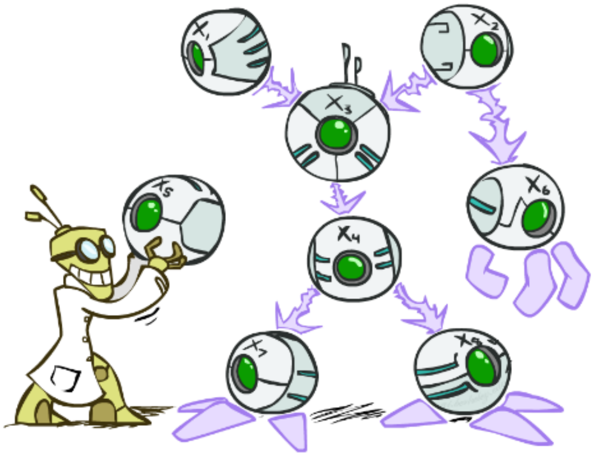
Announcement

- Midterm @March. 29 (in class)
  - Location: TBA
  - Format
    - Closed-book. You can bring an A4-size cheat sheet and nothing else.
    - Around 5 problems
  - Grade
    - 25% of the total grade

Bayesian Networks



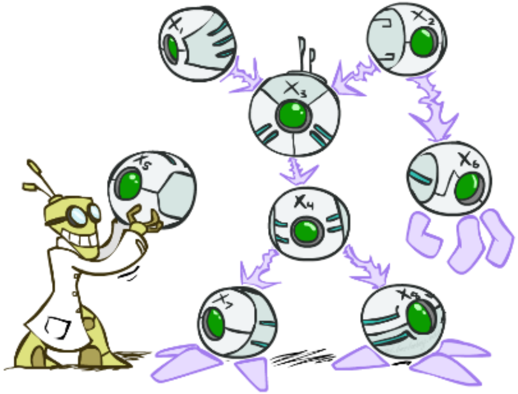
AIMA Chapter 14.1, 14.2, PRML Chapter 8

[Adapted from slides by Dan Klein and Pieter Abbeel at UC Berkeley (ai.berkeley.edu)
and from slides by Markus Svensén & Christopher M Bishop]

(rew)

Bayesian Networks: Big Picture

- Full joint distribution tables answer every question, but:
  - Size is exponential in the number of variables
  - Need gazillions of examples to learn the probabilities
  - Inference by enumeration (summing out hiddens) is too slow
- Bayesian networks:
  - Express all the conditional independence relationships in a domain
  - Factor the joint distribution into a product of small conditionals
  - Often reduce size from exponential to linear
  - Faster learning from fewer examples
  - Faster inference (linear time in some important cases)



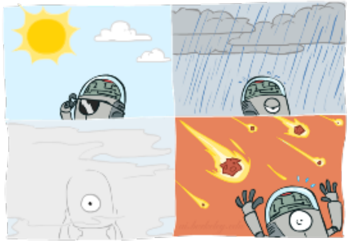
Bayesian Networks Syntax

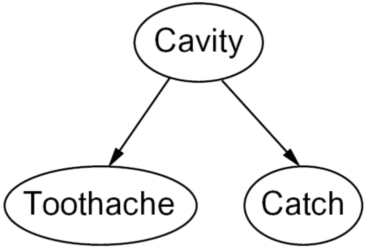
Bayesian Networks Syntax

- Nodes: variables (with domains)

- Arcs: interactions
  - Indicate “direct influence” between variables
  - For now: imagine that arrows mean direct causation (in general, they may not!)
  - Formally: encode conditional independence (more later)

- No cycle is allowed!

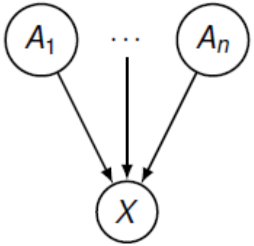




Bayesian Networks Syntax

- A directed, acyclic graph

- Conditional distributions for each node given its ***parent variables*** in the graph
  - ***CPT***: conditional probability table: each row is a distribution for child given a configuration of its parents
  - Description of a noisy “causal” process



$$P(X|A_1, \cdots, A_n)$$

A Bayes net = Topology (graph) + Local Conditional Probabilities

General formula for sparse BNs

- Suppose
  - n variables
  - Maximum domain size is d
  - Maximum number of parents is k
- Full joint distribution has size $O(d^n)$
- Bayes net has size $O(n \cdot d^{k+1})$
  - Linear scaling with n as long as causal structure is local

Bayesian Networks Semantics

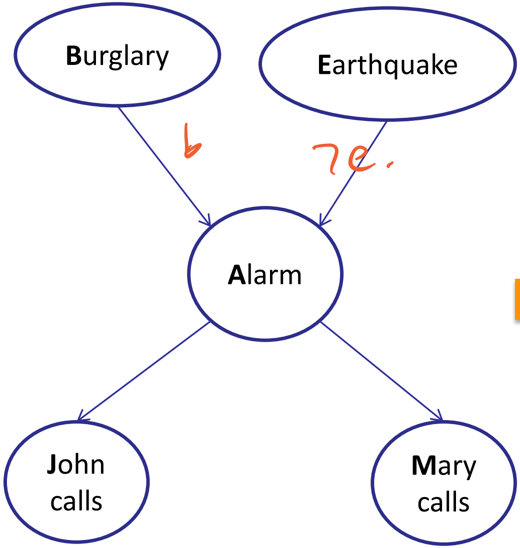
Bayesian networks global semantics

- Bayes nets encode joint distributions as product of conditional distributions on each variable:

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

Example

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

$P(b, \neg e, a, \neg j, \neg m) =$

$P(b)\ P(\neg e)\ P(a|b, \neg e)\ P(\neg j|a)\ P(\neg m|a)$

$= .001 \text{x} .998 \text{x} .94 \text{x} .1 \text{x} .3 = .000028$

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

A	P(M|A)	
	true	false
true	0.7	0.3
false	0.01	0.99

Probabilities in BNs

- Why are we guaranteed that setting

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

  results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$

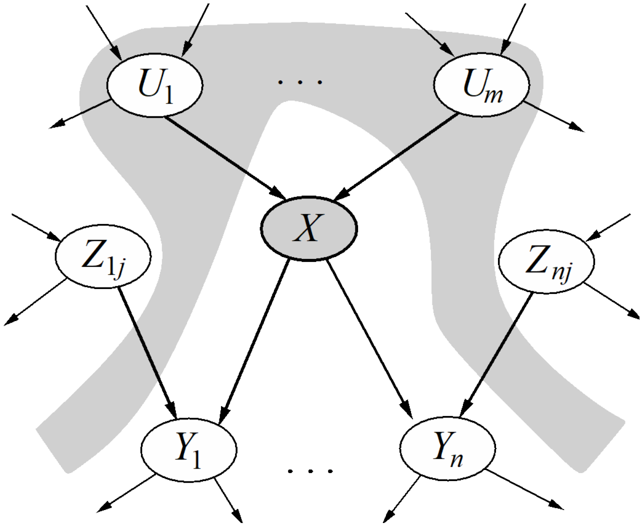
  i i-1

- Assume conditional independences: $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$
  - When adding node X_i, ensure parents "shield" it from other predecessors (no loop)

