabus overview

- **Representation:** to represent a distribution in a graphical language



- **Inference:** to answer queries using the joint distribution as our model of world

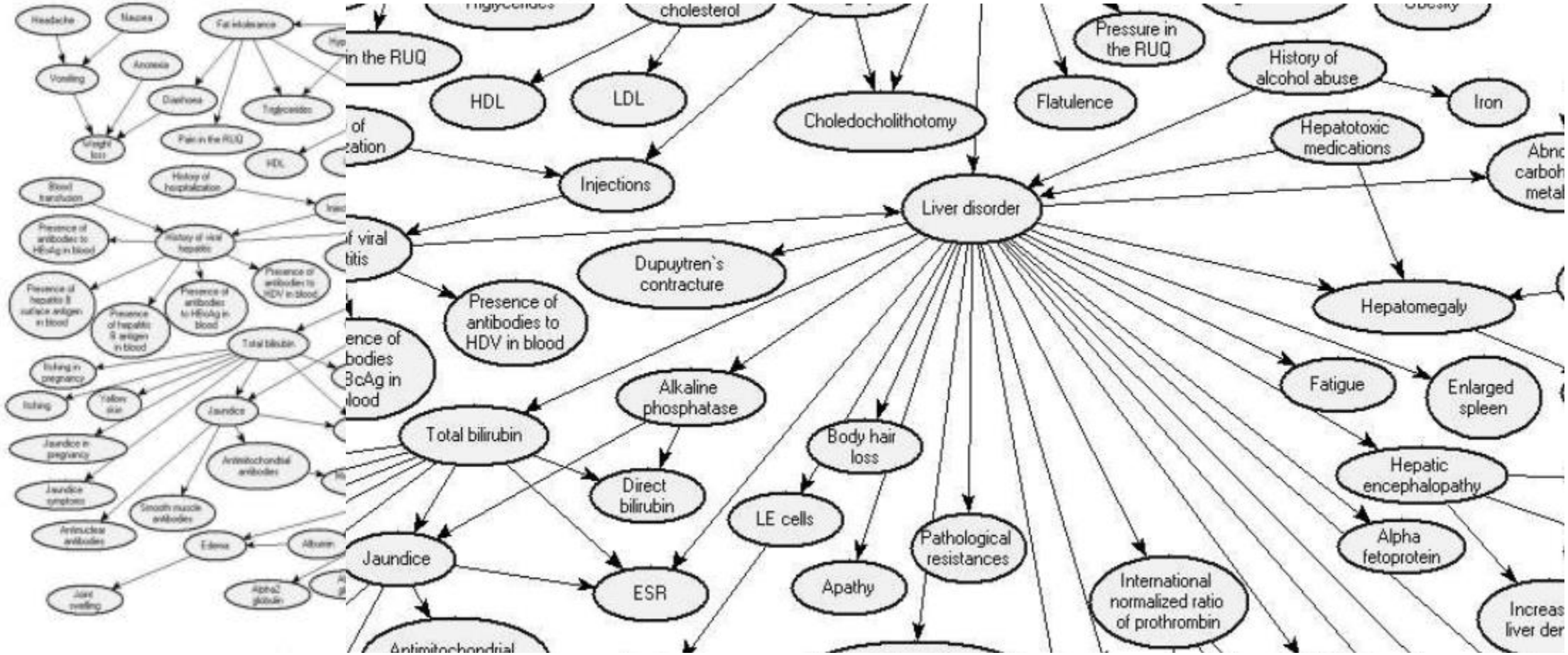
- **Learning:** learning graphical models from data

- Applications in bioinformatics

Outline

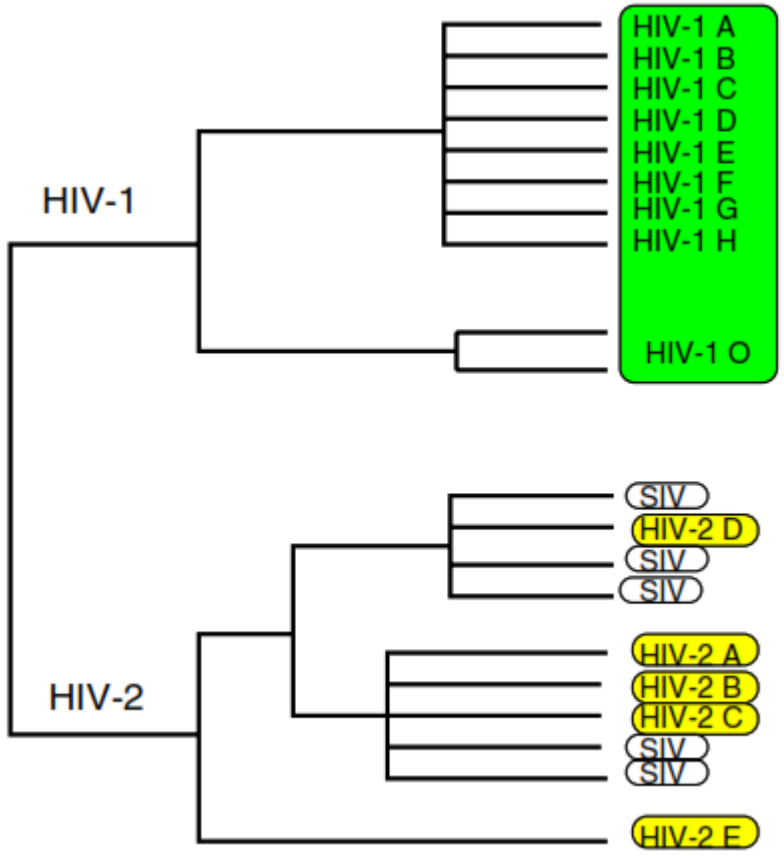
- Motivation: real-world applications of PGMs
- Recap: probability and statistics

