Graphical Model (I)

Wang Houfeng
Institute of Computational Linguistics
Peking University

Outline

- **Introduction**
- Directed Graph(BN)
- Undirected Graphical Model(MRF)

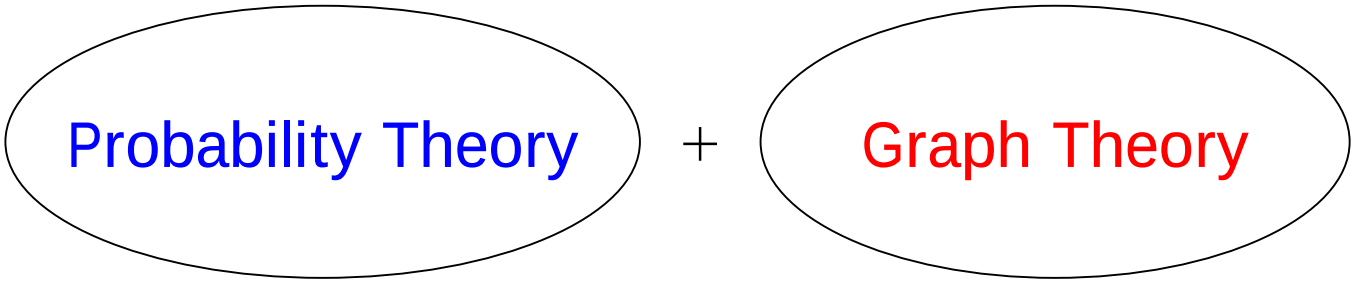
Motivation

- Graphical models define *probability distributions* over complex domains.
- These distributions are too complex to directly estimate or work with. Thus, we *factorize* the distribution - i.e. divide it into manageable parts.
  - ❖ these models allows us to *estimate* the probability of various events and to find the events which *maximize* that probability

Typical Applications

- These are particularly useful in NLP:
  - ❖ naive Bayes for document classification or topic detection
  - ❖ n-grams for language modelling
  - ❖ hidden Markov models (HMMs) for sequencing tasks (chunking, POS tagging, named entity recognition)
  - ❖ probabilistic context free grammars (P-CFGs) for syntax parsing
  - ❖ …

Graphical Model



- Modularity: a complex system is built by combining simpler parts(node).
- Probability theory: ensures consistency, provides interface models to data(node &edge).
- Graph theory: efficient general purpose algorithms.

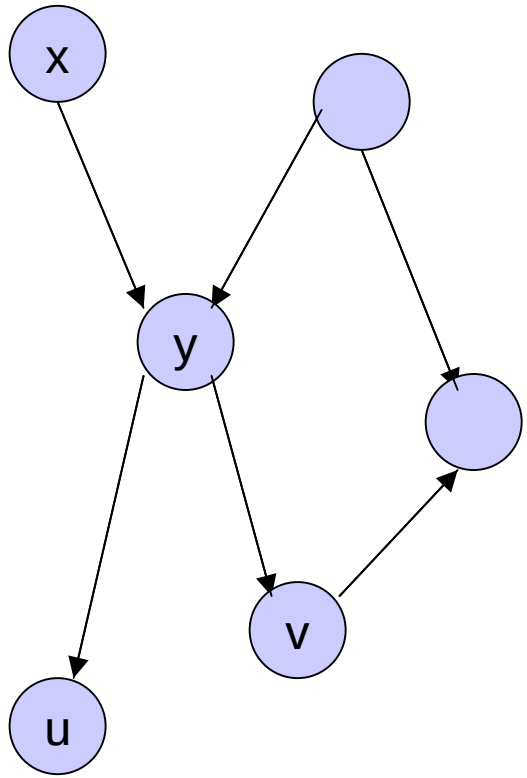
Graphical Model

Graphical representation of probabilistic relationship between a set of random variables.

Variables are represented by **nodes**.

- **Binary events**
- **Discrete variables**
- Continuous variables

Relationship is represented by (missing:marginizing) **edges**.



Undirected Graphical Model: Markov Random Fields.

Directed Graphical Model: Bayesian Networks.

Graphical Model

- Express the probabilistic dependency structure among a set of variables
- Consist of
  - a set of nodes, standing for variables
  - a set of edges, indicating dependency
  - a set of **functions** defined on the graph that defines a probability distribution

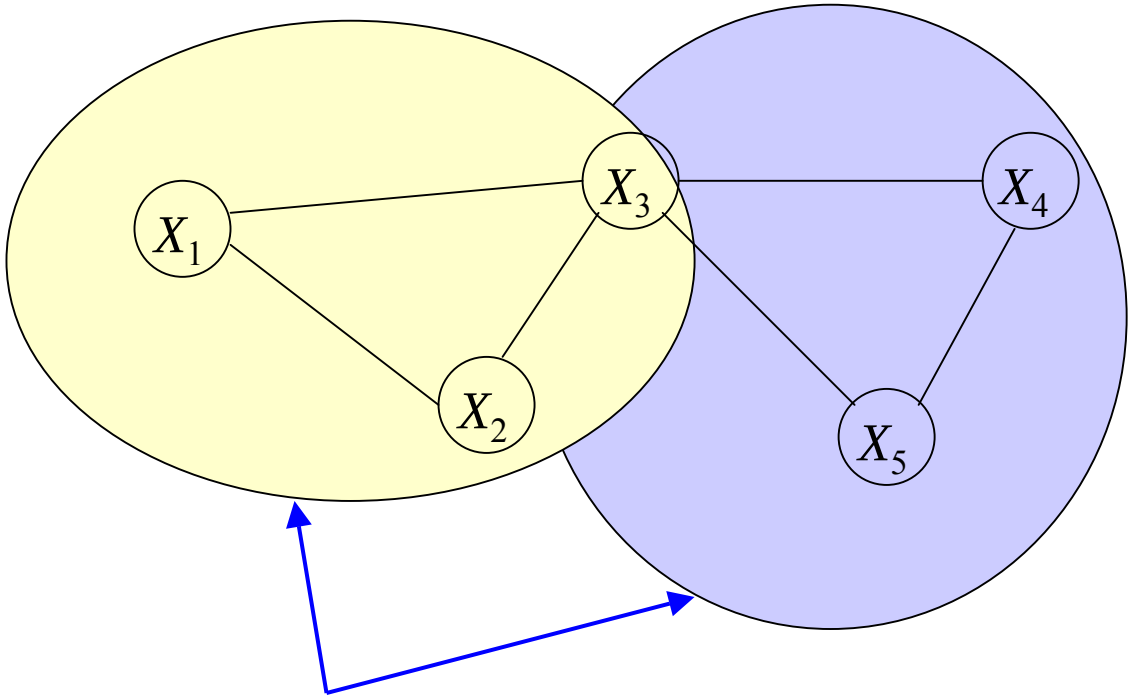
General…

- GMs represent families of probability distributions via graphs
  - directed, e.g. Bayesian networks
  - undirected, e.g. Markov random fields
  - combination, e.g. *Chain graphs*
- To need to specify:
  - **Semantics**: Bayesian network, Markov random field, ...
  - **Structure**: the graph itself
  - **Parameters** of local functions defining a probability distribution

Two classes of models

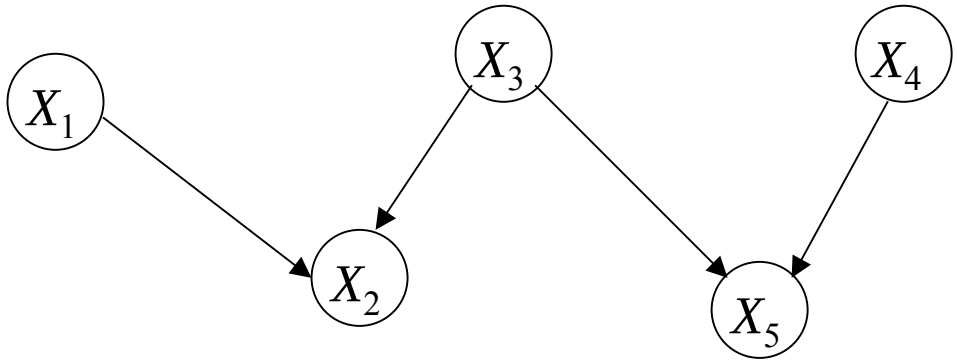
- **Bayesian networks(directed)**
  - Modeling asymmetric (causal) effects and dependencies
  - E.g. Naïve Bayes & HMM
- **Markov random fields (undirected)**
  - Modeling symmetric effects and dependencies among random variables;
  - AKA: Markov Networks
  - E.g. conditional random fields (CRFs), etc.

Undirected graphical models



- Consist of
  - a set of nodes
  - a set of edges (interdependence)

Directed graphical models



- Consist of
  - a set of nodes
  - a set of edges (conditional dependence)
  - a *conditional probability distribution* for each node, conditioned on its parents
- Constrained to directed acyclic graphs (**DAG**)

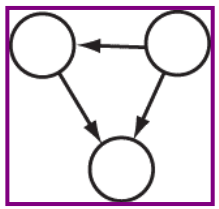
DAG

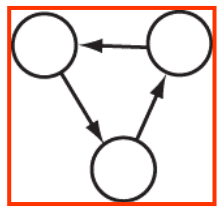
- Graph $G = (V, E)$, $V = \{X_1, X_2, \ldots, X_N\}$ (sometimes we write $V = \{v_1, \ldots, v_N\}$).

- $E = \{(X_i, X_j) : i \neq j\}$: set of directed edges.

Also, $(X_i, X_j) \in E$ means $X_i \rightarrow X_j$

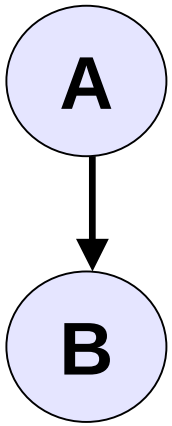
- G is acyclic, i.e, there are no directed cycles.