- So the network topology implies that certain conditional independencies hold

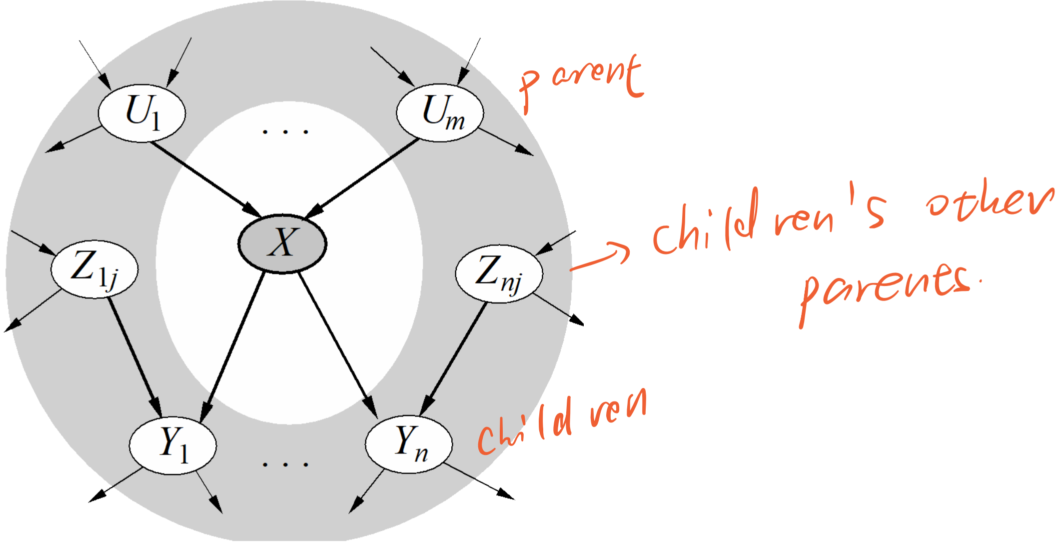
Conditional independence semantics

- ***Every variable is conditionally independent of its non-descendants given its parents***
- Conditional independence semantics <=> global semantics



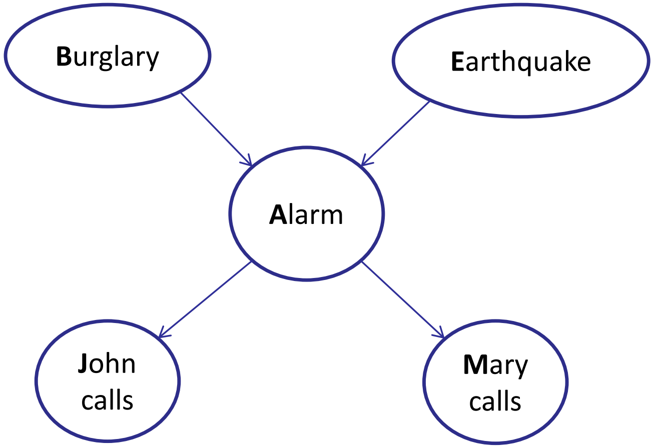
Markov blanket

- A variable's Markov blanket consists of parents, children, children's other parents
- ***Every variable is conditionally independent of all other variables given its Markov blanket***



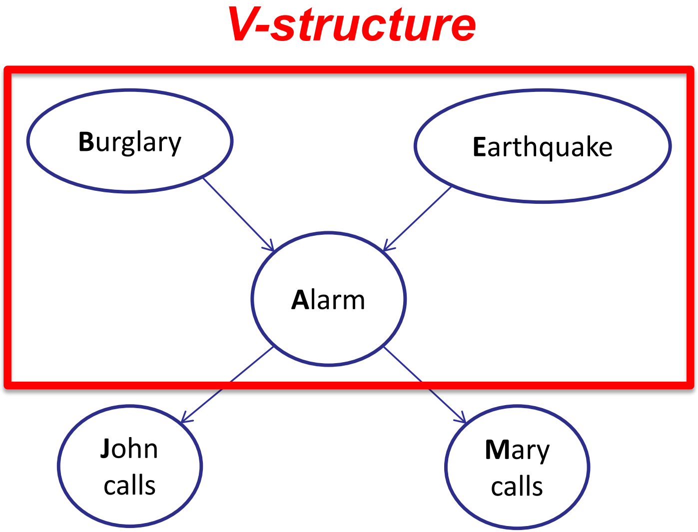
Example

- JohnCalls independent of Burglary given Alarm?
  - Yes
- JohnCalls independent of MaryCalls given Alarm?
  - Yes
- Burglary independent of Earthquake?
  - Yes

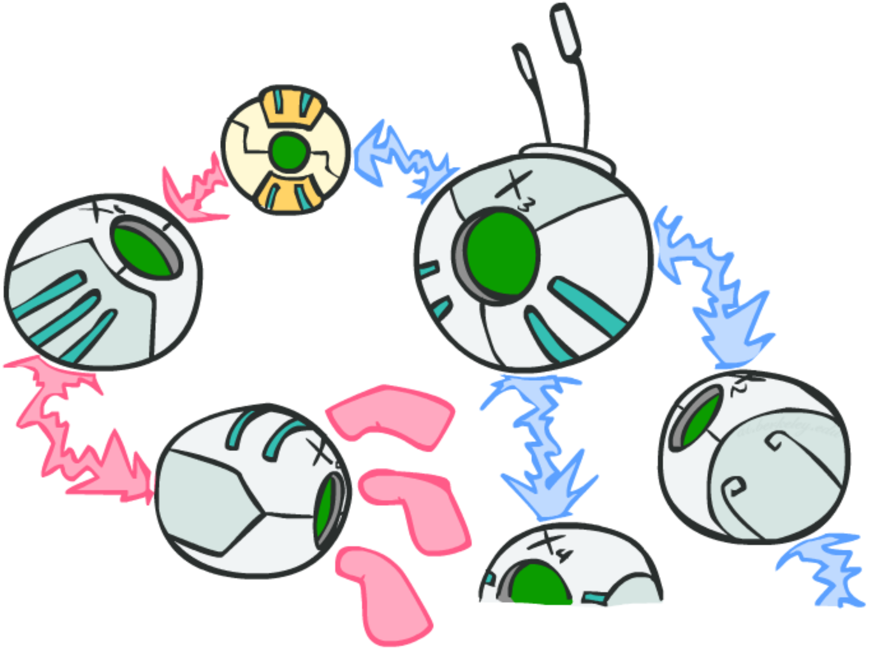


Example

- Burglary independent of Earthquake given Alarm?
  - NO!
  - Given that the alarm has sounded, both burglary and earthquake become more likely
  - But if we then learn that a burglary has happened, the alarm is ***explained away*** and the probability of earthquake drops back
- Burglary independent of Earthquake given JohnCalls?
- Any simple algorithm to determine conditional independence?



D-separation



D-separation: Outline

- Study independence properties for triples
- Analyze complex cases in terms of member triples
- D-separation: a condition / algorithm for answering such queries

Causal Chains

- This configuration is a "causal chain"

X: Low pressure   Y: Rain   Z: Traffic

Global semantics:

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z ? *No!*
- Guaranteed X independent of Z given Y?

$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)}$$

$$= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y)$$

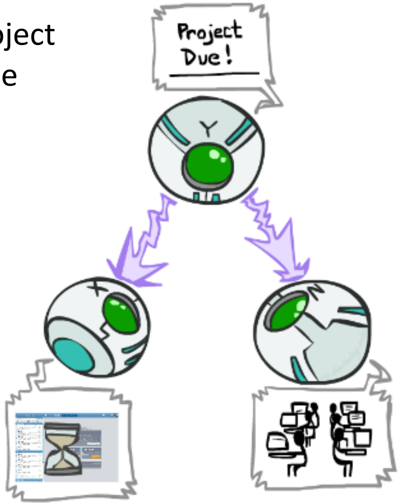
Yes!