Applications-Medical diagnosis system

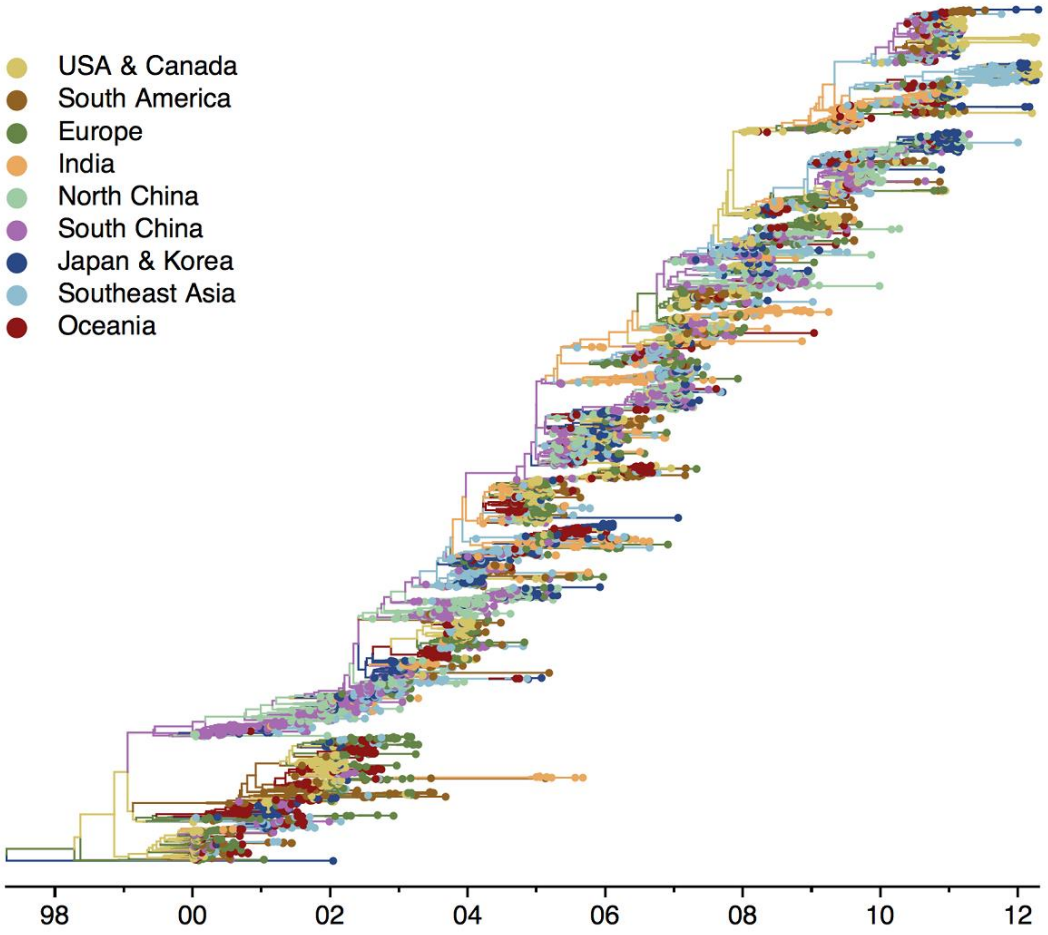


A Bayesian Network Model for Diagnosis of Liver Disorders *[Onisko ,1999]*

Applications- phylogenetic analysis



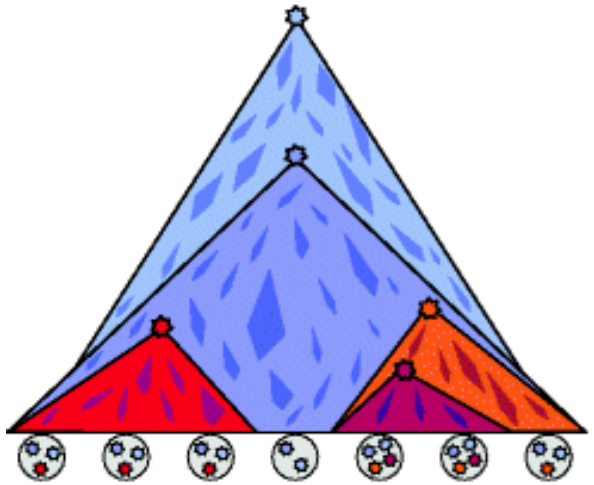
Cross-infection between humans and monkeys



Global influenza phylogeny

[Bedford et al, 2015]

Applications- tumor evolution

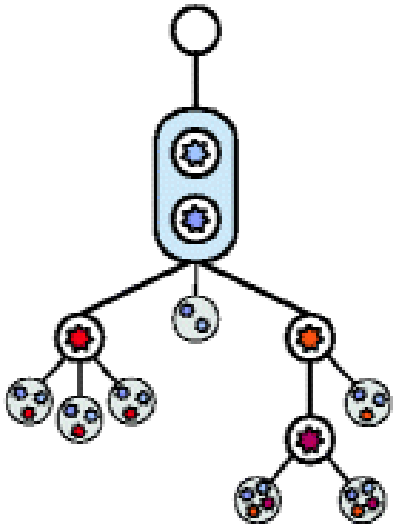


tumor evolution with time progressing downwards

cells

mutations	S_1	S_2	S_3	S_4	S_5	S_6	S_7
	1	1	-	1	1	-	1
	1	1	1	1	0	1	1
	1	1	1	0	0	0	0
	1	0	0	0	1	1	1
	0	0	-	0	0	1	0