Valid DAG: 

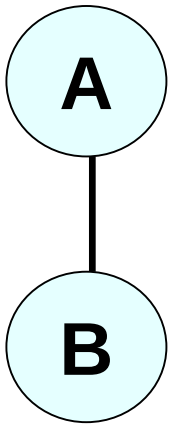
Invalid DAG: 

- DAG is also called a Bayesian network(BN)

Directed graphs vs. Undirected graphs



- Intuitively, the notion of **causality;**
- A $\rightarrow$ B, so the value of A directly determines the value of B **(one direct)**
- P(A,B) = P(B|A).P(A)



- Intuitively, the notion of correlation;
- A $\leftrightarrow$ B, so the values of A and B are interdependent **(Corelation)**

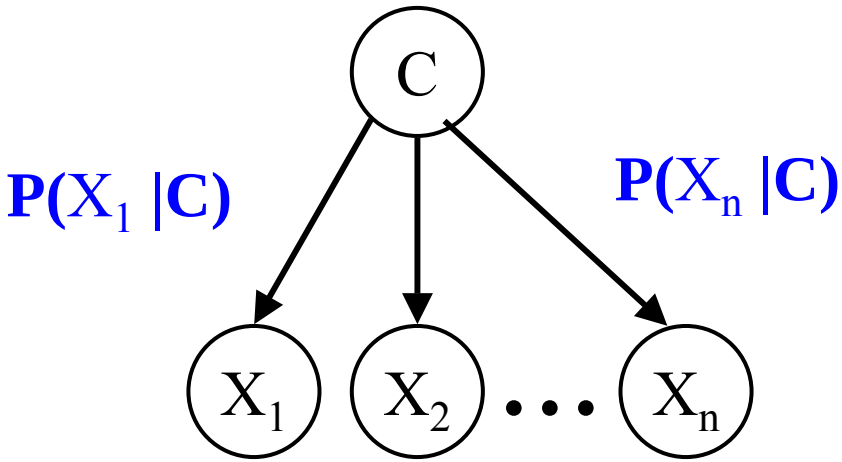
Typical Graphical Model

Generative Model

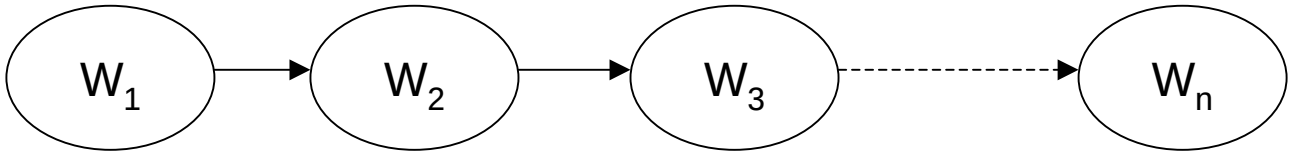
Discriminative Model

Typical Graphical Model

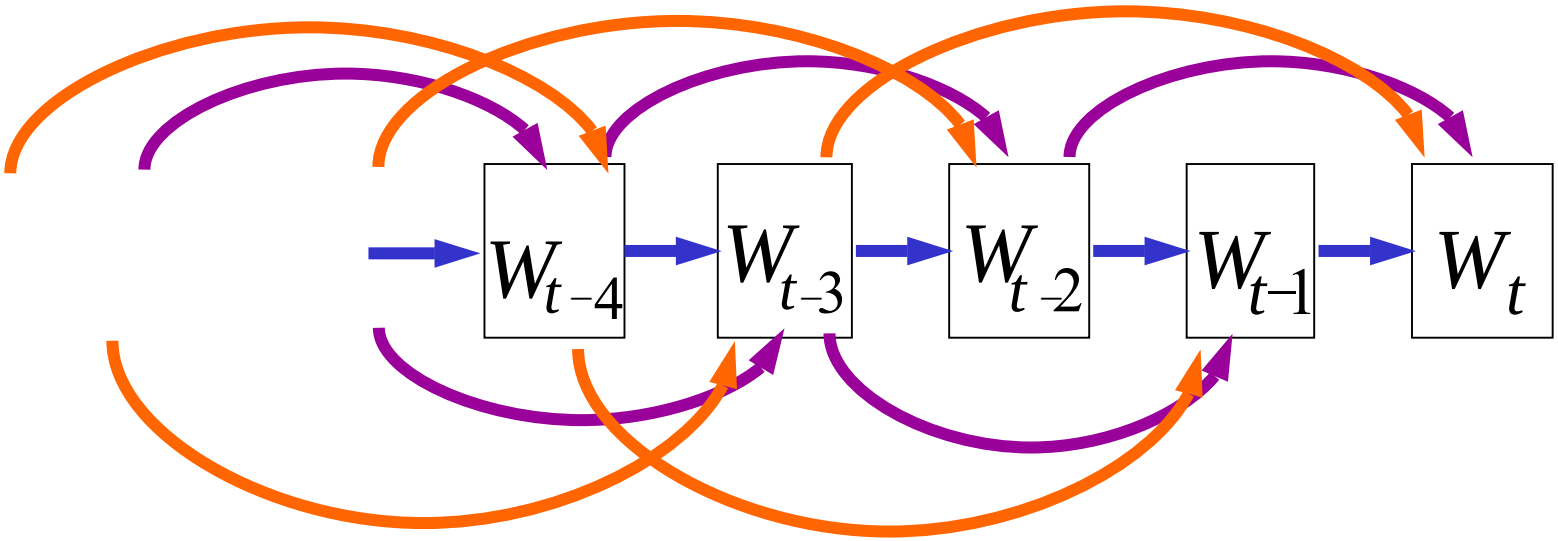
Naive Bayes



Typical Graphical Model

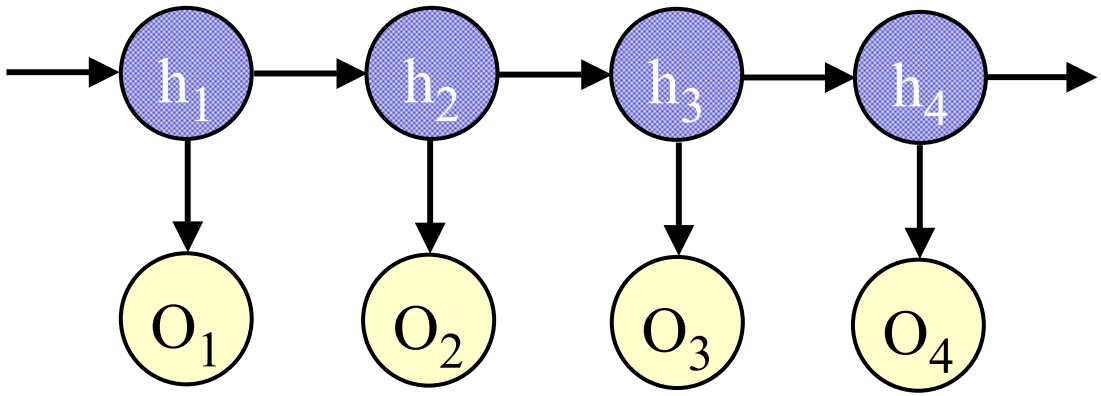


Language Model: bi-gram



Language Model: 4-gram

Typical Graphical Model

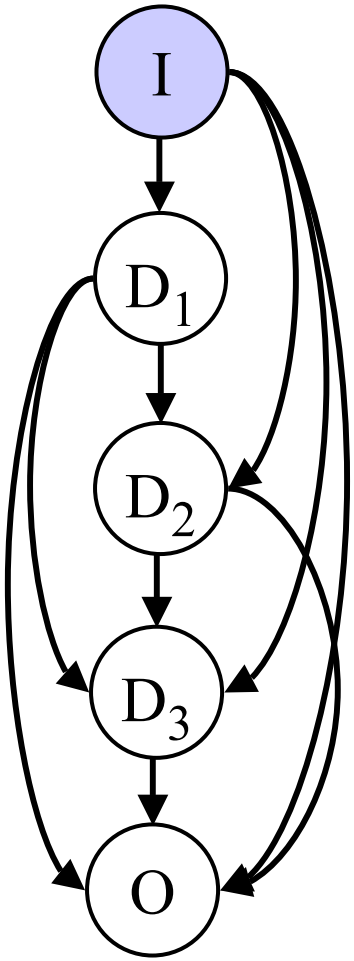


Hidden State

Observed Seq.

HMM

Generalized Decision Trees



- Generalized Probabilistic Decision Trees
  - **Each node is dependent on all nodes before it.**

Outline

- Introduction
- **Directed Graph(BN)**
- Undirected Graphical Model

Independent Random Variables

- X is independent of Y iff

  $P(X = x \mid Y = y) = P(X = x)$ for all values x and y

- If X and Y are independent then

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

- If any two random variables: $X_i, X_j \in U$ are independent, then

  $P(U) = P(X_1, \ldots X_n) = P(X_1)P(X_2)\ldots P(X_n).$

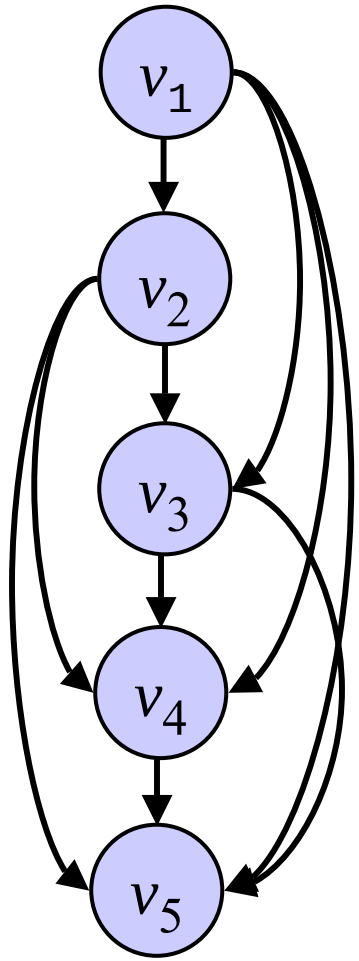
  Only n parameters is needed for this case

Independent Random Variables

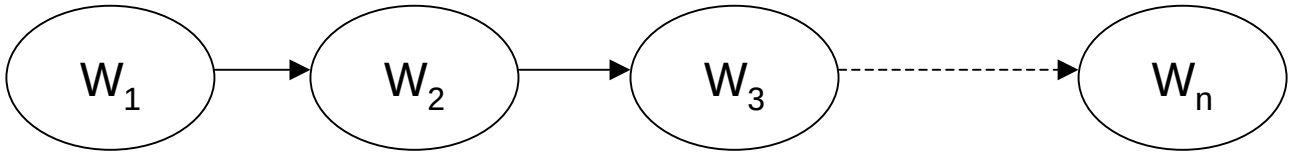
- Most of random variables of interest are not independent of each other.
- E.g. Each variable is dependent on the previous all variables in the sequence.

$$P(v_1, v_2, ..., v_n)$$

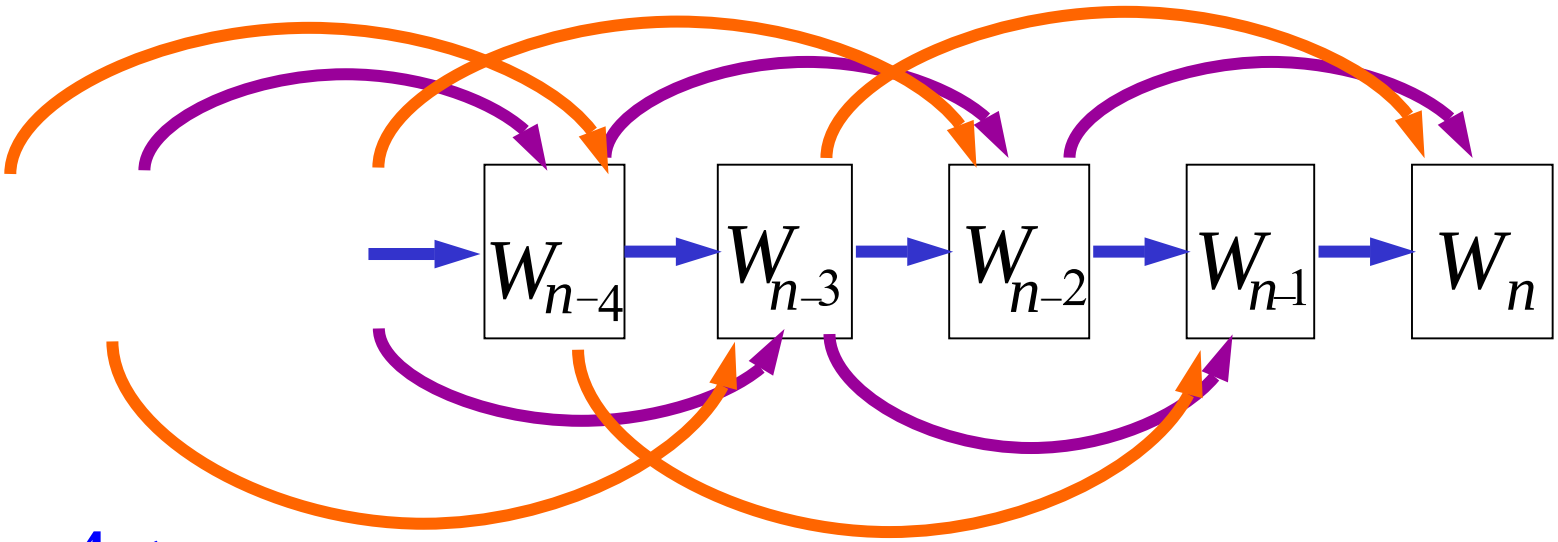
$$= P(v_n \mid v_{n-1}, v_{n-2}, ...v_1) P(v_{n-1} \mid v_{n-2}, ...v_1) \ ... \ P(v_2 \mid v_1) P(v_1)$$