- Evidence along the chain "blocks" the influence

Common Cause

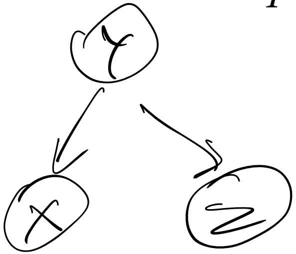
- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full



- Guaranteed X independent of Z ? *No!*
- Guaranteed X and Z independent given Y?

$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)}$$

$$= \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}$$

$$= P(z|y)$$

Yes!

- Observing the cause blocks influence between effects.

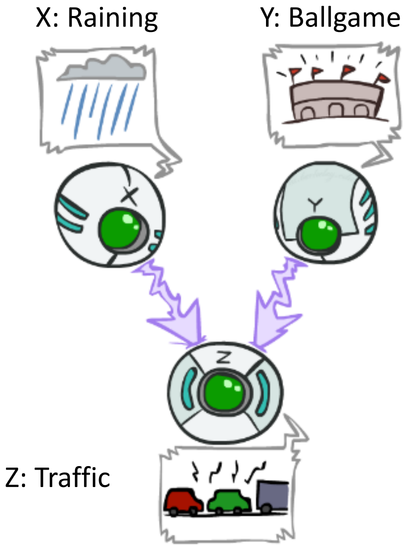
Global semantics:

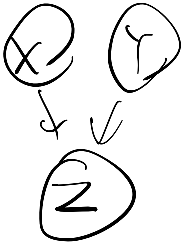
$$P(x,y,z) = P(y)P(x|y)P(z|y)$$

$P(y)P(x|y) \cdot P(y) \cdot P(z|y)$

Common Effect (v-structure)

- Last configuration: two causes of one effect (v-structures)





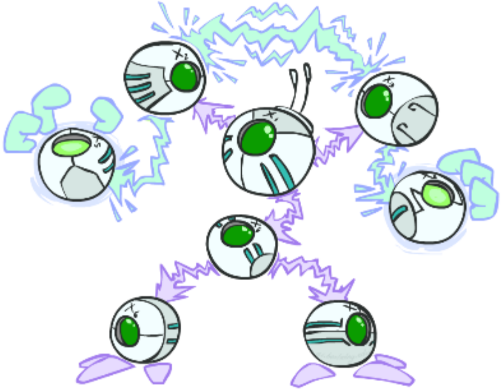
- Are X and Y independent?
  - *Yes*: the ballgame and the rain cause traffic, but they are not correlated
  - Still need to prove they must be (try it!)
- Are X and Y independent given Z?
  - *No*: seeing traffic puts the rain and the ballgame in competition as explanation.
- This is backwards from the other cases
  - Observing an effect activates influence between possible causes.

D-separation - the General Case

D-separation - the General Case

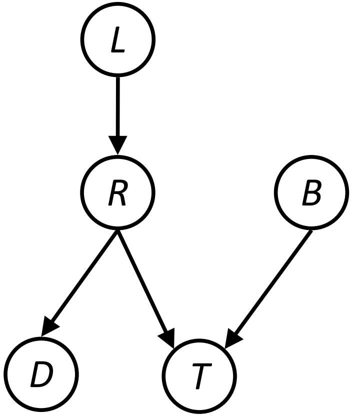
- General question: in a given BN, are two variables independent (given evidence)?

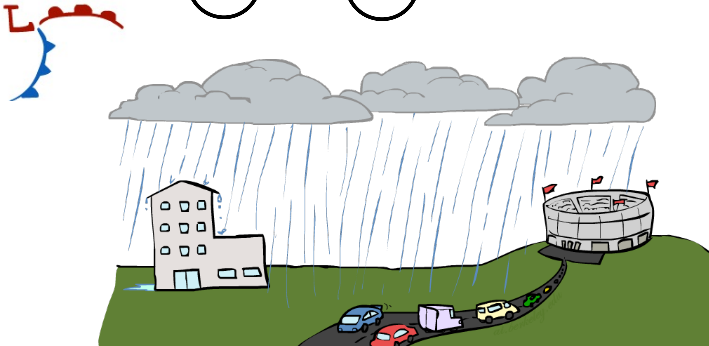
- Solution: analyze the graph; break the question into repetitions of the three canonical cases



Reachability

- Recipe: shade evidence nodes, look for paths in the resulting graph
- Attempt 1: if two nodes are **not** connected by any undirected path not blocked by a shaded node, they are conditionally independent
- Almost works, but not quite
  - Where does it break?
  - Answer: the v-structure at T doesn't count as a link in a path unless "active"





Active / Inactive Paths

重叠

- Question: X, Y, Z are non-intersecting subsets of nodes. Are X and Y conditionally independent given Z?

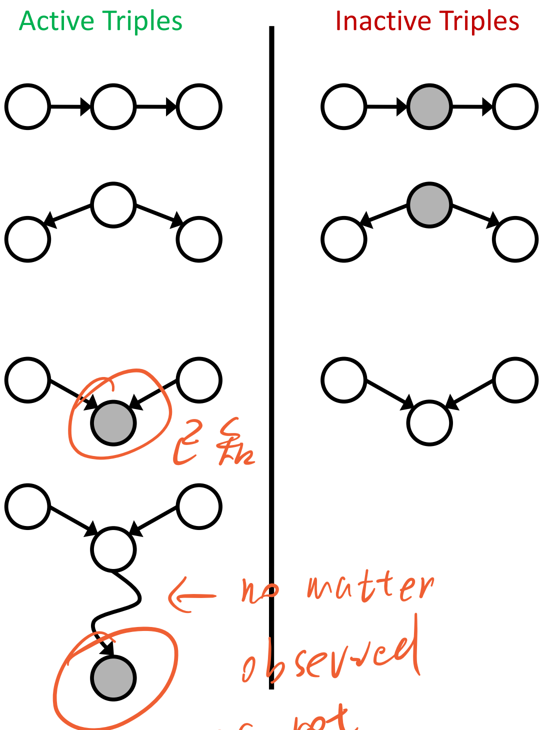
- A triple is active in the following three cases
  - Causal chain $A \to B \to C$ where B is unobserved (either direction)
  - Common cause $A \leftarrow B \to C$ where B is unobserved
  - Common effect (aka v-structure) $A \to B \leftarrow C$ where B *or one of its descendents* is observed

- A path is active if each triple along the path is active
- A path is blocked if it contains a single inactive triple

$X \to Y$

- If all paths from X to Y are blocked, then X is said to be "**d-separated**" from Y by Z
- If d-separated, then X and Y are conditionally independent given Z

directed!!



D-Separation

- Query: $X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$?