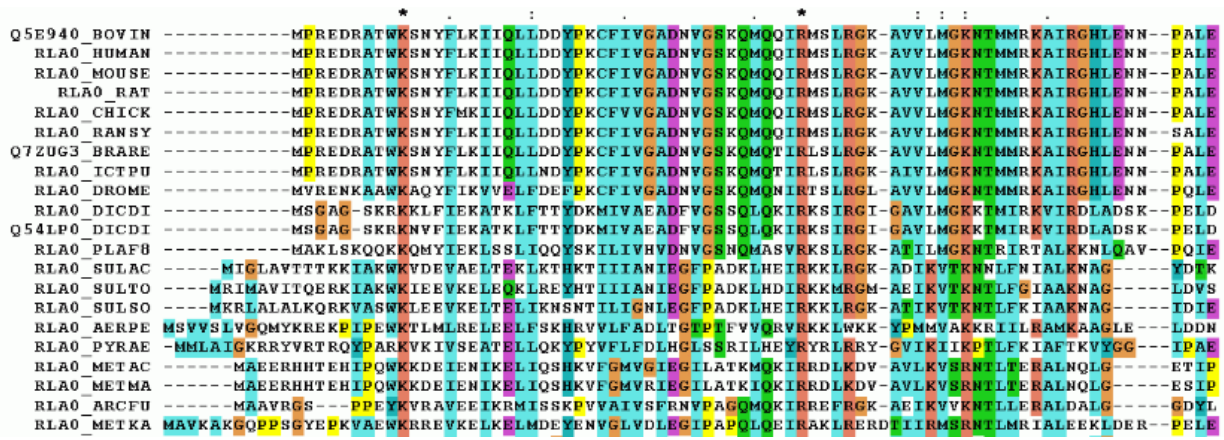
Mutation data

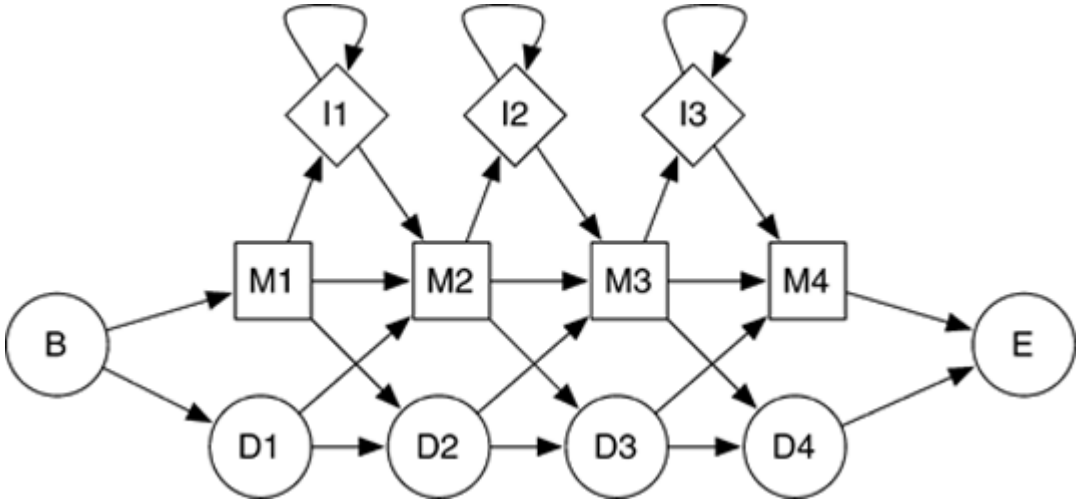


Mutation tree

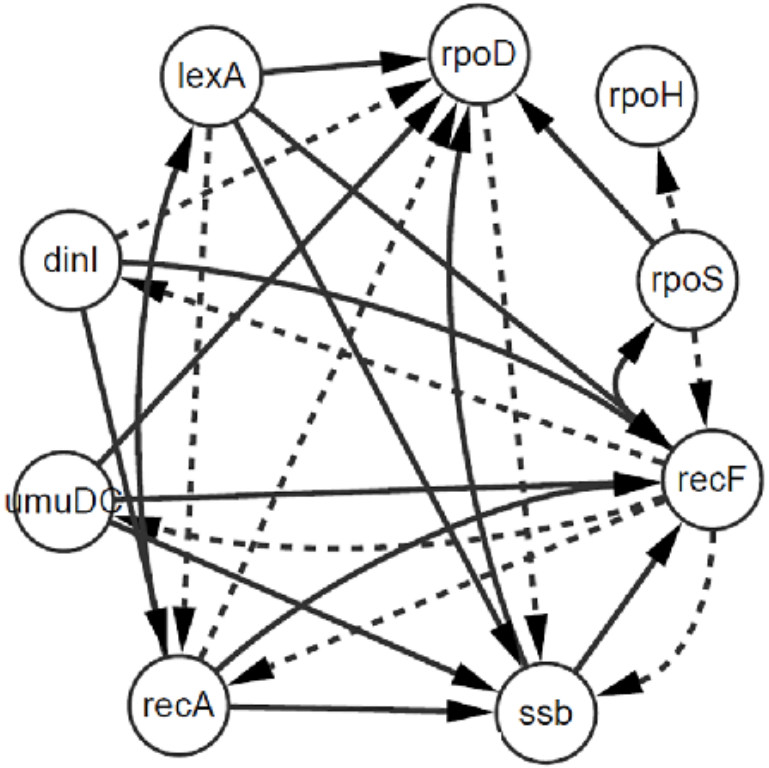
[Jahn et al, 2016]