Language Model again



2-gram: $P(w_1, w_2, ..., w_n) = P(w_1)\prod_{i=2}^{n} P(w_i \mid w_{i-1})$



4-gram:

$$P(w_1, w_2, ..., w_n) = P(w_3 \mid w_2, w_1)P(w_2 \mid w_1)P(w_1)\prod_{i=4}^{n} P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3})$$

Conditional Independence

- A more suitable notion is that of **conditional independence**.
- X and Y are conditional independent given Z iff :

$P(X=x|Y=y,Z=z) = P(X=x|Z=z)$ for all values x,y,z

- Notation: $I(X,Y|Z)$ or $X \perp\!\!\!\perp Y \mid Z$

  That is, the values of Y does not change prediction of X once we know the value of Z

- A different form: X and Y are conditional independent given Z iff :

$P(X=x,Y=y|Z=z) = P(X=x|Z=z)\,P(Y=y|Z=z)$ for all values x,y,z

Rule

- If X and Y are conditional independent given Z ,then:

  P(X,Y,Z)

  = P(X|Y,Z) P(Y,Z)

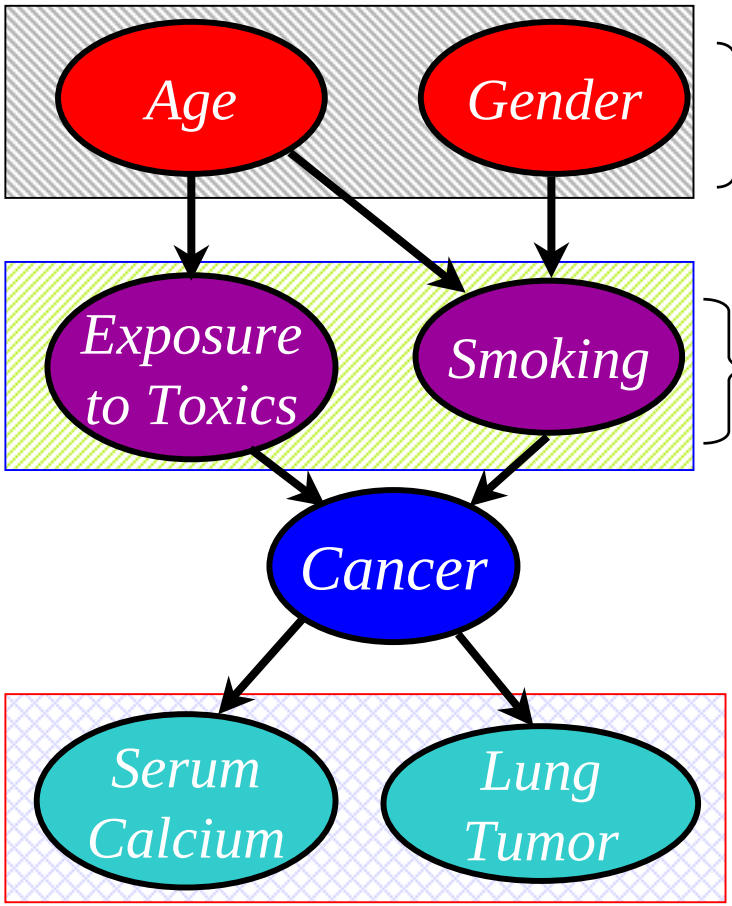
  = P(X|Y,Z) P(Y|Z) P(Z)

  = P(X|Z) P(Y|Z) P(Z)

- Graphical Models can help answer conditional independent queries

An example

- A variable (node) is conditionally independent of its non-descendants given its parents.



Non-Descendants

Parents

Cancer is independent of ***Age*** and ***Gender*** given ***Exposure to Toxics*** and ***Smoking***.

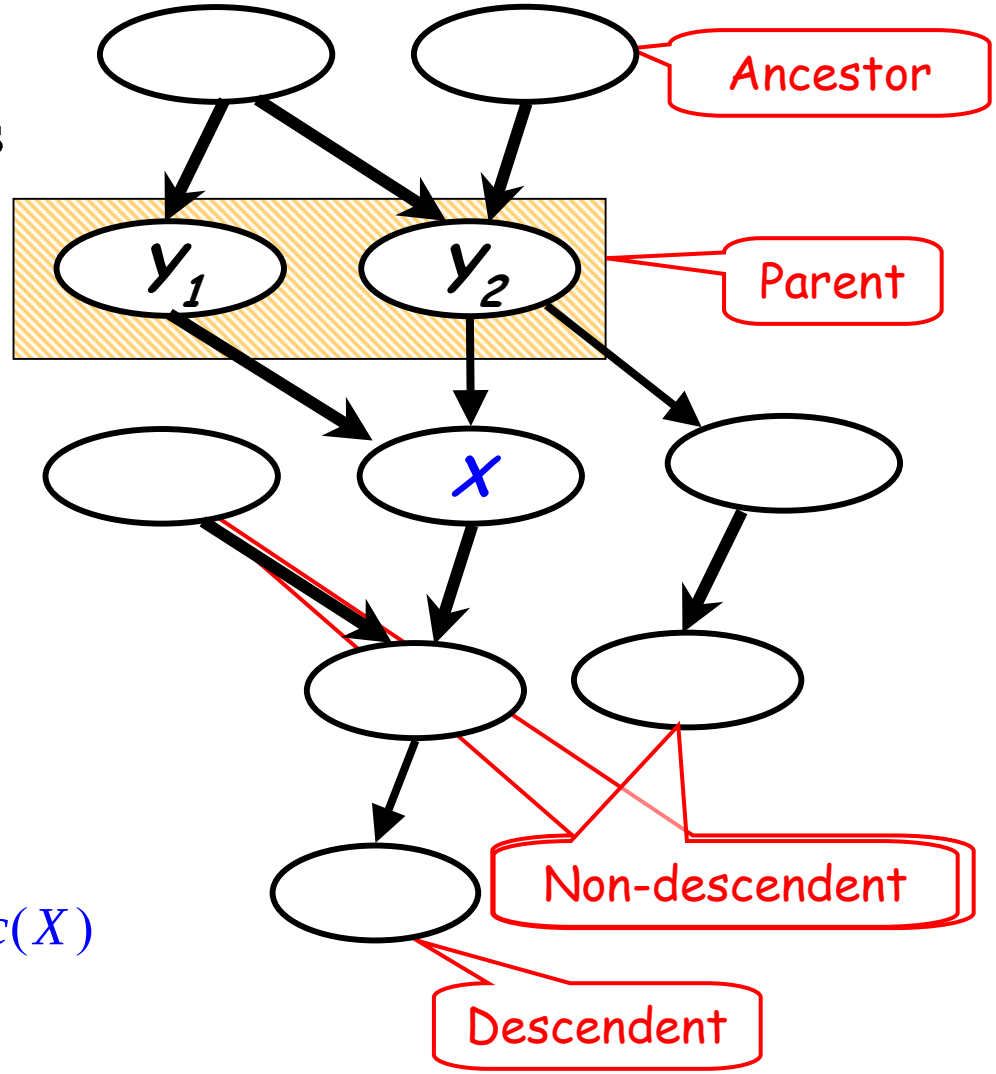
Descendants:Y is the **descendent** of X if there is a path from X to Y and each step follows the direct.

Complex Case

- For each variable X and parents pa(X) exists a conditional probability
  - **P(X | pa(X))**
- Each random variable X, is independent of its **non-descendents**, given its parents Pa(X)
- Formally,

$$X \perp\!\!\!\perp No - Desc(X) \mid Pa(X)$$

where, $Ancestor(x) \in No - Desc(X)$



Conditional Independence

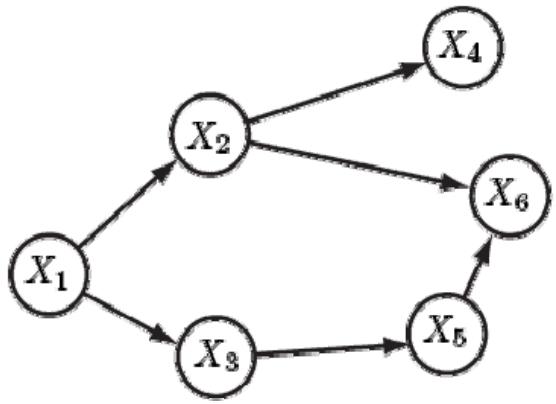
- Ex: is $X_4 \perp\!\!\!\perp \{X_1, X_3\} | X_2$?

$$p(x_{1:4}) = \sum_{x_5, x_6} p(x_{1:6})$$

marginalize (sum) over $X_{5,6}$

$$= p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)\sum_{x_5} p(x_5 \mid x_3)\sum_{x_6} p(x_6 \mid x_2, x_5)$$

$$= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)$$



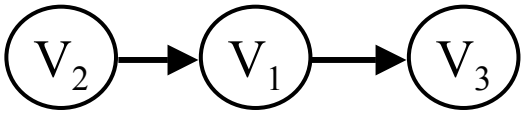
$$p(x_{1:3}) = p(x_1)p(x_2|x_1)p(x_3|x_1)$$

$$p(x_4|x_{1:3}) = \frac{p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)}{p(x_1)p(x_2|x_1)p(x_3|x_1)} = p(x_4|x_2)$$

Three Cases

- Three 3-node examples and their conditional independence.

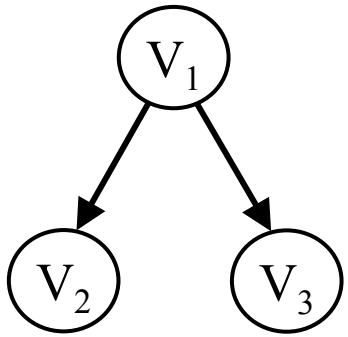
Case 1: Linear



$$V_2 \perp\!\!\!\perp V_3 \mid V_1$$

$$V_2 \neg\perp\!\!\!\perp V_3$$

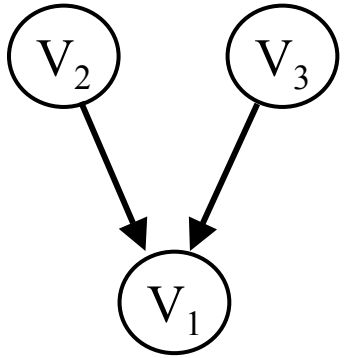
Case 2: Diverging



$$V_2 \perp\!\!\!\perp V_3 \mid V_1$$

$$V_2 \neg\perp\!\!\!\perp V_3$$

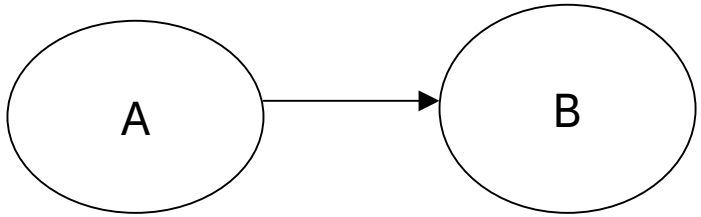
Case 3: Converging



$$V_2 \perp\!\!\!\perp V_3$$

$$V_2 \neg\perp\!\!\!\perp V_3 \mid V_1$$

Simple Case

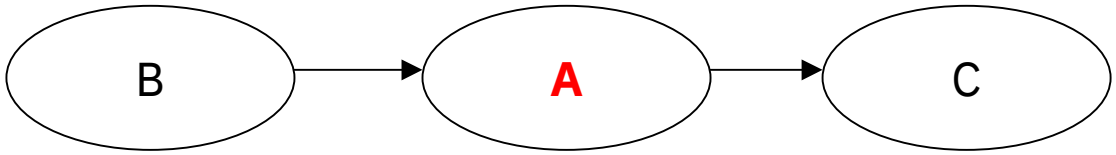


- **Dependency** is described by the **conditional probability** P(B|A)
- Knowledge about A: priori probability P(A)
- Calculate the joint probability of the A and B

$$P(A,B)=P(B|A)P(A)$$

Case 1: Chain (Serial Connection)

Markov Chain(**head-to-tail via A**)



$B \perp\!\!\!\perp C \mid A$ why?