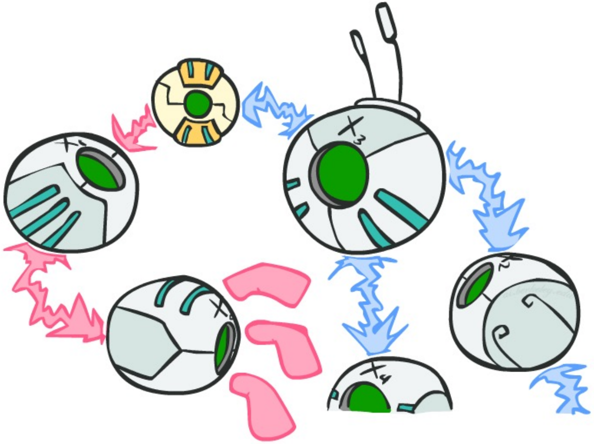
已知这些时

- Check all (undirected!) paths between X_i and X_j
  - If one or more active, then independence not guaranteed

$$X_i \not\perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

  - Otherwise (i.e. if all paths are inactive), then independence is guaranteed

$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$



all inactive

Example

find path triple on path.

$R \perp\!\!\!\perp B$ Yes

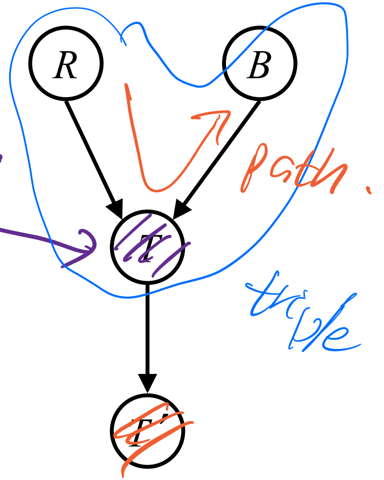
if blocked (inde?)

$R \perp\!\!\!\perp B | T$ observed

R, B, T inactive

no.

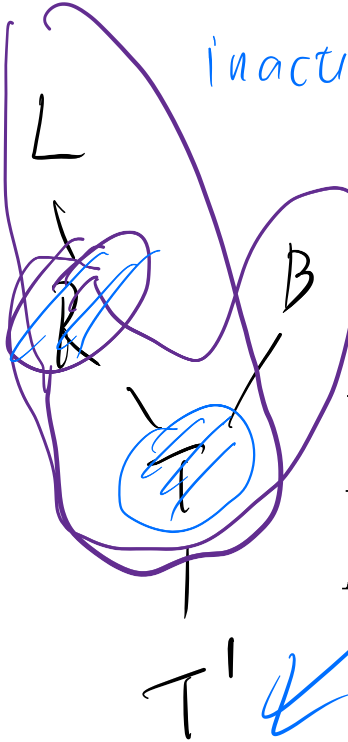
$R \perp\!\!\!\perp B | T'$



Example

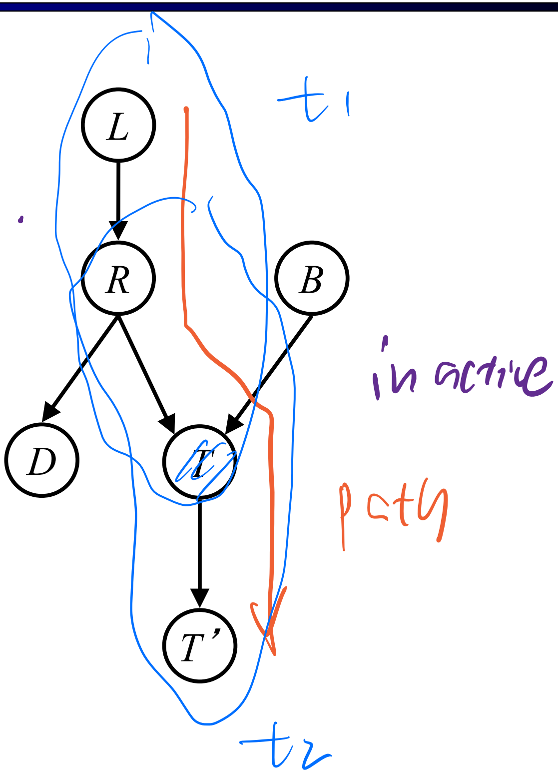
triple

inactive



$L \perp\!\!\!\perp T' \mid T$	Yes
$L \perp\!\!\!\perp B$	Yes
$L \perp\!\!\!\perp B \mid T$	
$L \perp\!\!\!\perp B \mid T'$	
$L \perp\!\!\!\perp B \mid T, R$	Yes

inactive blocks.



Example

- Variables:
  - R: Raining
  - T: Traffic
  - D: Roof drips
  - S: I'm sad
- Questions:

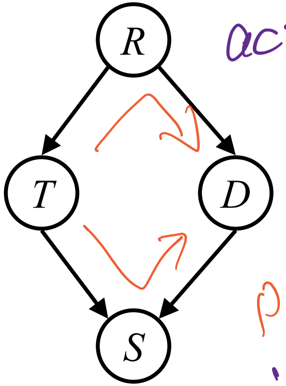
$$T \perp\!\!\!\perp D$$ both path inactive

$$T \perp\!\!\!\perp D | R$$ Yes

$$T \perp\!\!\!\perp D | R, S$$

$T \perp\!\!\!\perp D$:

path 1: active all



path 2: inactive

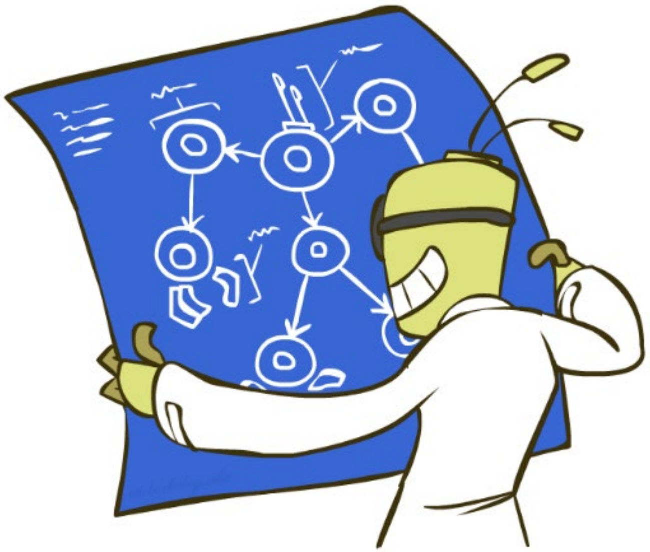
Structure Implications

- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

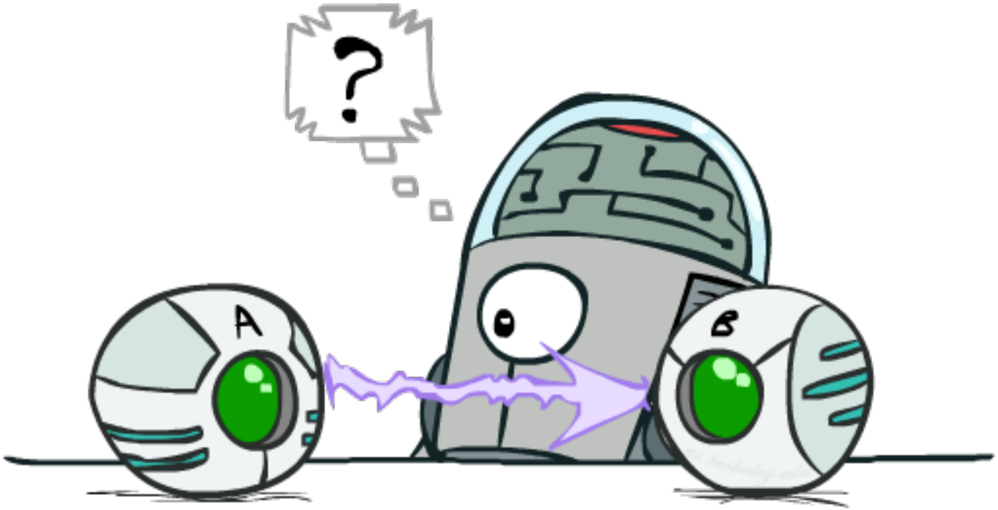
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