Applications- multiple sequence alignment





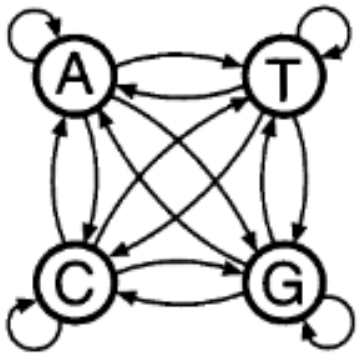
Applications- reconstruction of gene regulatory network



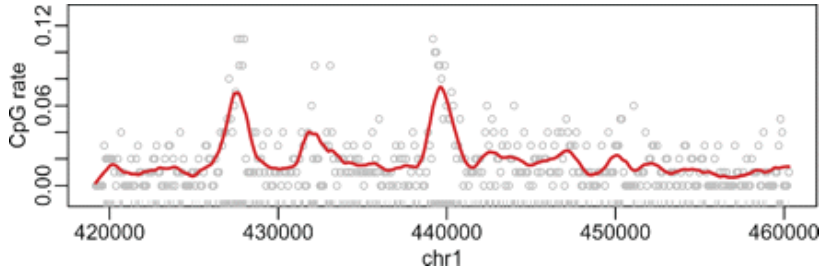
inferred SOS DNA repair network using dynamic Bayesian network [Liu et al, 2016]

Applications- genome annotation

- Methylation of **CpG islands** plays an important role in regulating gene expression.
- Identifying CpG islands by graphical models



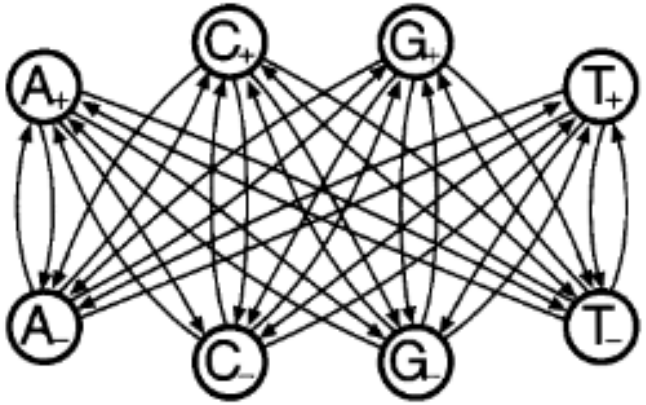
A Markov chain for DNA



[Wo et al, 2010]

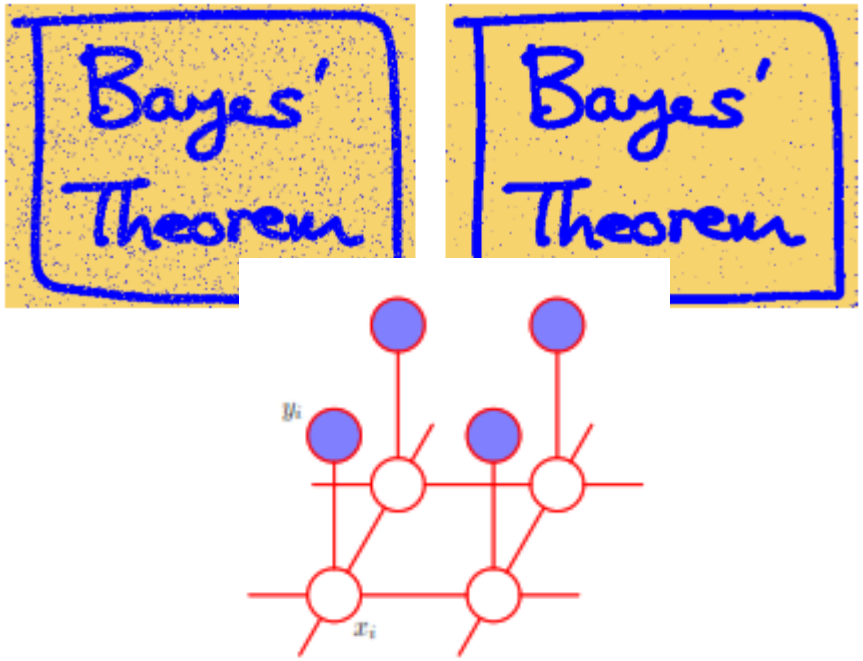
CpG Island

background



[Durbin, 2005]