$B \neg\perp\!\!\!\perp C$ why

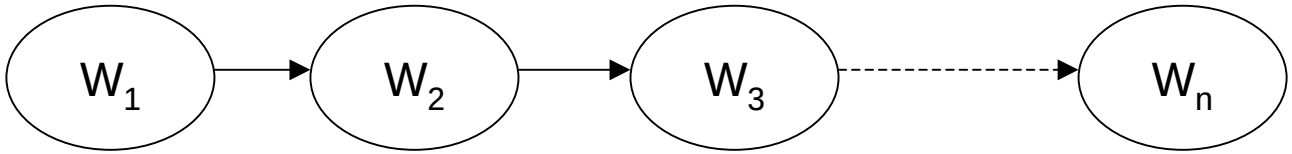
- $P(A,B,C)=\mathbf{P(C|A)}\ \mathbf{P(A|B)P(B)}$

$$\Rightarrow P(B,C)=P(B)\sum_A P(C\mid A)P(A\mid B)\neq P(B)P(C)\Rightarrow B \neg\perp\!\!\!\perp C$$

$$\Rightarrow P(B,C\mid A)=\frac{P(A,B,C)}{P(A)}=\frac{P(B)P(A\mid B)P(C\mid A)}{P(A)}$$

$$=\frac{P(A,B)P(C\mid A)}{P(A)}=P(B\mid A)P(C\mid A)\quad\Rightarrow\quad B \perp\!\!\!\perp C \mid A$$

Language Model: 2-gram



In general, for a sentence: S=$w_1, w_2, ..., w_n$

$$P(w_1, w_2, ..., w_n) = P(w_n \mid w_1, w_2 ..., w_{n-1}) P(w_1, w_2 ..., w_{n-1})$$

Considering conditional independence,

$$P(w_n \mid w_1, w_2 ..., w_{n-1}) = P(w_n \mid w_{n-1})$$

So, $P(w_1, w_2, ..., w_n) = P(w_n \mid w_{n-1}) P(w_1, w_2 ..., w_{n-1})$

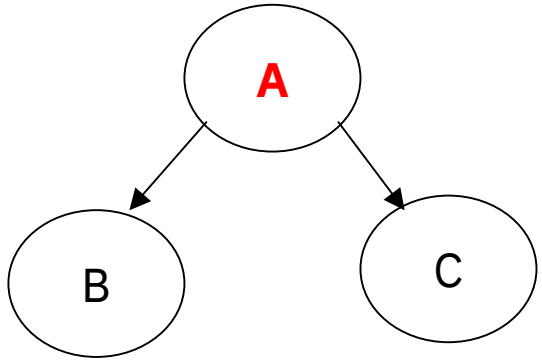
$$P(w_1, w_2, ..., w_{n-1}) = P(w_{n-1} \mid w_{n-2}) P(w_1, w_2 ..., w_{n-2})$$

.....

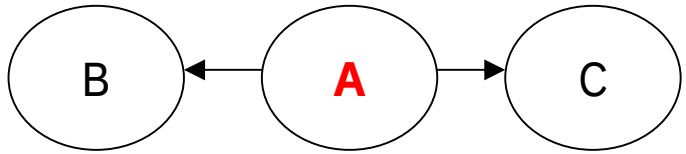
$$P(w_1, w_2, ..., w_n) = P(w_1) \prod_{i=2}^{n} P(w_i \mid w_{i-1}) \quad \text{(2-gram)}$$

Case2: Diverging Connection

Λ-structure: Also Markov chain(tail-to-tail via A)



or



$B \perp\!\!\!\perp C \mid A$ why

$B \neg\perp\!\!\!\perp C$ why

- $P(A,B,C)=\boldsymbol{P(C|A)}\ \boldsymbol{P(B|A)P(A)}$

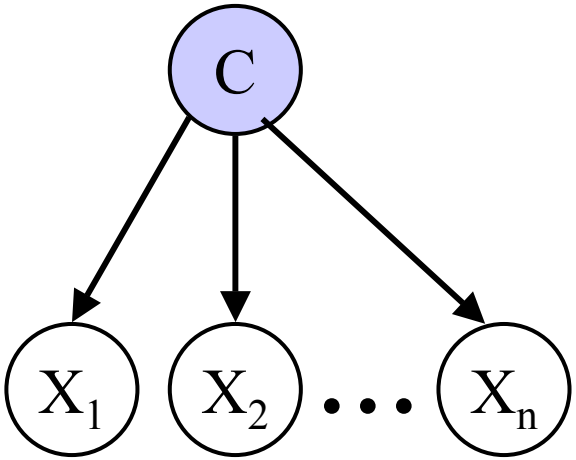
$$\Rightarrow P(B,C)=\sum_A P(B\mid A)P(C\mid A)P(A)\neq P(B)P(C)\Rightarrow B \neg\perp\!\!\!\perp C$$

- $P(A,B,C)=\boldsymbol{P(C|B,A)P(B,A)}=\boldsymbol{P(C|B,A)}\ \boldsymbol{P(B|A)P(A)}$

$$\Rightarrow P(C\mid B,A) = P(C\mid A) \Rightarrow B \perp\!\!\!\perp C \mid A$$

An Example: Naïve Bayes

Naive Bayes



$$P(c, x_1, x_2, \ldots x_n)$$

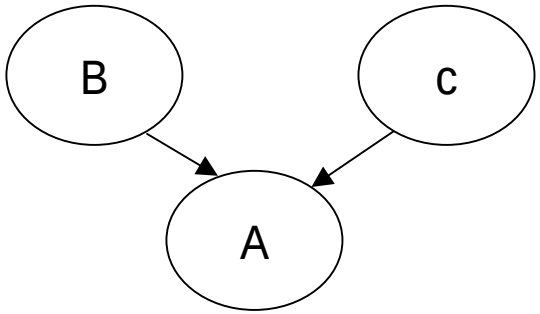
$$= P(c) \bullet \prod_{i=1}^{n} P(x_i \mid c)$$

Generative Model:

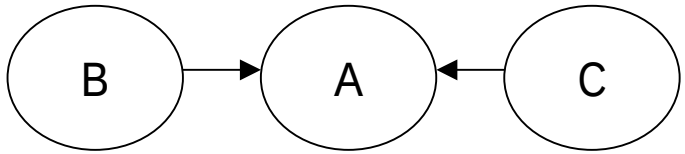
The set of words are generated by the category c

Case 3: Converging Connection

V-structure(head-to-head via A)



or



$B \perp\!\!\!\perp C$ why

$B \neg\perp\!\!\!\perp C \mid A$ why

$$P(A,B,C)=\mathbf{P(C)\ P(B)P(A|B,C)}$$

$$P(B,C) = \sum_A P(B)P(C)P(A \mid B,C)$$

$$= P(B)P(C)\sum_A P(A \mid B,C) = P(B)P(C) \Rightarrow B \perp\!\!\!\perp C$$

$$\Rightarrow P(B,C \mid A) = \frac{P(A,B,C)}{P(A)} = \frac{P(B)P(C)P(A \mid B,C)}{P(A)}$$

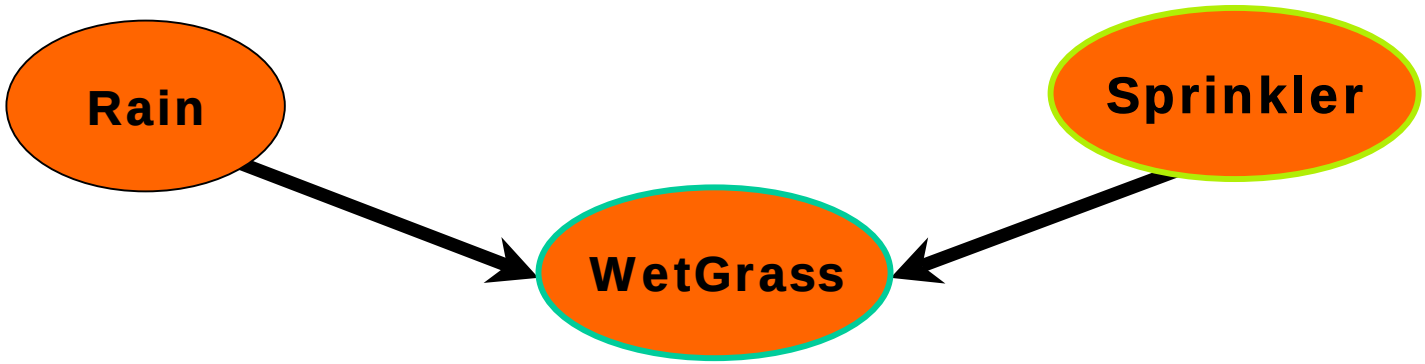
$$\neq P(B \mid A)P(C \mid A) \Rightarrow B \neg\perp\!\!\!\perp C \mid A$$

Independence

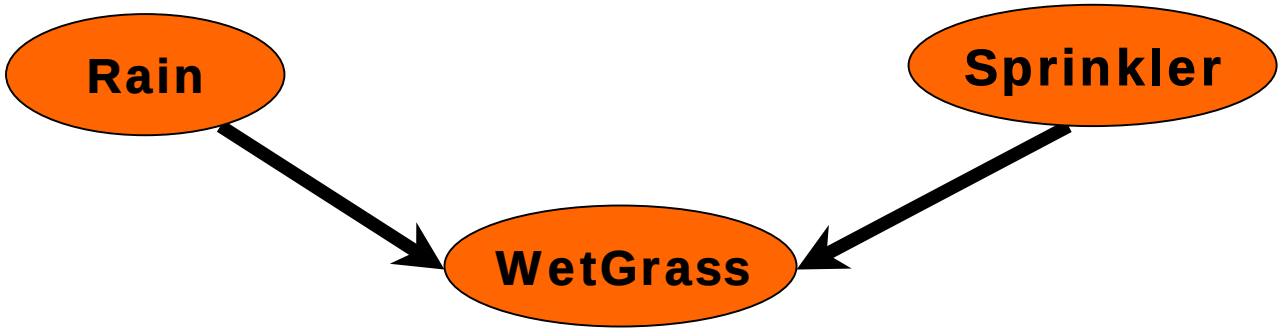
- case1=case2
  - A tail-to-tail node or a head-to-tail node leaves a path unblocked unless it is observed in which case it blocks the path
- case3 has opposite behavior from both case1 and cace2
  - A head-to-head node blocks a path if it is unobserved, but once the node, and/or at least one of its descendants, is observed the path becomes unblocked.

Explaining away

- If anything is known about the consequence, then information on one possible cause may tell us something about the other causes.