- Conditional independence semantics <=> global semantics

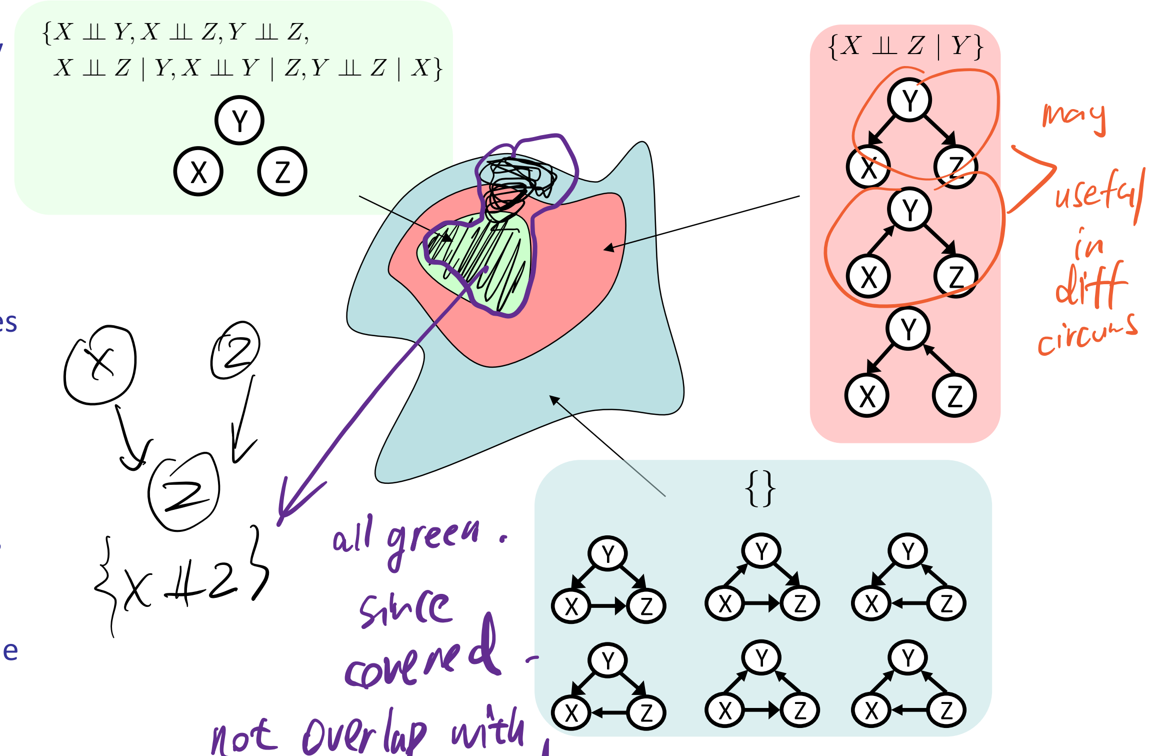


Node Ordering



Topology Limits Distributions

- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences
- (There might be more independence)
- Adding arcs increases the set of distributions, but has several costs
- Full conditioning can encode any distribution



need

Bayes Nets Representation Summary

- Bayes nets compactly encode joint distributions (by making use of conditional independences!)

- Guaranteed independencies of distributions can be deduced from BN graph structure

- D-separation gives precise conditional independence guarantees from graph alone

- A Bayes net's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution

related with value

d-separation not

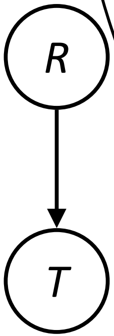
Bayes Nets

Some special value may cause extra independence

- Representation
- Probabilistic Inference
  - Enumeration (exact, exponential complexity)
  - Variable elimination (exact, worst-case exponential complexity, often better)
  - Probabilistic inference is NP-complete
- Conditional Independences
- Sampling
- Learning from data

Example: Traffic

- Causal direction



$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

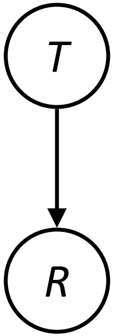
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T,R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

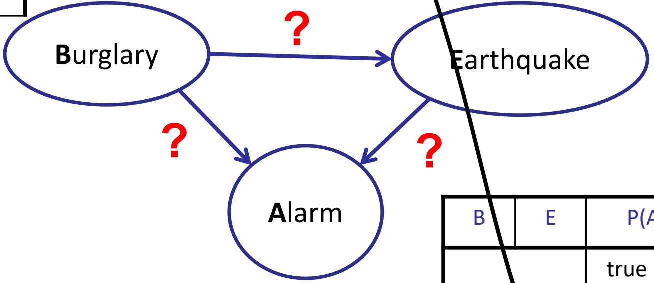
$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

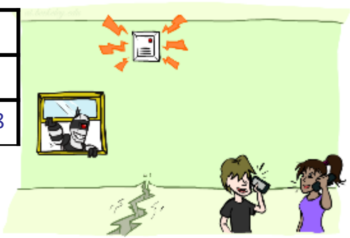
Example: Burglary

- Burglary
- Earthquake
- Alarm

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998



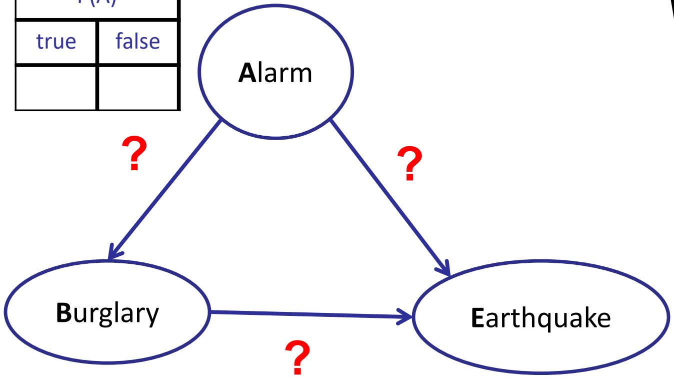
B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

2 edges, 6 free parameters

Example: Burglary

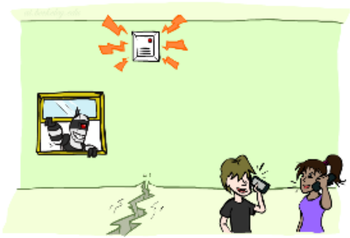
- Alarm
- Burglary
- Earthquake

P(A)	
true	false



A	P(B|A)	
	true	false
true		
false		

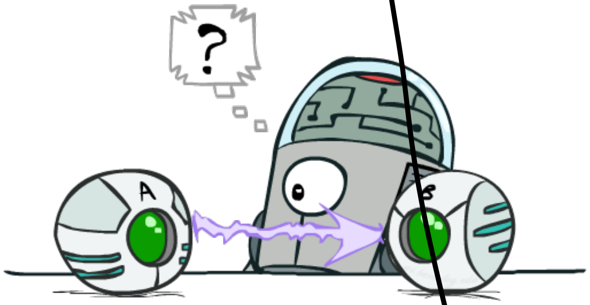
A	B	P(E|A,B)	
		true	false
true	true		
true	false		
false	true		
false	false		



3 edges, 7 free parameters

Causality?