Other applications



Markov random fields

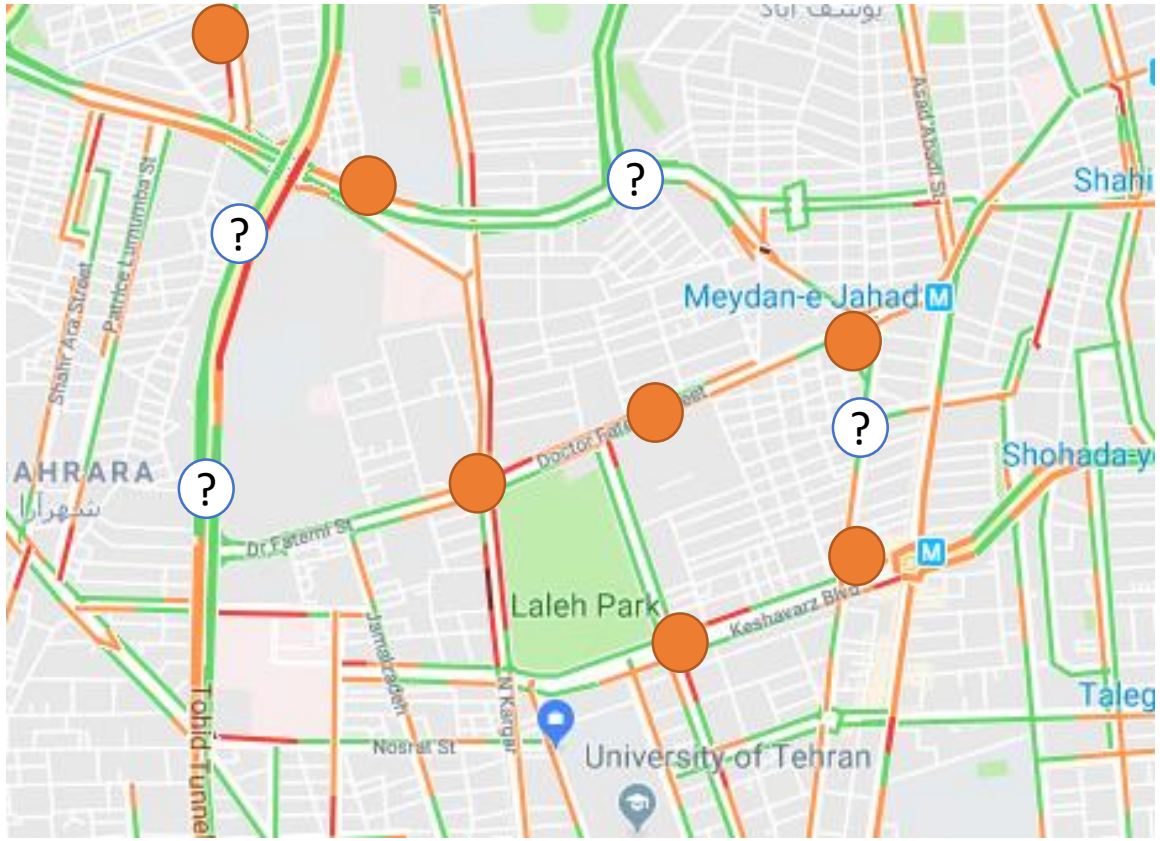
Image analysis

"Do you really like graphical models?"

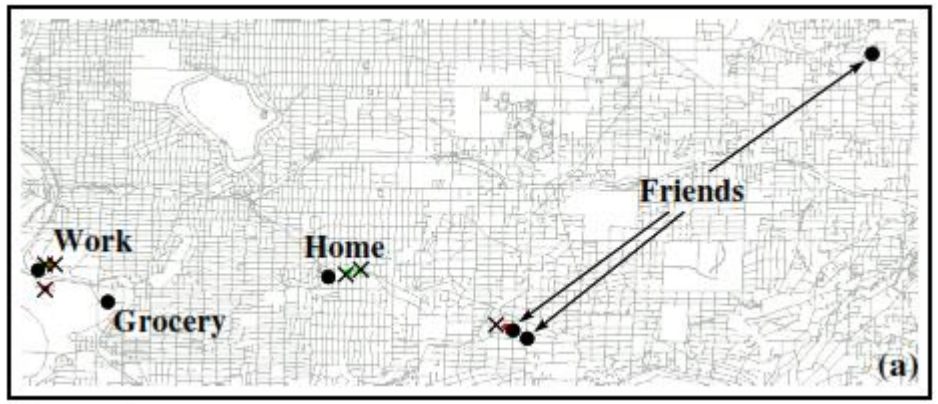
By hidden Markov models

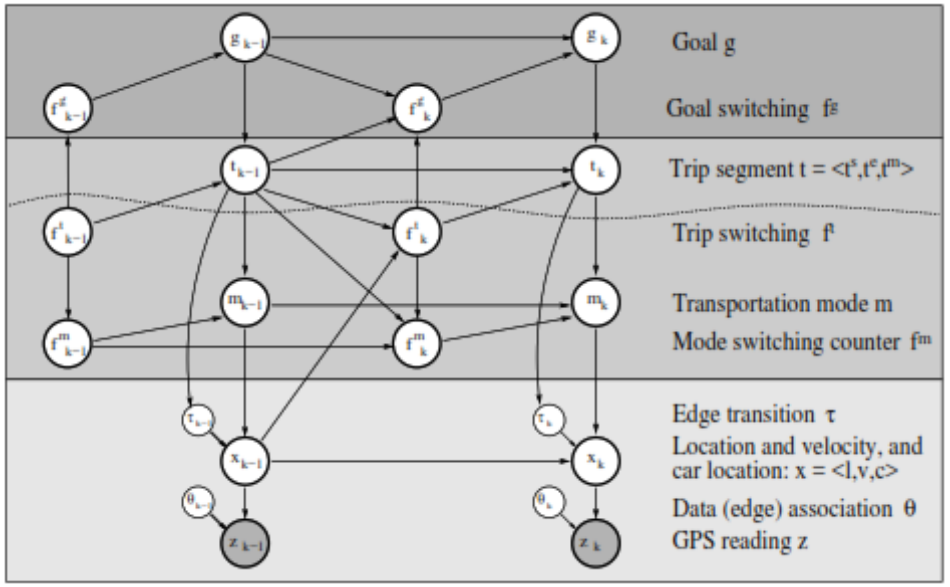
Speech recognition

Other applications-2



Estimating unmonitored road speeds from monitored speeds



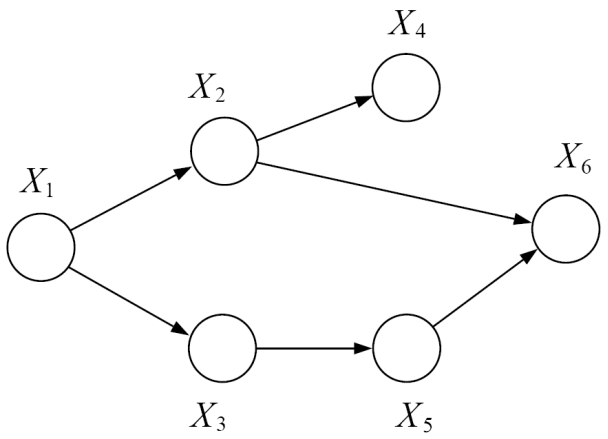


Hierarchical graphical model representing person's outdoor movement. Data: raw GPS coordinates