Explaining away

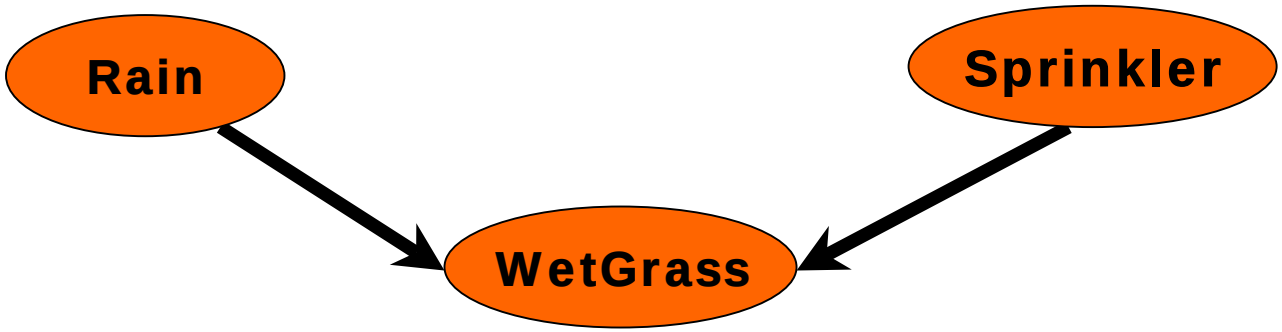


$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

Assume grass will be wet if and only if it rained last night, or if the sprinklers were left on:

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Explaining away



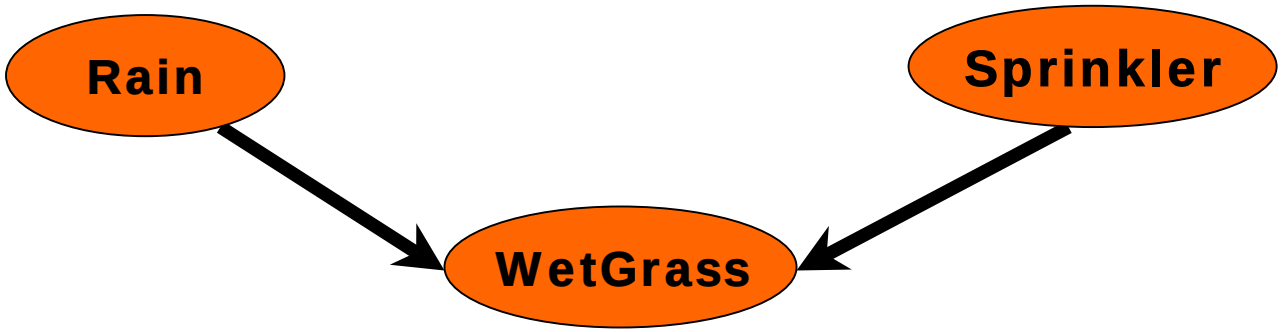
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(w,r)}{p(w)} = \frac{P(w \mid r)P(r)}{\sum_{r',s'} P(w \mid r',s')P(r',s')}$$

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

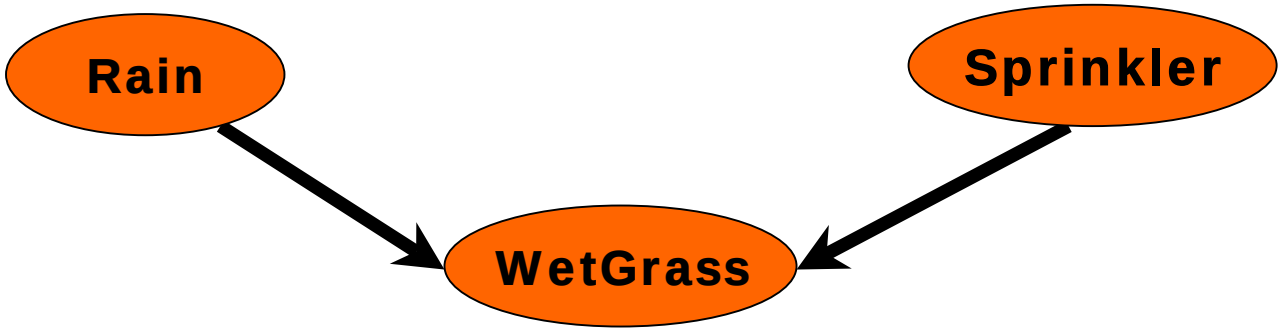
$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$

$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{P(r,s) + P(r,\neg s) + P(\neg r,s)}$$

Explaining away



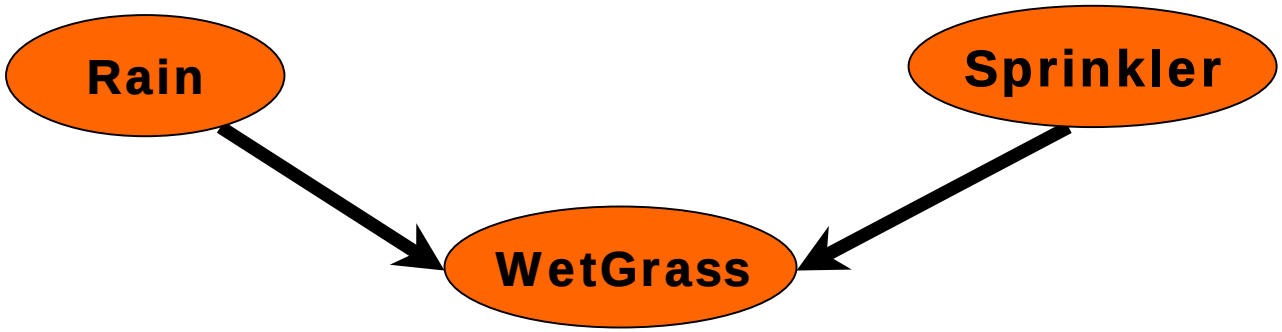
$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{P(r) + P(\neg r, s)}$$

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$

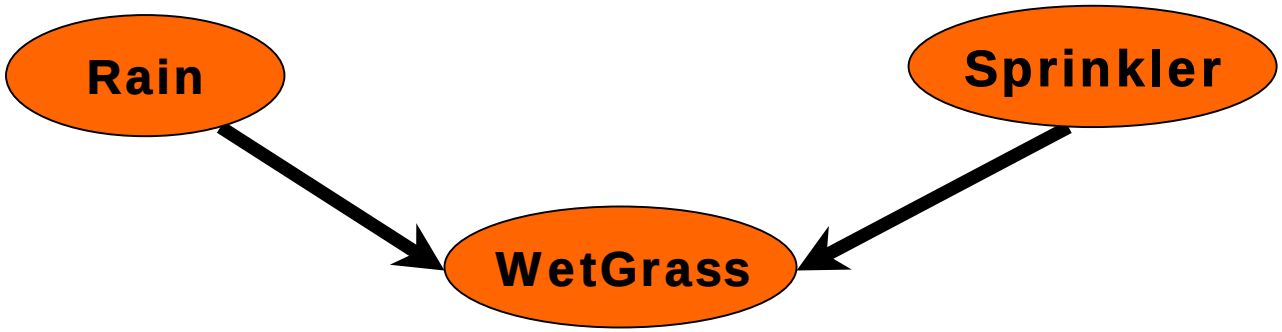
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet:

$$P(r \mid w) = \frac{P(r)}{P(r) + P(\neg r)P(s)} > P(r)$$

Between 1 and $P(s)$

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$

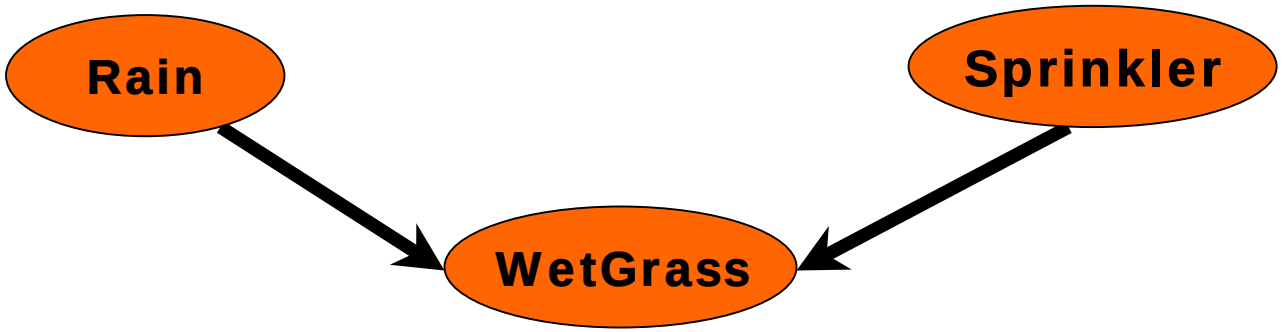
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet and sprinklers were left on:

$$P(r \mid w,s) = \frac{P(w \mid r,s)P(r \mid s)}{P(w \mid s)}$$

Both terms = 1

Explaining away



$$P(R,S,W) = P(R)P(S)P(W \mid S,R)$$

$$P(W = w \mid S,R) = 1 \quad \text{if } S = s \text{ or } R = r$$

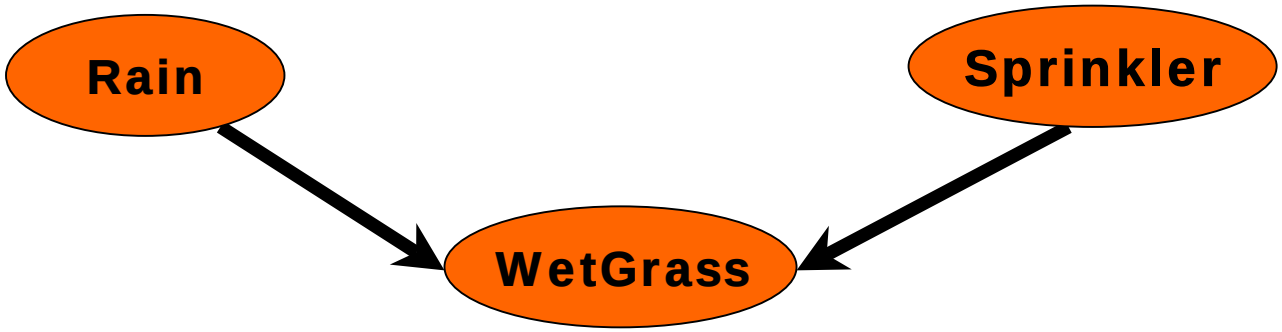
$$= 0 \quad \text{if } R = \neg r \text{ and } S = \neg s.$$

Compute probability it rained last night, given that the grass is wet and sprinklers were left on:

$$P(r \mid w,s) = P(r \mid s) = P(r)$$

Explaining away

- Given the consequence(**WetGrass**);
- Information on one cause(**Sprinkler**) tell us something about the other causes(**Rain**)-decreasing the probability of **'Rain'.**



$$P(r \mid w) = \frac{P(r)}{P(r) + P(\neg r)P(s)} > P(r)$$

V

$$P(r \mid w, s) = P(r \mid s) = P(r)$$

"Discounting" to prior probability.