- When Bayes nets reflect the true causal patterns:
  (e.g., Burglary, Earthquake, Alarm)
  - Often simpler (fewer parents, fewer parameters)
  - Often easier to assess probabilities
  - Often more robust: e.g., changes in frequency of burglaries should not affect the rest of the model!



- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Umbrella*
  - End up with arrows that reflect correlation, not causation

- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence:
    $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$

Example Application: Topic Modeling

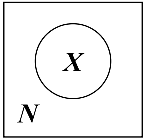
[Containing some materials from slides by David Blei at Princeton and by Hongning Wang at UVa]

Introduction

- A large body of text available online
  - It is difficult to find and discover what we need.
- Topic models
  - Approaches to discovering the main themes of a large unstructured collection of documents
  - Can be used to automatically organize, understand, search, and summarize large electronic archives
  - Latent Dirichlet Allocation (LDA) is the most popular

Plate Notation

- Representation of repeated subgraphs in a Bayesian network



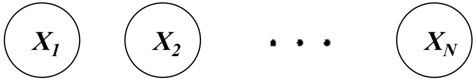


Plate Notation

- Representation of repeated subgraphs in a Bayesian network

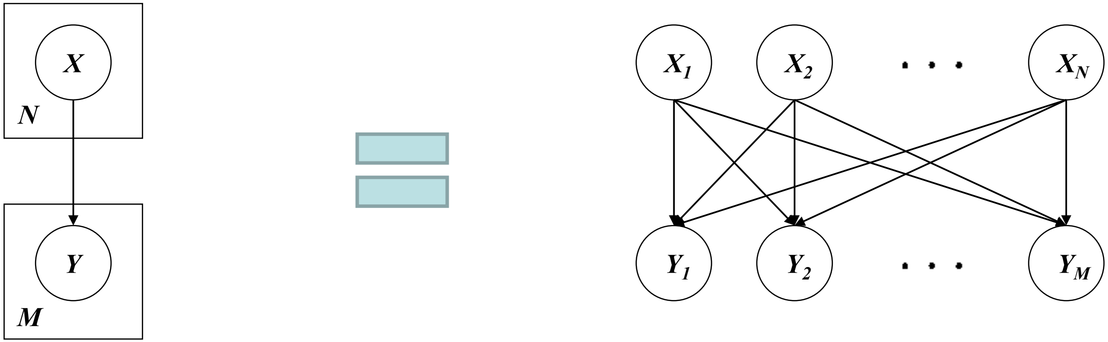
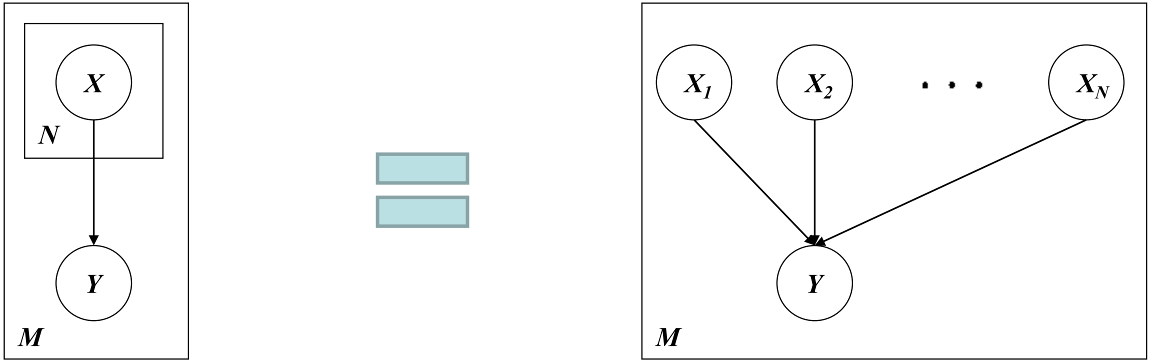
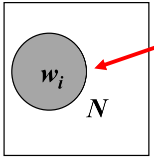


Plate Notation

- Representation of repeated subgraphs in a Bayesian network



How to generate a document



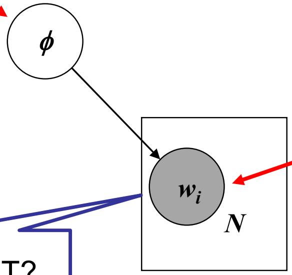
word ***e.g.*****, “how”, “to”, “get”, …**

How to generate a document

distribution over words

e.g., [0.1, 0.3, 0.2, …]

how	0.1
to	0.3
get	0.2
…	…



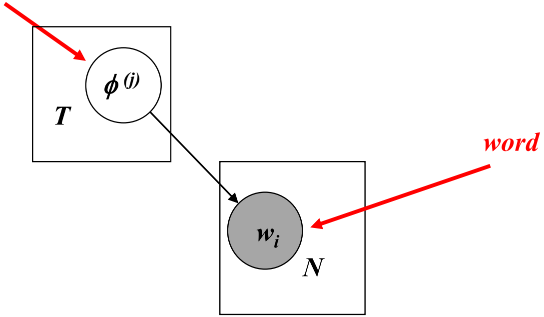
word

What is the CPT?