[L. Liao, D. Fox, and H. Kautz. AAAI-04]

Key ideas

- **Representation:** represent the problem as a collection of random variables $X_1, \dots, X_n$ with joint distribution $P(X_1, \dots, X_n)$ with some conditional independences.



- **Inference:** compute conditional probabilities given some evidences $P(X_i \mid E = e)$.

- **Learning:** estimate the parameters and structure of a Bayesian network from data

Recap: probability and statistics

Random variables

- *A random variable is a variable whose possible values (D) are outcomes (Ω) of a random phenomenon (from Wikipedia).*
- A random variable X is a mapping $X: \Omega \rightarrow D$.

- Example: a random variable *intelligence* that takes as values either *high* or *low*.

- Random variables can be either
  - Categorical or discrete
  - Or continuous

Examples

- Binary-valued random variable: “biased coin flips”
  D={H, T}
  P(X = H) = p. Hence, P(X=T) = 1-p

  The distribution of such a random variable is called a Bernoulli distribution and denoted by $X \sim \mathrm{Bernoulli}(p)$

- Categorical random variable: “*k*-sided dice”
  $D = \{1, \dots, k\}$
  $P(X = i) = p_i$ subject to $\sum_{i=1}^{k} p_i = 1$

  The distribution of such a random variable is called a multinomial distribution and denoted by $X \sim \mathrm{Mult}(p)$

Joint distribution

- An example of joint distribution P(intelligence, grade)

		Intelligence		
		low	*high*	
Grade	*A*	0.07	0.18	0.25
	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

- Marginal distribution:

$$P(G = A) = P(G = A, I = low) + P(G = A, I = high) = 0.07 + 0.18 = 0.25$$

- **Question:** suppose $X_1, \dots, X_n$ are binary-valued variables. How many parameters do we need to specify the joint distribution?
  - $2^n - 1$ parameters for 2^n different assignments of values $x_1, \dots, x_n$.

Marginal distributions

Discrete case: $$P(X) = \sum_{Y} P(X,Y), \quad P(Y) = \sum_{X} P(X,Y)$$

Continuous case: $$P(X) = \int_{Y} P(X,Y)\, dY, \quad P(Y) = \int_{X} P(X,Y)\, dX$$

- Given the joint distribution $P(X_1, \dots, X_n)$
- Then

$$\mathrm{P}(\mathrm{X_i} = \mathrm{x_i}) = \sum_{x_1,\dots,x_{i-1},x_{i+1},\dots,x_n} P(X_1 = x_1, \dots, X_n = x_n)$$

- **Question:** if all $\mathrm{X_i}$ binary: How many terms?

Conditional probabilities

- The conditional probability of *X* and *Y* is

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

- *G:* overexpression of an oncogene
- *C*: the presence of a tumor
- *P(G, C)*: Prob. of oncogene overexpression and the presence of a tumor.
- *P(G | C)*: Prob. of oncogene overexpression in cancer patients (can be assessed by counting).
- *P(C | G):* Prob. of cancer given gene expression measurement **(might be difficult to assess)**.

Bayes' rule (important)