D-Separation

- Two nodes A and B is said to be d-separated by a connection set of nodes Z if one of the following holds:

1. The node **$z \in Z$,** connecting A and B, meet case1(head-to-tail) or case2(tail-to-tail)

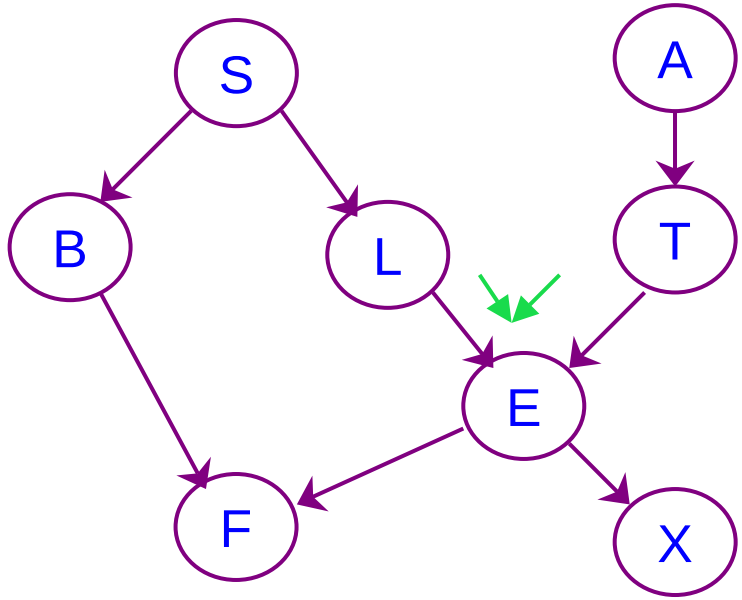
2. Any node **z' ,** connecting A and B, meet case3(head-to-head) but **$z' \notin (Z \cup \text{descendent}(Z))$**.

Z d-separates A and B if it d-separates every chain from A to B.

Z d-separates two sets of nodes, X and Y, if it d-separates every chain from a node in X to a node in Y.

D-Separation - Examples

Consider the following DAG:



I(X, A | T) and I(B, L | S, E)

Note, however, that the DAG **does not entail**:

I (B, L | S, F) since F unblocks the chain {B,F,E,L}

Nor I(L, T | X) since X unblocks the chain {L,E,T}

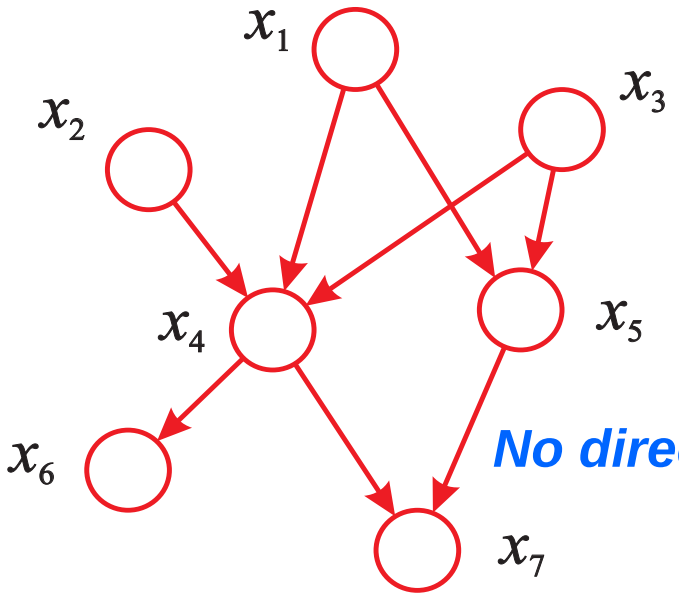
Nor I(F, T | E) since E unblocks the chain {F,B,S,L,E,T} **why?**

Factorization: Decomposition

- Joint distribution (**n nodes**)

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(\, x_i \mid pa_i \,)$$

where, pa_i *is the parents of node* x_i



No directed cycles

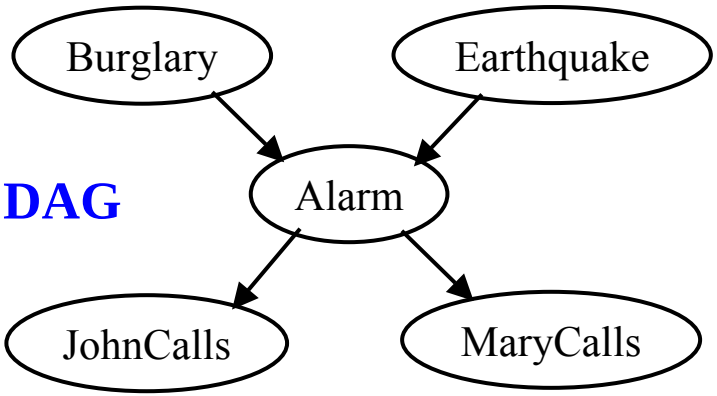
Parents of x_4

$$P(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$
$$= P(x_1) \bullet P(x_2) \bullet P(x_3)$$
$$\bullet P(x_4 \mid x_1, x_2, x_3)$$
$$\bullet P(x_5 \mid x_1, x_3) \bullet P(x_6 \mid x_4)$$
$$\bullet P(x_7 \mid x_4, x_5)$$

Bayesian Network (general)

- Two component: BN=(G, P)
  - **Directed acyclic graph (G)**
    - Nodes correspond to random variables
    - (Missing) links encode independences
  - **Parameters (P)**
    - Local conditional Probability Distribution for every variable and its parents: P(X | pa(X))

DAG



Local CPD

P(A|B,E)

B	E	T	F
T	T	.95	.05
T	F	.94	.06
F	T	.29	.71
F	F	.001	.999

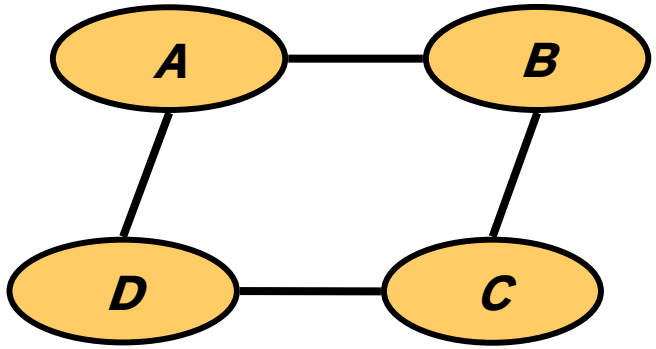
Outline

- Introduction
- Directed Graph(BN)
- **Undirected Graphical Model**

Undirected Graphical Model

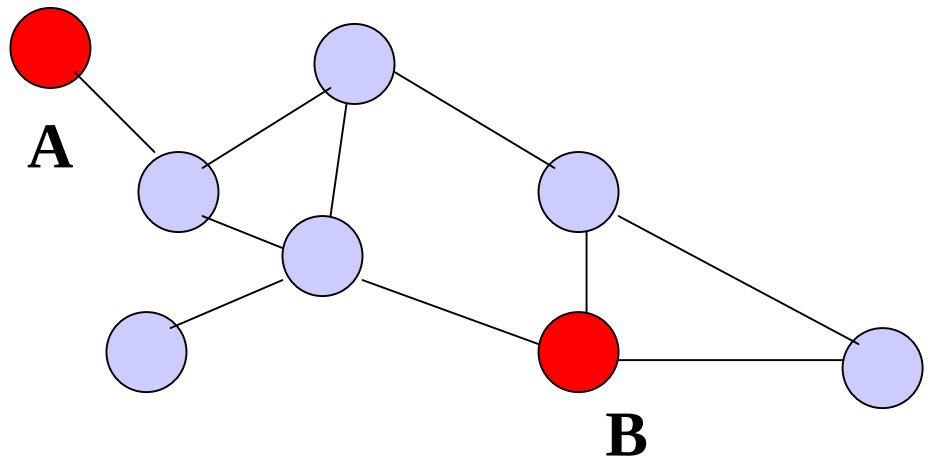
- Three names:
  - Undirected graphical model (U-GM)
  - Markov random field (MRF)
  - Markov network (MN)
- Definition: MRF is a graphical representation of a probability distribution using undirected graph. Given a graph G=(V,E), where V is a set of nodes (random variable) and $E=\{(v_i,v_j)|\ v_i,v_j \in V\}$ is a set of edges linking a pair of nodes and the each edge is undirected.

Markov Random Fields



Pairwise Markov Property

- Two nodes in the network that are not directly connected can be made **independent** given all other nodes. **(point to point)**



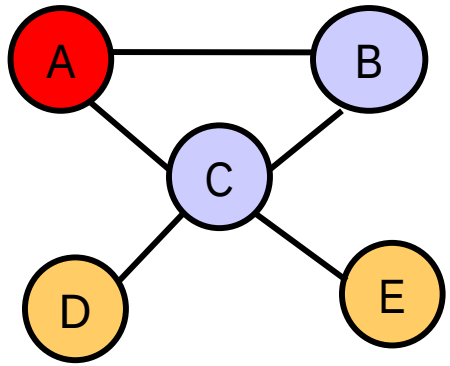
Local Markov Property

- **Local Markov property**: X is independent of all other nodes given its neighbors.
- Let G be an **undirected graph,** Let U is the set of all nodes in G, N_i be the **set** of **neighbors** of X_i , the independence:
  $I(X_i ; U - N_i - \{X_i\} \mid N_i)$. **(point to set)**

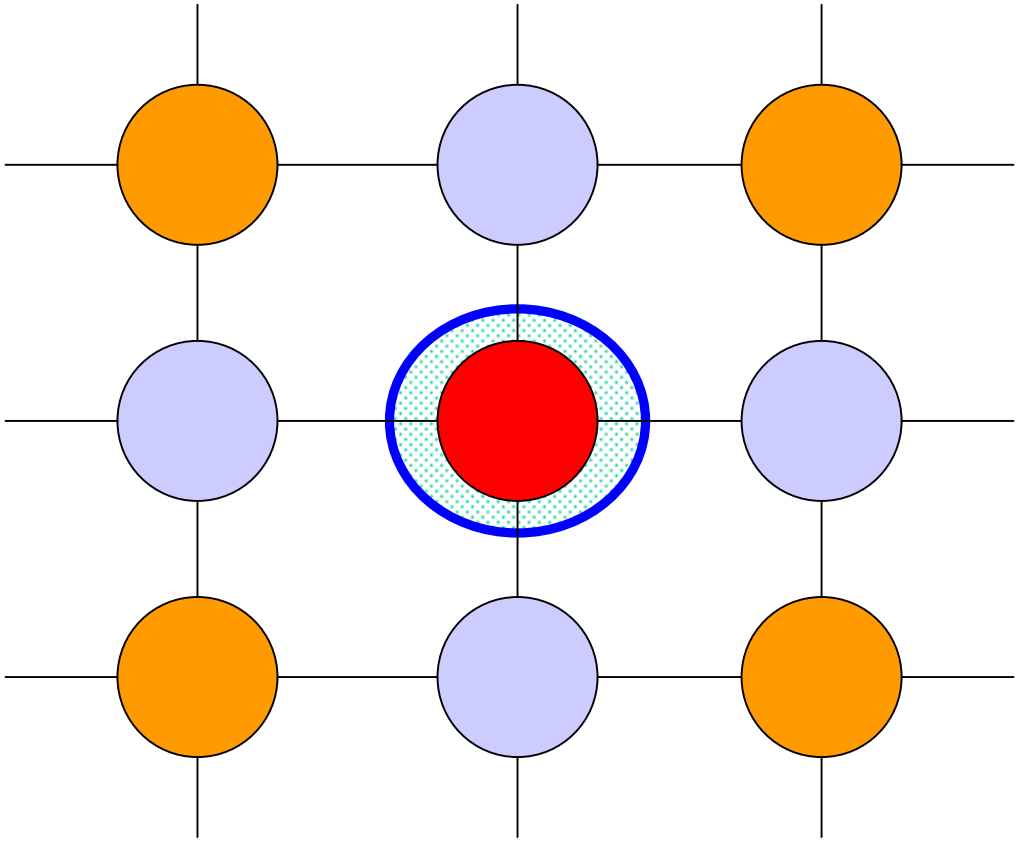
Example:

This graph implies that

$$A \perp\!\!\!\perp \{D, E\} \mid \{B, C\}$$



Neighbors & Independence



A node is conditionally independent of all others given its neighbours.

Markov Blanket

- The **Markov blanket** of a node, X, in a Markov Random Fields, is the set of its neighbors in the graph (nodes that have an edge connecting to X).
- Every node in a Markov Random Fields is conditionally independent of every other node given its Markov blanket—Local Markov property

So, $p(x \mid all\_other\_nodes)$

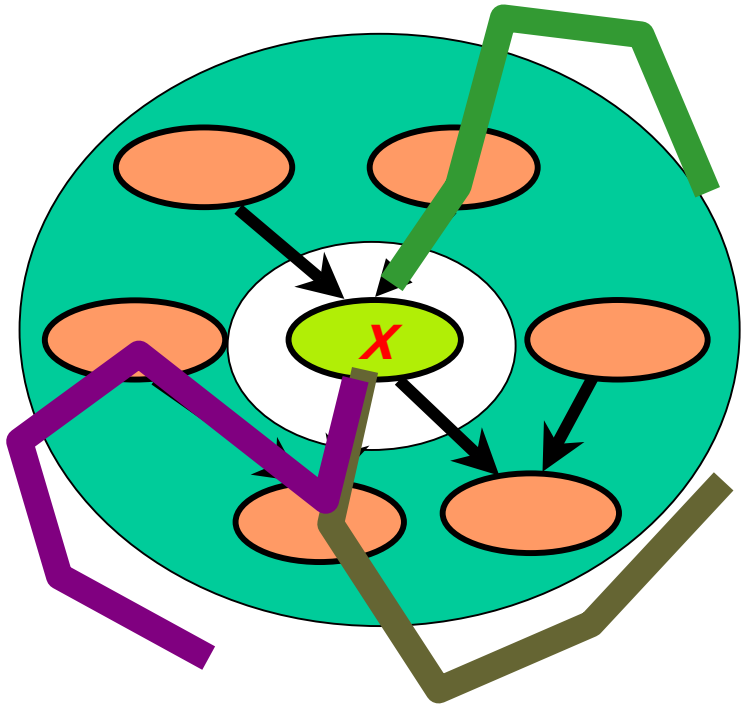
$= p(x \mid all\_other\_nodes - neighbor(x), neighbor(x))$

$= p(x \mid neighbor(x))$

Markov Blanket for DAG

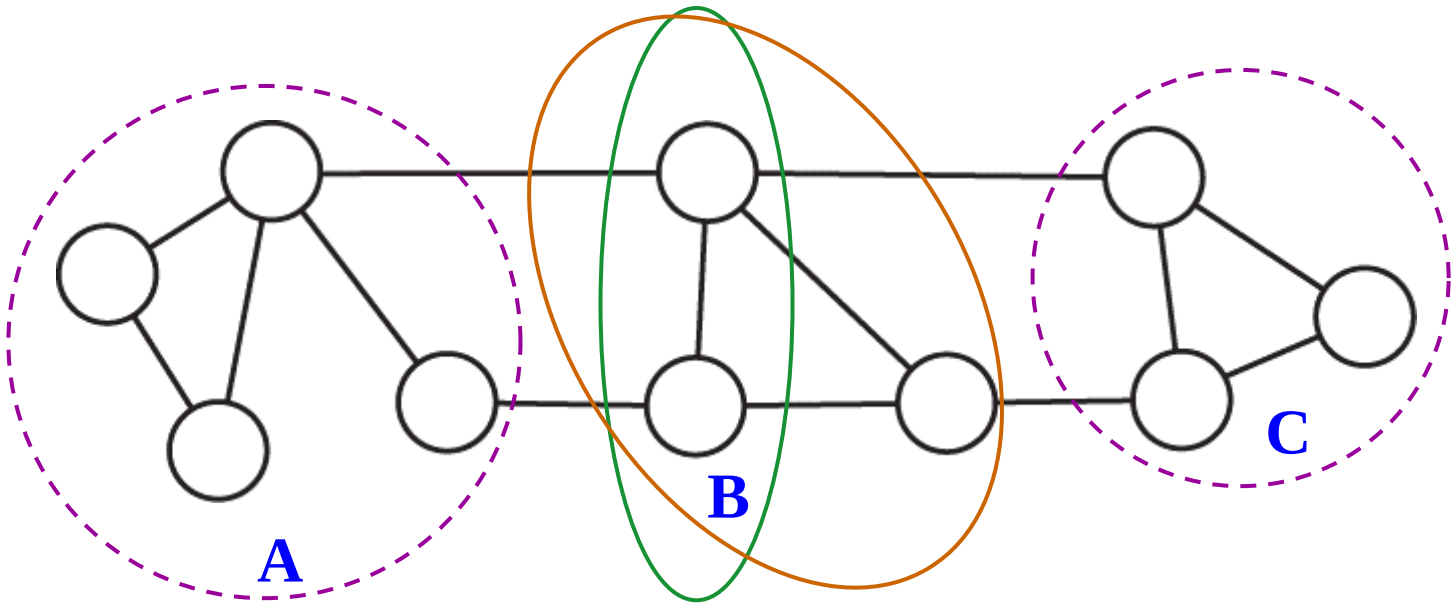
Three types of Paths:

- Upward paths
  - Blocked by **parents**
- Downward paths
  - Blocked by **children**
- Sideway paths
  - Blocked by "**spouses**"



Global Markov Property

- Global Morkov Property(G): $A \perp\!\!\!\perp C \mid B$ if B separates A from C (all paths from any node in A to any node in C go through some node in B) **(set to set)**



Maximal Cliques

- Clique: subset of nodes where each pair connected
- Maximal clique: no node can be added & remain a clique

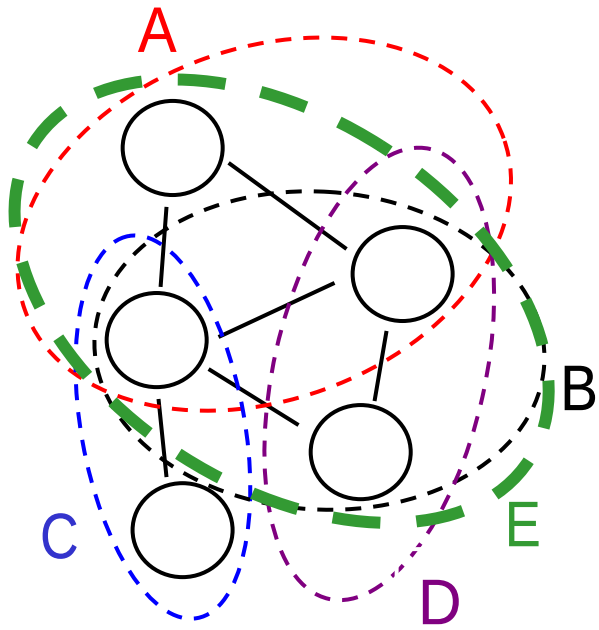
A is a maximal clique

B is a maximal clique

C is a maximal clique

D is a clique but non-maximal clique

Is E a clique?



Factorization: Potential function

- The joint distribution is product of non-negative functions over all maximal *cliques* of the graph

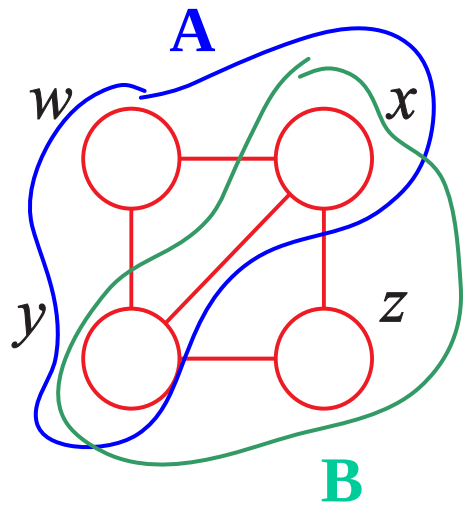
$Given: G = (V, E)$

$$P(V) = \frac{1}{Z} \prod_{C} \psi_C(x_C)$$

Z is the normalization constant: $Z = \sum_{x \in V} \prod_{C} \psi_C(x_C)$

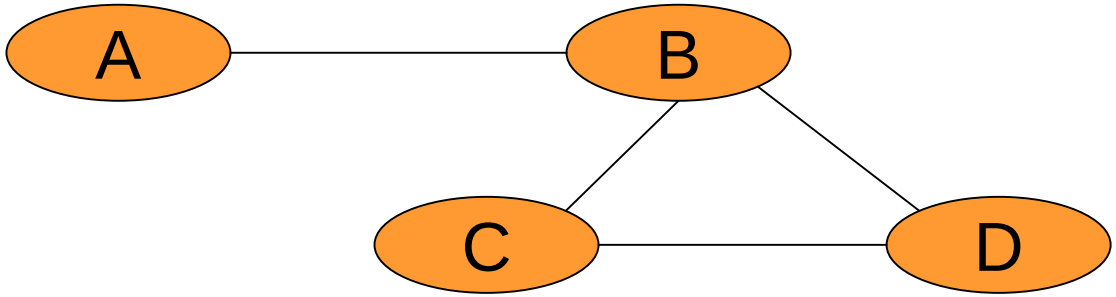
C is a maximal clique and the set of variables in C is x_C

$\psi_C(x_C)$ is the potential function satifying $\psi_C(x_C) \geq 0$



Alternative form: $$P(V) = \frac{1}{Z} e^{\sum_i g(C_i)}$$

Example



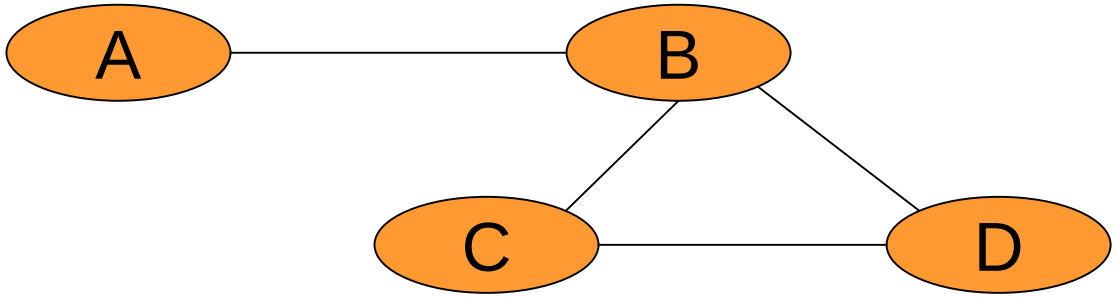
- Potential functions defined over cliques

$$P(X) = \frac{1}{Z}\prod_{c}\psi_c(X_c) \qquad Z = \sum_{X}\prod_{c}\psi_c(X_c)$$

$$\psi(A,B) = \begin{cases} 3.7 & \text{if A and B} \\ 2.1 & \text{if A and } \overline{\text{B}} \\ 0.7 & \text{otherwise} \end{cases}$$

$$\psi(B,C,D) = \begin{cases} 2.3 & \text{if B and } \overline{\text{C}} \text{ and D} \\ 5.1 & \text{otherwise} \end{cases}$$

Cont.