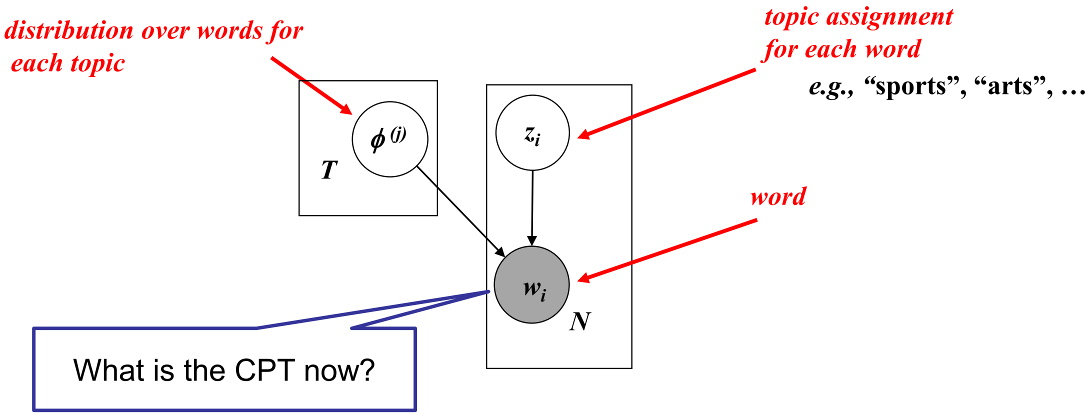
How to generate a document

distribution over words for each topic

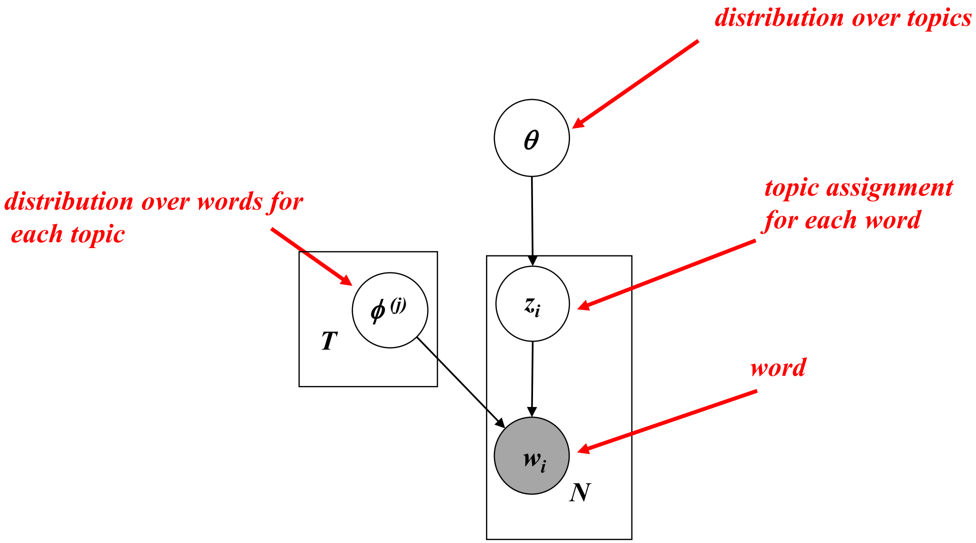
e.g., **sports: [0.1, 0.3, 0.2, …]**
arts: [0.2, 0.3, 0.1, …]
science: [0.3, 0.4, 0.1, …]
……