- Since

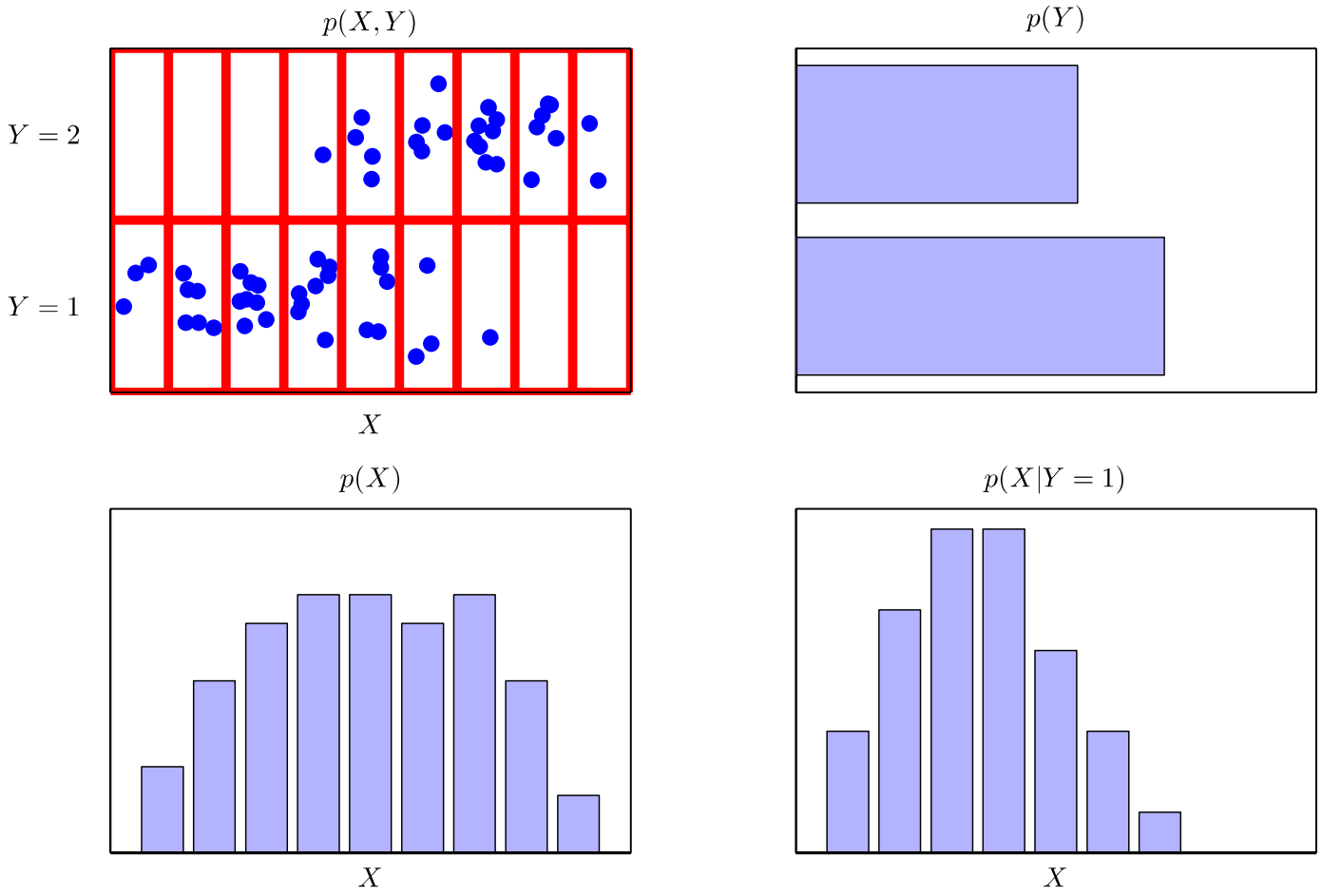
$$P(G,C) \;=\; P(G|C)P(C) \;=\; P(C|G)P(G)$$

- Hence

$$P(C|G) \;=\; \frac{P(G|C)P(C)}{P(G)}$$

- Hence, the diagnostic conditional probability P(C|G) can be computed without having to be determined **explicitly**.

Example: joint, marginal and conditional probabilities



Chain rule (important)

$$P(X_1, \dots, X_n) = P(X_1)P(\, X_2 \mid X_1 \,) \dots, P(X_n \mid X_1, \dots, X_{n-1})$$

Independent random variables

- For *n* independent random variables $X_1, \dots, X_n$ we have

$$P(X_1, \dots, X_n) = P(X_1) \dots P(X_n)$$

- **Question:** how many parameters in this case?

- Something in between full dependence and full independence assumptions?

Conditional independence

- It is not often that we encounter two independent random variables. A more common situation is when two random variables are independent given an additional random variable.

- Proposition: $X \perp Y \mid Z \quad \Leftrightarrow \quad P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

- An alternative way to prove conditional independence of X and Y given Z

$$P(X \mid Y, Z) = P(X \mid Z)$$

Why conditional independence is useful?

- By chain rule

$$P(X_1, \dots, X_n) = P(X_1)P(X_2 \mid X_1) \dots, P(X_n \mid X_1, \dots, X_{n-1})$$

Question: how many parameters?

- Suppose $X_i \perp X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n \mid X_1$ for all $i \in \{2, \dots, n\}$. Then

$$P(X_1, \dots, X_n) = ?$$

Why conditional independence is useful?

- By chain rule

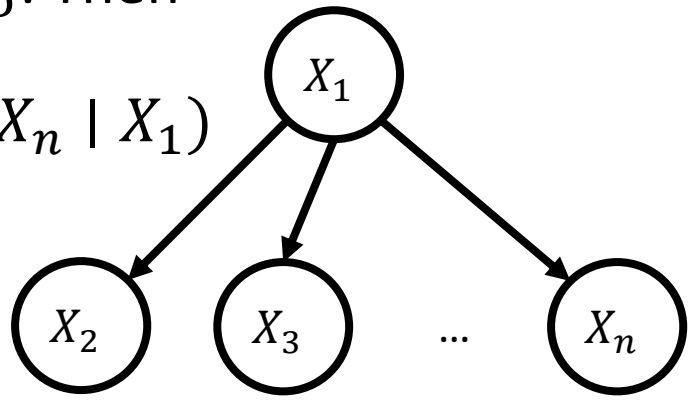
$$P(X_1, \dots, X_n) = P(X_1)P(X_2 \mid X_1) \dots, P(X_n \mid X_1, \dots, X_{n-1})$$

Question: how many parameters?

- Suppose $X_i \perp X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n \mid X_1$ for all $i \in \{2, \dots, n\}$. Then

$$P(X_1, \dots, X_n) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1) \dots, P(X_n \mid X_1)$$

Question: how many parameters?



Querying a distribution

- Probability queries
  - **Evidence**: a subset *E* of random variables in the model, and an instantiation *e* to these variables.
  - **Query variables:** a subset Y of random variables.
  - Query: $P(Y \mid E = e)$
  - **Question:** $\mathrm{P}(\, G \mid i^0 \,)$?

- Map queries:
  - To find the most likely assignment to the variables W given the evidence E=e

    $$MAP(W|e) = \arg\max_{w} P(w \mid e)$$
  - **Question:** $\mathrm{MAP}(D, G \mid i^1)$?

I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^0	g^2	0.168
i^0	d^0	g^3	0.126
i^0	d^1	g^1	0.009
i^0	d^1	g^2	0.045
i^0	d^1	g^3	0.126
i^1	d^0	g^1	0.252
i^1	d^0	g^2	0.0224
i^1	d^0	g^3	0.0056
i^1	d^1	g^1	0.06
i^1	d^1	g^2	0.036
i^1	d^1	g^3	0.024

Continuous random variables

- Probability density function *f* for continuous random variable X

$$\int_x f(x)dx = 1$$

- Cumulative distribution *P*

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

- Similarly for multiple random variables

$$P(a_1 \leq X_1 \leq b_1, \dots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \dots \int_{a_n}^{b_n} f(x_1, \dots, x_n)dx_1 \dots dx_n$$

Project 1

- Application of Bayesian networks on gene expression data

- Deadline: Esfand 19, 1397 (March 9, 2019)

Coursework

- **Grading based on**
  - Data analysis projects: 25% (two students per group)
  - Problem sets: 25%
  - Reading assignments: 10%
  - Final exam: 40%

- **Reading assignments:**
  - Deadline: the day before the lecture by 17:00.

- **Late submission policy**
  - In total, 5 late days for problem sets. 3 late days for projects. 6 late days for reading assignments.
  - No further extension. Zero point for submitting assignments after the deadline.
  - Regardless, in order to pass the course **you have to submit all assignments.**

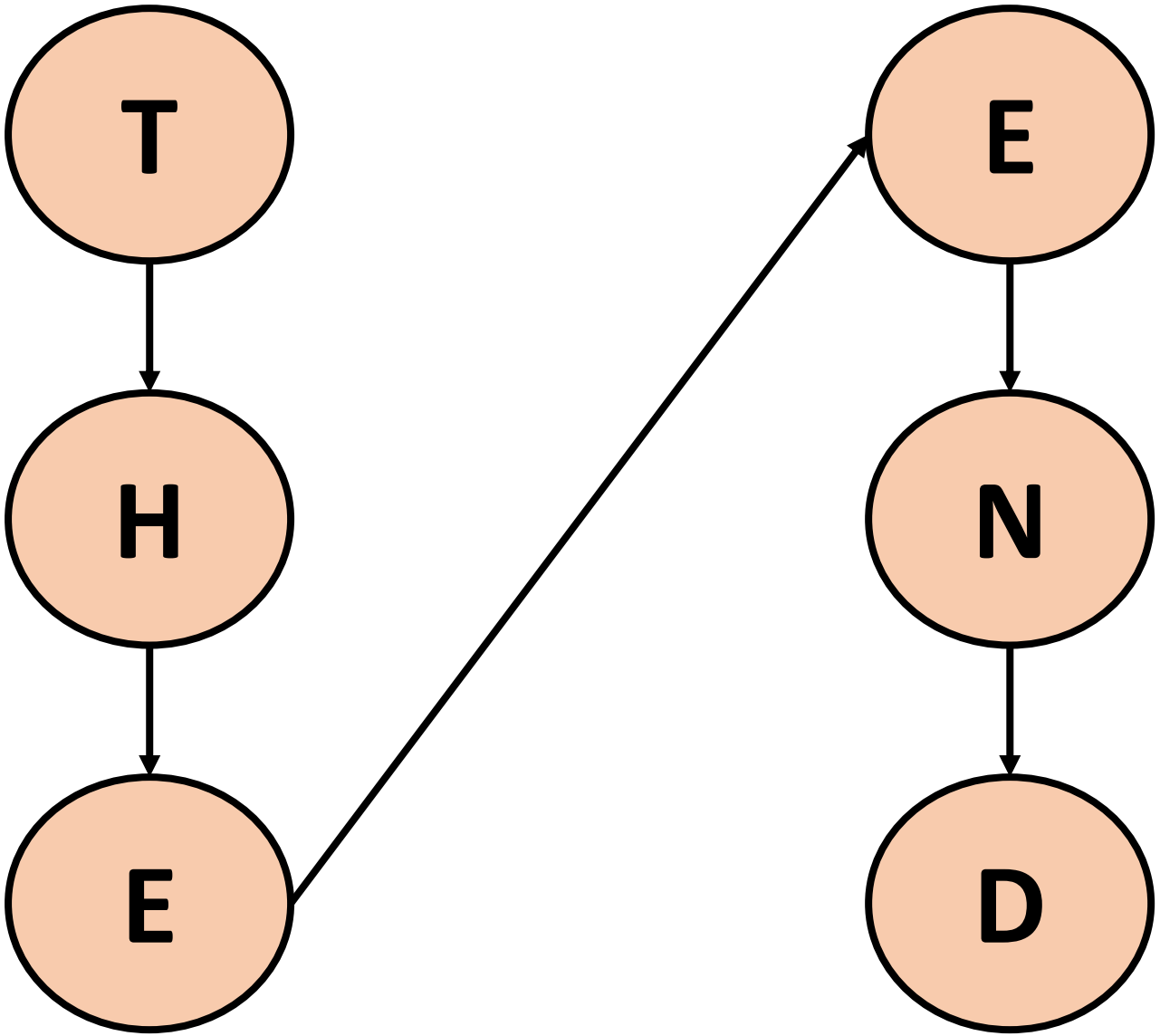
- Discussing assignments is encouraged, but you must turn in your own solutions.

Bibliography

1. Daphne Koller and Nir Friedman. Probabilistic graphical models: principles and techniques, 2011, MIT press.

2. Dirk Husmeier. Probabilistic Modeling in Bioinformatics and Medical Informatics, 2010, Springer.

3. Christopher M. Bishop. Pattern Recognition and Machine Learning, 2016, Springer.


T
H
E
E
N
D