- Potential functions defined over cliques

$$P(X) = \frac{1}{Z}\exp\left(\sum_i w_i f_i(X)\right) \qquad Z = \sum_X \exp\left(\sum_i w_i f_i(X)\right)$$

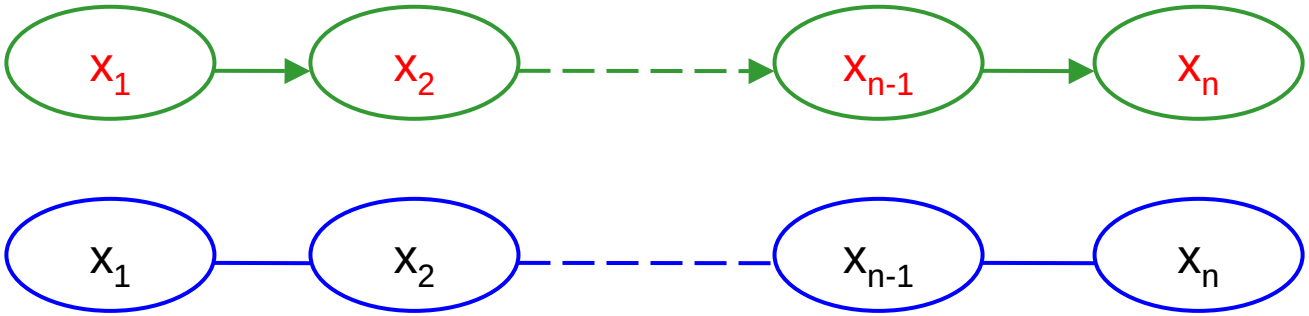
Weight of Feature *i*

Feature *i*

$$f(A,B) = \begin{cases} 1 & \text{if A and B} \\ 0 & \text{otherwise} \end{cases}$$

$$f(B,C,D) = \begin{cases} 1 & \text{if B and } \overline{\text{C}} \text{ and D} \\ 0 & \end{cases}$$

Relation to Directed Graphs



for(a): $p(\mathbf{x}) = p(x_1)\,p(x_2 \mid x_1)\,p(x_3 \mid x_2)\ldots p(x_n \mid x_{n-1})$

for(b): $p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\ldots\psi_{n-1,n}(x_{n-1}, x_n)$

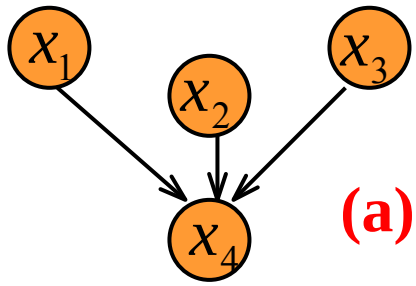
let: $\psi_{1,2}(x_1, x_2) = p(x_1)\,p(x_2 \mid x_1)$, $\psi_{2,3}(x_2, x_3) = p(x_3 \mid x_2)$,

$\ldots\ \psi_{n-1,n}(x_{n-1}, x_n) = p(x_n \mid x_{n-1})$. Note that $Z = 1$ in this case.

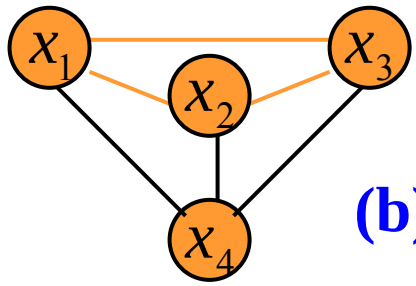
*p(**x**) for(a) equals to p(**x**) for(b)*

*So, **(a)*** can be converted to ***(b)*** *equally*

Complex case: more than one parent



(a)



(b)

for(a): $p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4 \mid x_1, x_2, x_3)$

Obviously, $p(x_4 \mid x_1, x_2, x_3)$ contains 4 variables: x_1, x_2, x_3 and x_4.

So, the four variables must be contained **in a single clique** in order to form a **clique potential**. Three edges must be added.

The process of '***marrying the parents***' has become as ***moralization***, and the undirected graph, after removing arrows, is called ***moral graph.***

Moralized Graphs

- Given a DAG ***G***, we define the **moralized graph** of ***G*** to be an undirected graph ***U*** such that
  - if $X \rightarrow Y$ in ***G***, then X -- Y in ***U***
  - if $X \rightarrow Y \leftarrow Z$ in ***G***, then X -- Z in ***U***
  - no other edges are in ***U***

Conversion

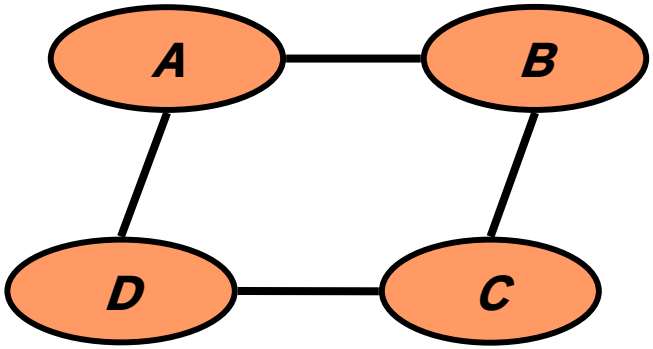
- Constructing the moral graph:
  - Add additional undirected edges between all pair of parents for each node in the graph: **marrying parents**;
  - Drop the arrows on original arcs;
  - Obtain the moral graph.
- Clique potential:
  - Initialize all of clique potentials of the moral graph to 1;
  - Take each conditional distribution factor in the original directed graph and multiply it into one of the clique potentials.
  - Set 1 to Z;

MRF vs. Bayesian Networks

- The transformation to a Moral graph loses information about **independencies**
- Thus, Markov networks cannot model "explaining away"

Example: Expressive Power

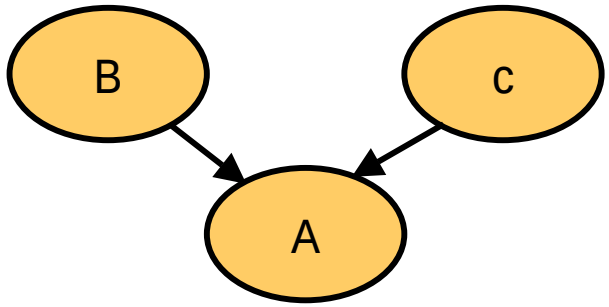
- Can we always convert Directed ↔ Undirected ?
  - **No.**



$$A \perp\!\!\!\perp C \mid \{B, D\}$$

$$B \perp\!\!\!\perp D \mid \{A, C\}$$

- No directed model can represent these and only these independencies !



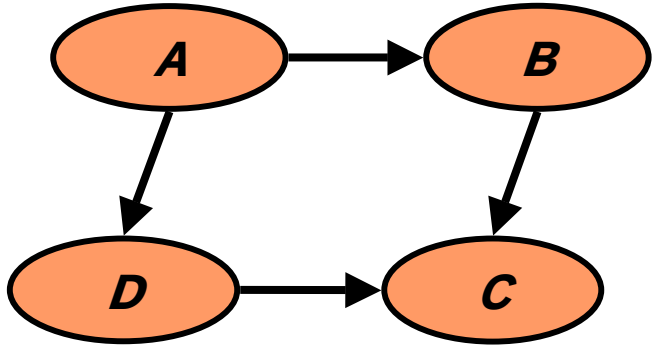
$$B \perp\!\!\!\perp C$$

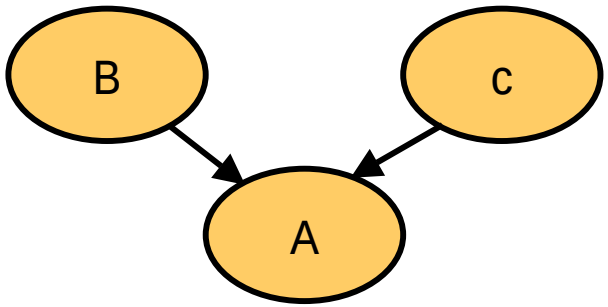
- No undirected model can represent these and only these independencies !

More

$B \perp\!\!\!\perp D \mid A$

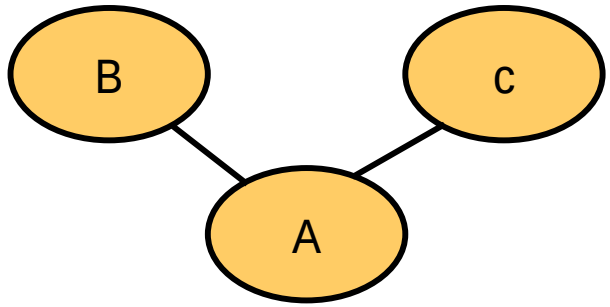
$B \neg\perp\!\!\!\perp D \mid (A, C)$





$B \perp\!\!\!\perp C$

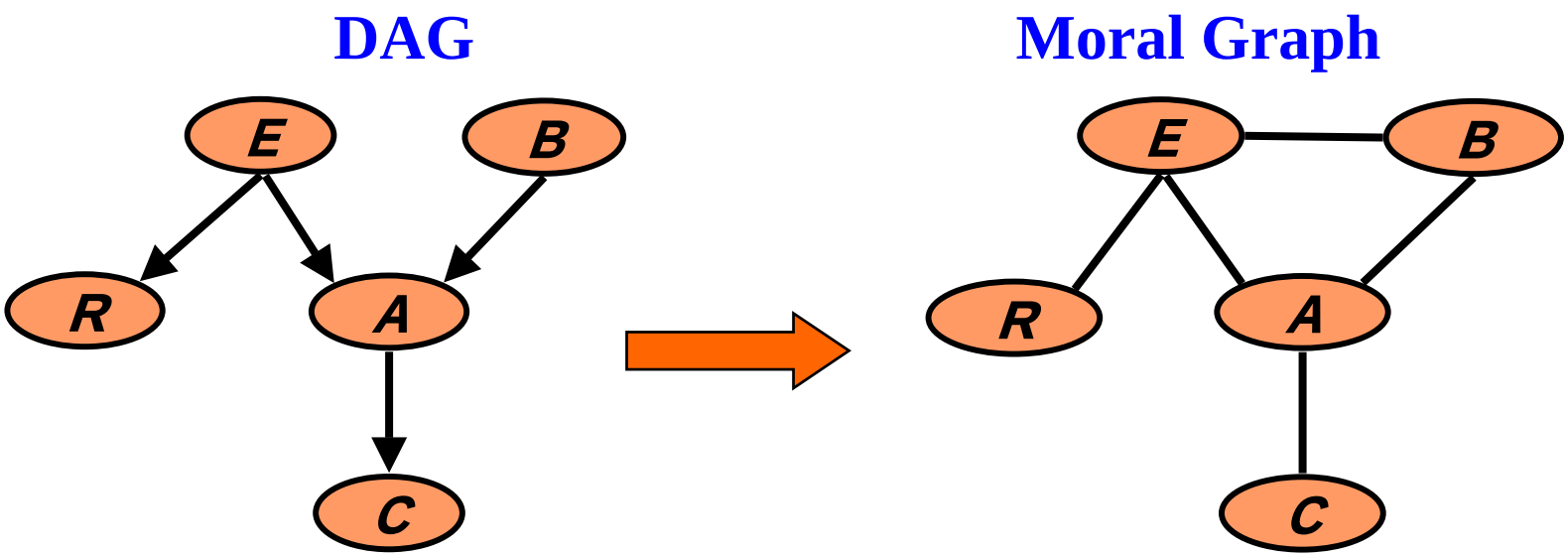
$B \neg\perp\!\!\!\perp C \mid A$



$B \neg\perp\!\!\!\perp C$

$B \perp\!\!\!\perp C \mid A$

Cont.



$B \perp\!\!\!\perp E$ in DAG is not satisfied by the moral graph

Relationship between Directed & Undirected Models