How to generate a document



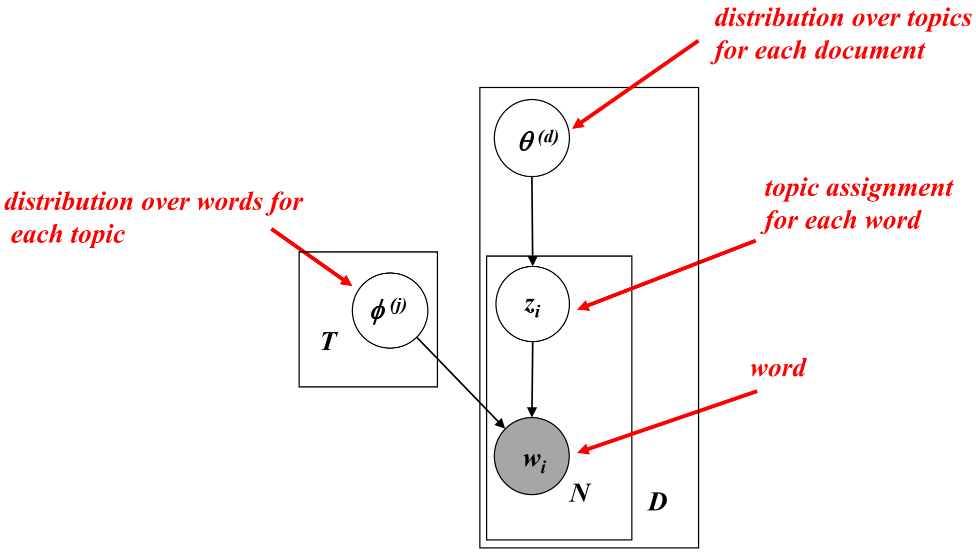
How to generate a document



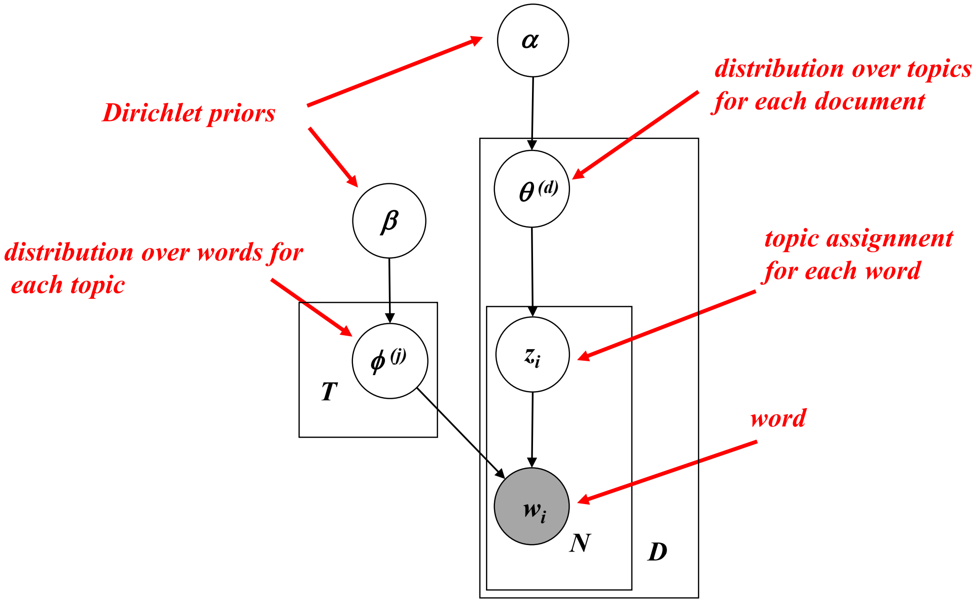
e.g., **[0.2, 0.3, 0.2, …]**

sports	**0.1**
arts	**0.3**
science	**0.2**
…	**…**

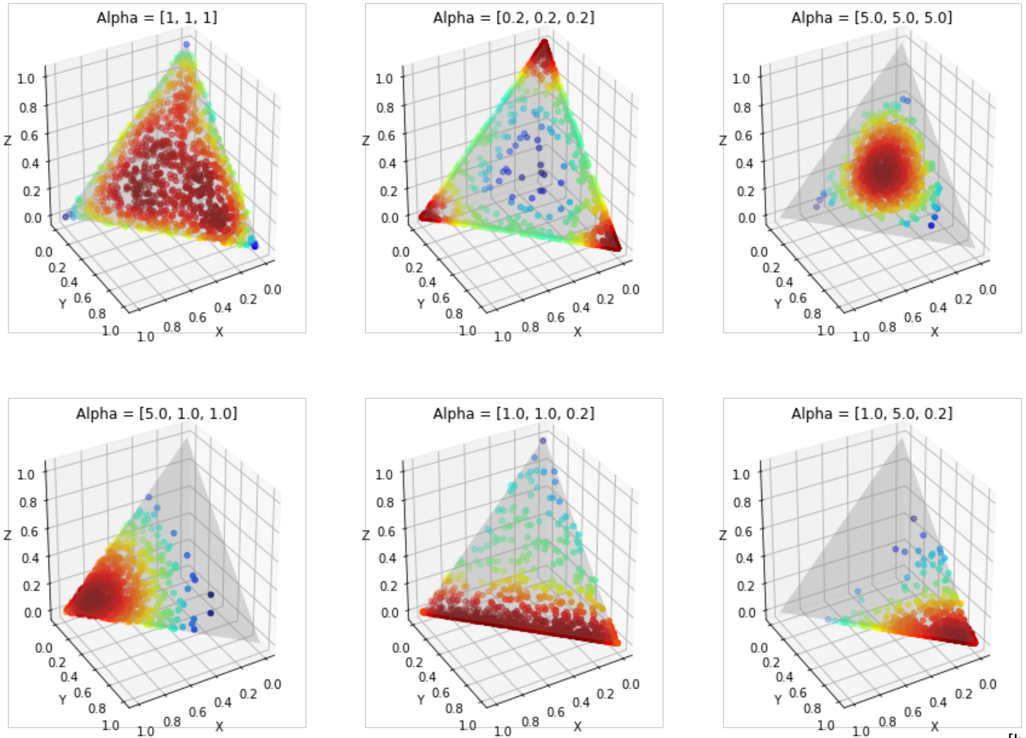
How to generate documents



How to generate documents

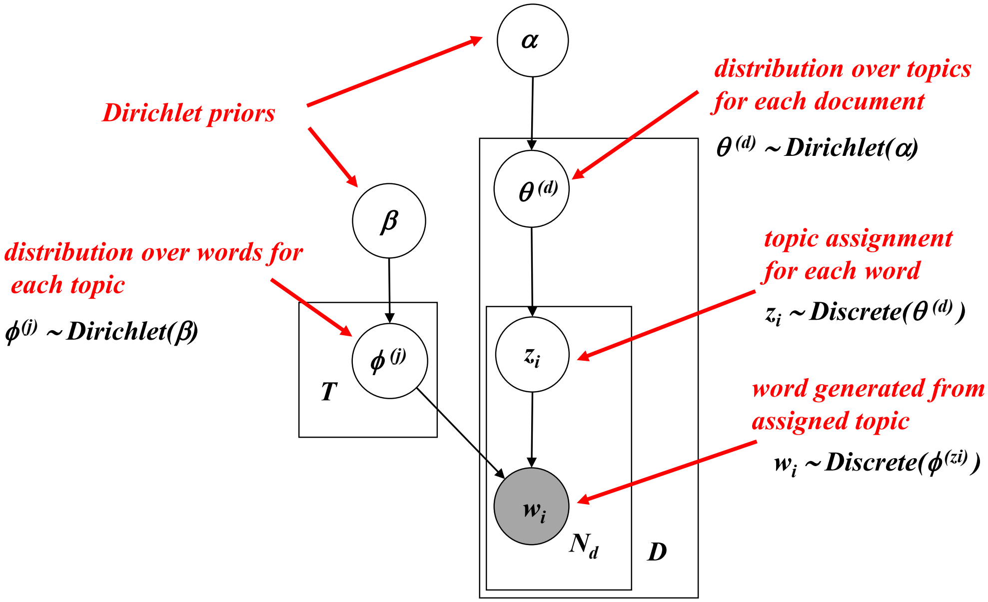


Dirichlet Distribution

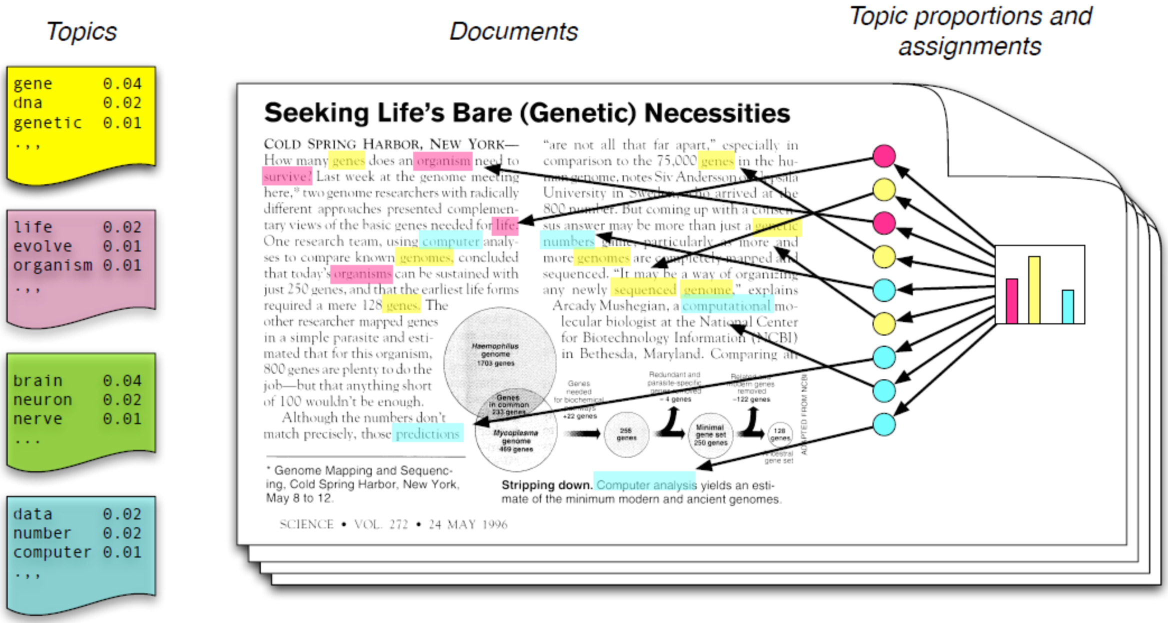


[by Thushan Ganegedara at USydney]

Latent Dirichlet Allocation (LDA)

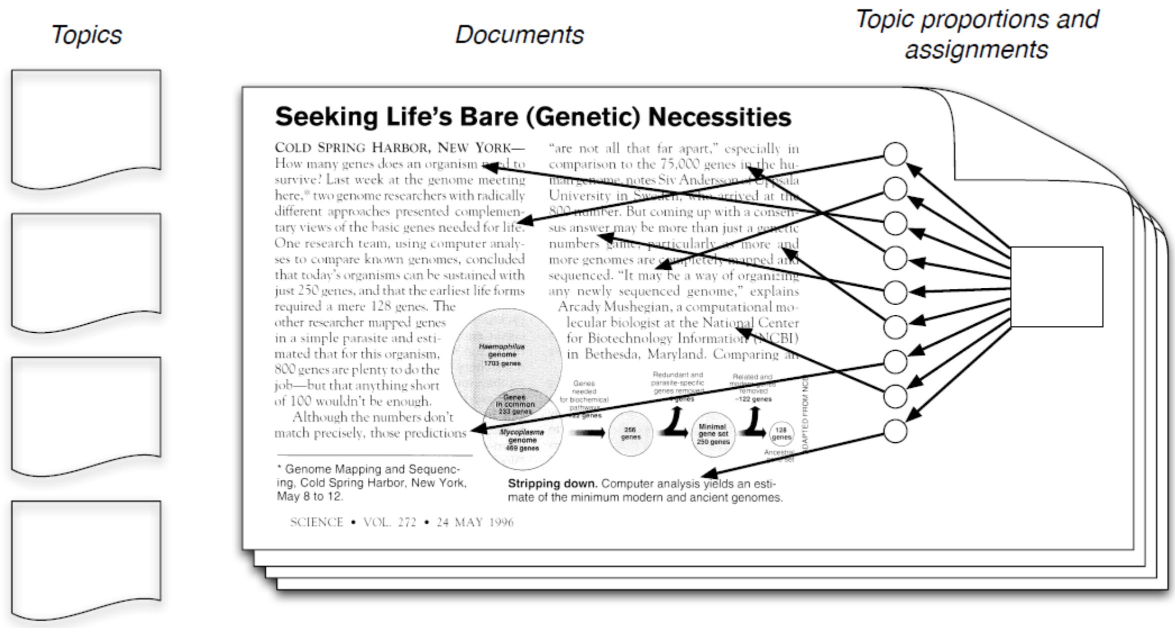


Illustration



- Each **topic** is a distribution of words; each **document** is a mixture of corpus-wide topics; and each **word** is drawn from one of those topics.

Illustration



- In reality, we only observe documents. The other structures are hidden variables that must be inferred. (We will discuss inference later.)

Topics inferred by LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI