CS7015 (Deep Learning) : Lecture 18

Markov Networks

Mitesh M. Khapra

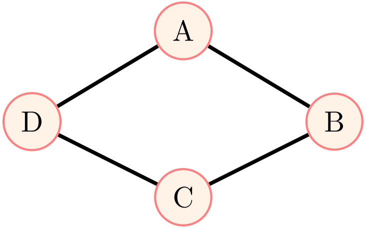
Department of Computer Science and Engineering
Indian Institute of Technology Madras

Acknowledgments

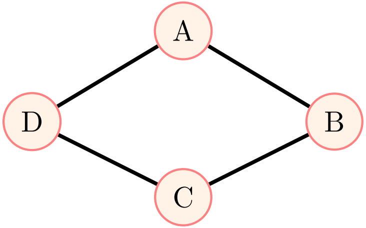
- Probabilistic Graphical models: Principles and Techniques, Daphne Koller and Nir Friedman

Module 18.1: Markov Networks: Motivation

- To motivate undirected graphical models let us consider a new example

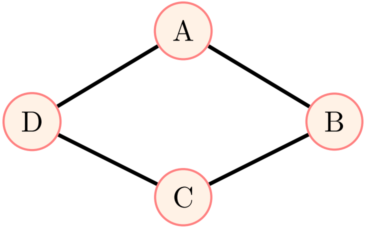


- To motivate undirected graphical models let us consider a new example

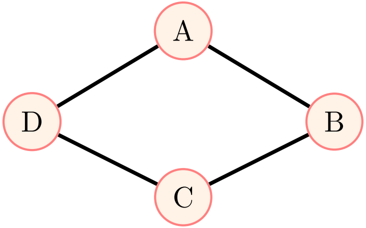


- A, B, C, D are four students

- To motivate undirected graphical models let us consider a new example

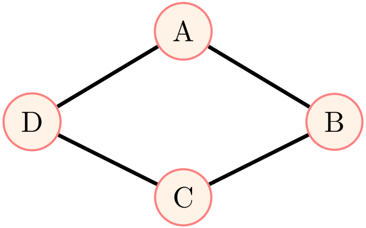


- A, B, C, D are four students
- A and B study together sometimes
- To motivate undirected graphical models let us consider a new example



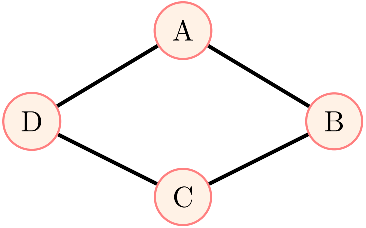
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes

- To motivate undirected graphical models let us consider a new example



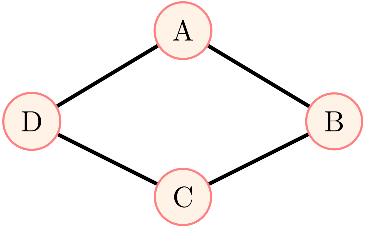
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes

- To motivate undirected graphical models let us consider a new example



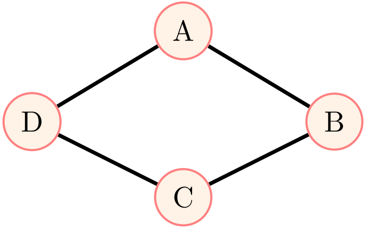
- To motivate undirected graphical models let us consider a new example

- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes



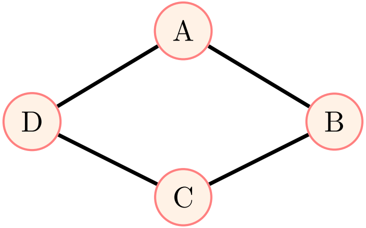
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together

- To motivate undirected graphical models let us consider a new example



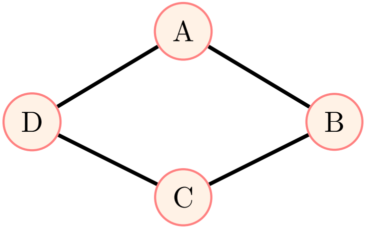
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- To motivate undirected graphical models let us consider a new example



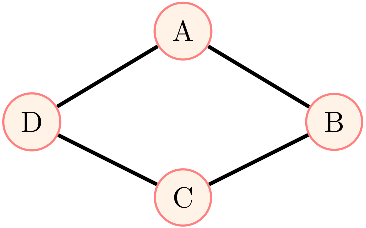
- To motivate undirected graphical models let us consider a new example

- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together



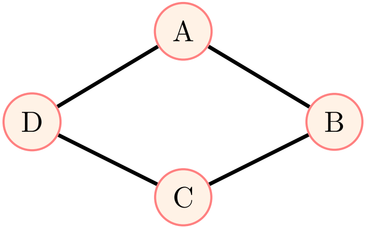
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- To motivate undirected graphical models let us consider a new example
- Now suppose there was some misconception in the lecture due to some error made by the teacher



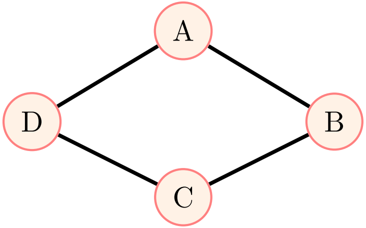
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- To motivate undirected graphical models let us consider a new example
- Now suppose there was some misconception in the lecture due to some error made by the teacher
- Each one of A, B, C, D could have independently cleared this misconception by thinking about it after the lecture



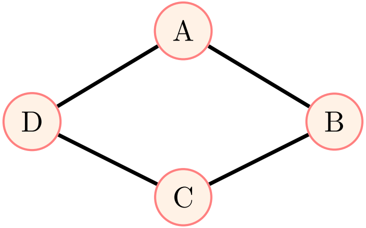
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- To motivate undirected graphical models let us consider a new example
- Now suppose there was some misconception in the lecture due to some error made by the teacher
- Each one of A, B, C, D could have independently cleared this misconception by thinking about it after the lecture
- In subsequent study pairs, each student could then pass on this information to their partner



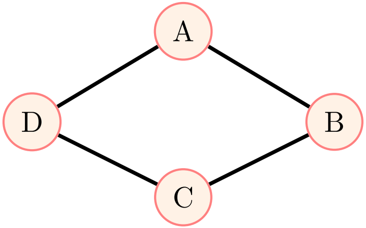
- We are now interested in knowing whether a student still has the misconception or not

- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together



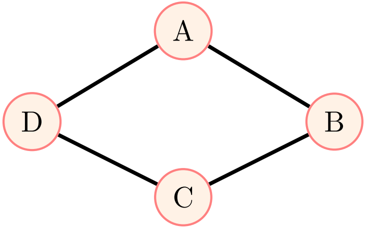
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- We are now interested in knowing whether a student still has the misconception or not
- Or we are interested in $P(A, B, C, D)$



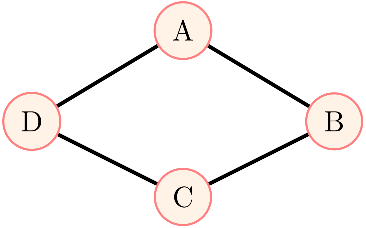
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- We are now interested in knowing whether a student still has the misconception or not
- Or we are interested in $P(A, B, C, D)$
- where A, B, C, D can take values 0 (no misconception) or 1 (misconception)



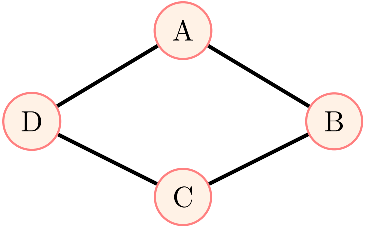
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- We are now interested in knowing whether a student still has the misconception or not
- Or we are interested in $P(A, B, C, D)$
- where A, B, C, D can take values 0 (no misconception) or 1 (misconception)
- How do we model this using a Bayesian Network ?



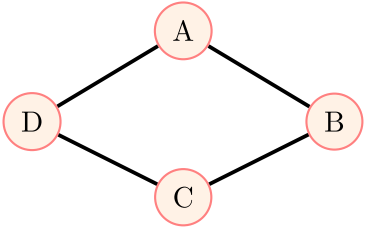
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- First let us examine the conditional independencies in this problem



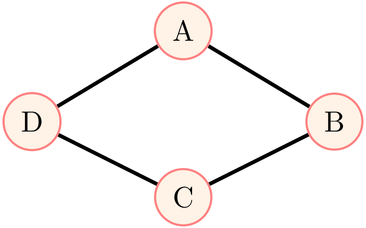
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- First let us examine the conditional independencies in this problem
- $A \perp C|\{B, D\}$ (because A & C never interact)



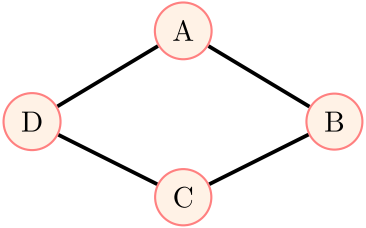
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- First let us examine the conditional independencies in this problem
- $A \perp C | \{B, D\}$ (because A & C never interact)
- $B \perp D | \{A, C\}$ (because B & D never interact)



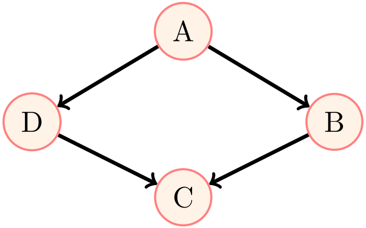
- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- First let us examine the conditional independencies in this problem
- $A \perp C | \{B, D\}$ (because A & C never interact)
- $B \perp D | \{A, C\}$ (because B & D never interact)
- There are no other conditional independencies in the problem

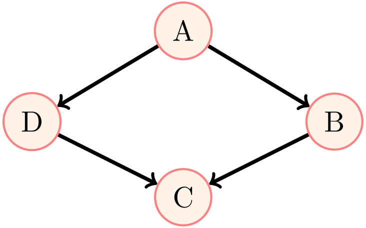


- A, B, C, D are four students
- A and B study together sometimes
- B and C study together sometimes
- C and D study together sometimes
- A and D study together sometimes
- A and C never study together
- B and D never study together

- First let us examine the conditional independencies in this problem
- $A \perp C | \{B, D\}$ (because A & C never interact)
- $B \perp D | \{A, C\}$ (because B & D never interact)
- There are no other conditional independencies in the problem
- Now let us try to represent this using a Bayesian Network

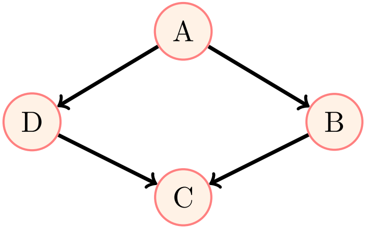


- How about this one?



- How about this one?
- Indeed, it captures the following independencies relation

$$A \perp C | \{B, D\}$$



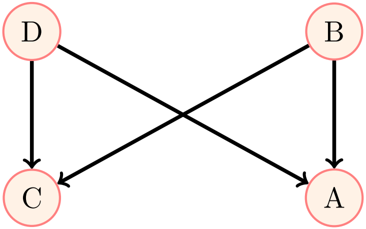
- How about this one?
- Indeed, it captures the following independencies relation

$$A \perp C | \{B, D\}$$

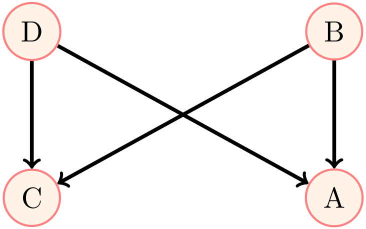
- But, it also implies that

$$B \not\perp D | \{A, C\}$$

- Let us try a different network

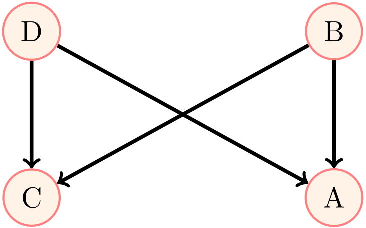


- Let us try a different network



- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

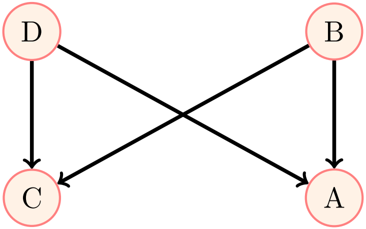


- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

- But

$$B \perp D (\text{unconditional})$$



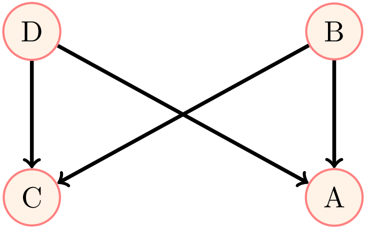
- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

- But

$$B \perp D (\text{unconditional})$$

- You can try other networks



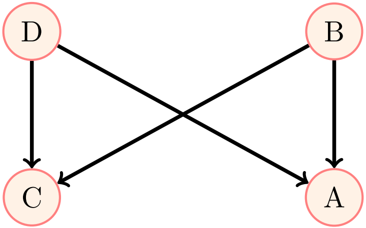
- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

- But

$$B \perp D(\text{unconditional})$$

- You can try other networks
- Turns out there is no Bayesian Network which can exactly capture independence relations that we are interested in



- **Perfect Map**: A graph G is a Perfect Map for a distribution P if the independance relations implied by the graph are exactly the same as those implied by the distribution

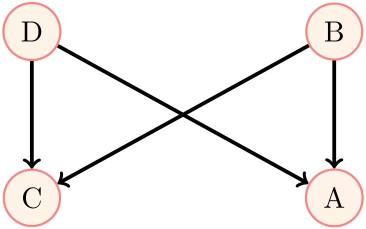
- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

- But

$$B \perp D (\text{unconditional})$$

- You can try other networks
- Turns out there is no Bayesian Network which can exactly capture independence relations that we are interested in



- **Perfect Map**: A graph G is a Perfect Map for a distribution P if the independance relations implied by the graph are exactly the same as those implied by the distribution

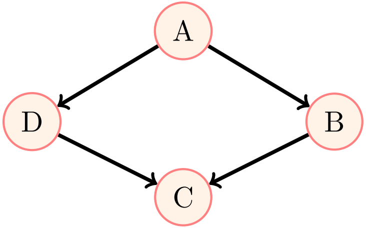
- Let us try a different network
- Again

$$A \perp C | \{B, D\}$$

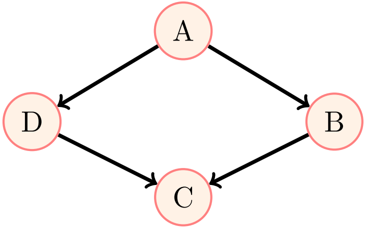
- But

$$B \perp D(\text{unconditional})$$

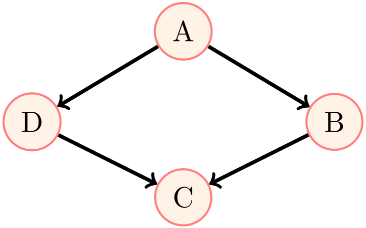
- You can try other networks
- Turns out there is no Bayesian Network which can exactly capture independence relations that we are interested in
- There is no Perfect Map for the distribution



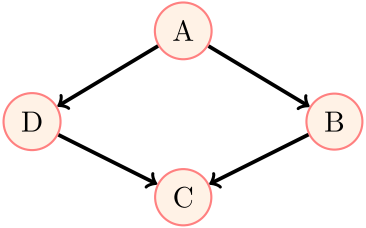
- The problem is that a directed graphical model is not suitable for this example



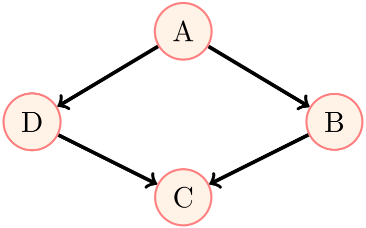
- The problem is that a directed graphical model is not suitable for this example
- A directed edge between two nodes implies some kind of direction in the interaction



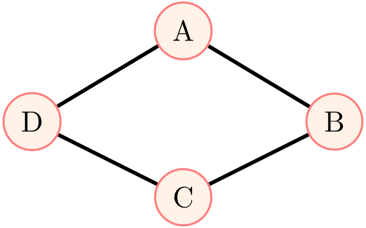
- The problem is that a directed graphical model is not suitable for this example
- A directed edge between two nodes implies some kind of direction in the interaction
- For example $A \rightarrow B$ could indicate that A influences B but not the other way round



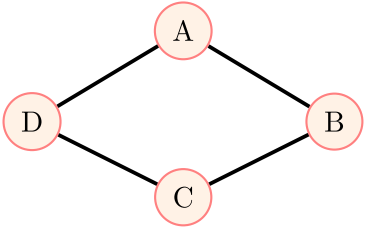
- The problem is that a directed graphical model is not suitable for this example
- A directed edge between two nodes implies some kind of direction in the interaction
- For example $A \rightarrow B$ could indicate that A influences B but not the other way round
- But in our example $A\&B$ are equal partners (they both contribute to the study discussion)



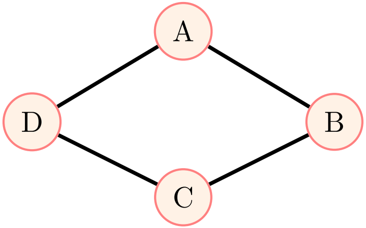
- The problem is that a directed graphical model is not suitable for this example
- A directed edge between two nodes implies some kind of direction in the interaction
- For example $A \to B$ could indicate that A influences B but not the other way round
- But in our example $A\&B$ are equal partners (they both contribute to the study discussion)
- We want to capture the strength of this interaction (and there is no direction here)



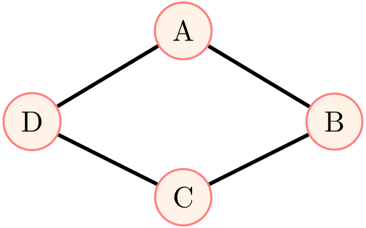
- We move on from Directed Graphical Models to Undirected Graphical Models



- We move on from Directed Graphical Models to Undirected Graphical Models
- Also known as **Markov Network**

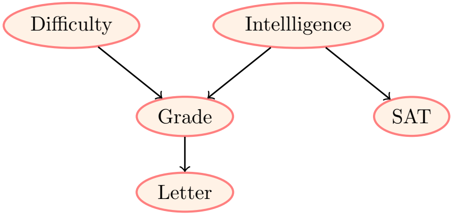


- We move on from Directed Graphical Models to Undirected Graphical Models
- Also known as **Markov Network**
- The Markov Network on the left exactly captures the interactions inherent in the problem



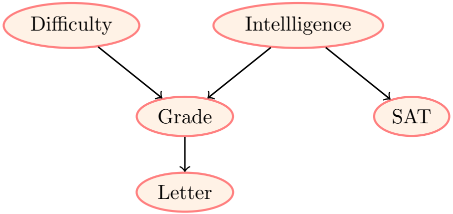
- We move on from Directed Graphical Models to Undirected Graphical Models
- Also known as **Markov Network**
- The Markov Network on the left exactly captures the interactions inherent in the problem
- But how do we parameterize this graph?

Module 18.2: Factors in Markov Network



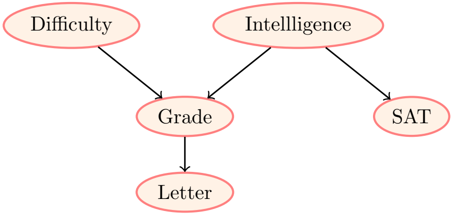
- Recall that in the directed case the factors were Conditional Probability Distributions (CPDs)

$$P(G,S,I,L,D) = P(I)P(D)P(G|I,D)P(S|I)P(L|G)$$



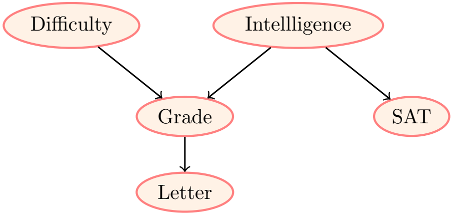
$$P(G, S, I, L, D) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)$$

- Recall that in the directed case the factors were Conditional Probability Distributions (CPDs)
- Each such factor captured interaction (dependence) between the connected nodes



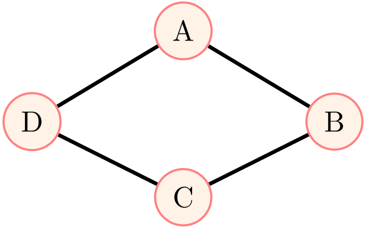
$$P(G,S,I,L,D) = P(I)P(D)P(G|I,D)P(S|I)P(L|G)$$

- Recall that in the directed case the factors were Conditional Probability Distributions (CPDs)
- Each such factor captured interaction (dependence) between the connected nodes
- Can we use CPDs in the undirected case also ?

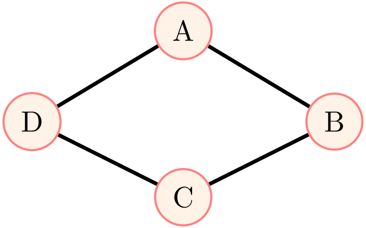


$$P(G,S,I,L,D) = P(I)P(D)P(G|I,D)P(S|I)P(L|G)$$

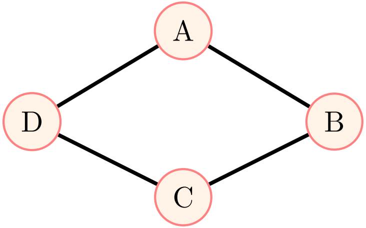
- Recall that in the directed case the factors were Conditional Probability Distributions (CPDs)
- Each such factor captured interaction (dependence) between the connected nodes
- Can we use CPDs in the undirected case also ?
- CPDs don't make sense in the undirected case because there is no direction and hence no natural conditioning (Is $A|B$ or $B|A$?)



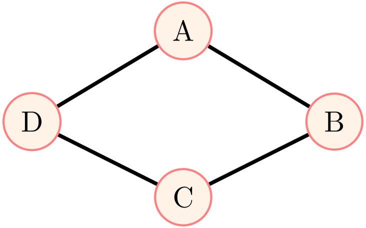
- So what should be the factors or parameters in this case



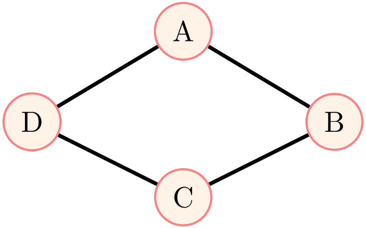
- So what should be the factors or parameters in this case
- **Question:** What do we want these factors to capture ?



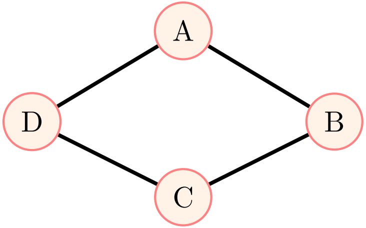
- So what should be the factors or parameters in this case
- **Question:** What do we want these factors to capture ?
- **Answer:** The affinity between connected random variables



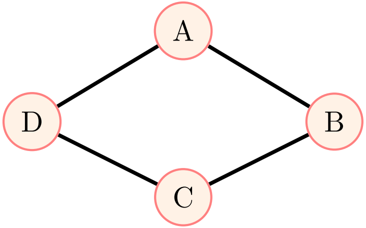
- So what should be the factors or parameters in this case
- **Question:** What do we want these factors to capture ?
- **Answer:** The affinity between connected random variables
- Just as in the directed case the factors captured the conditional dependence between a set of random variables, here we want them to capture the affinity between them



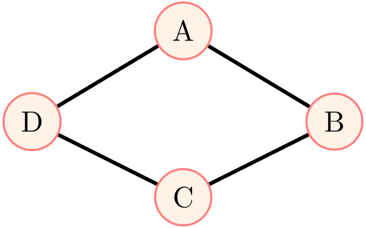
- However we can borrow the intuition from the directed case.



- However we can borrow the intuition from the directed case.
- Even in the undirected case, we want each such factor to capture interactions (affinity) between connected nodes

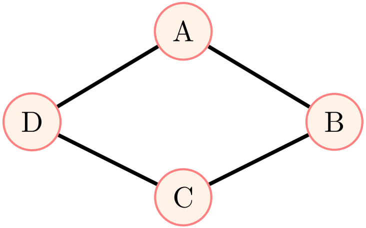


- However we can borrow the intuition from the directed case.
- Even in the undirected case, we want each such factor to capture interactions (affinity) between connected nodes
- We could have factors $\phi_1(A, B)$, $\phi_2(B, C)$, $\phi_3(C, D)$, $\phi_4(D, A)$ which capture the affinity between the corresponding nodes.



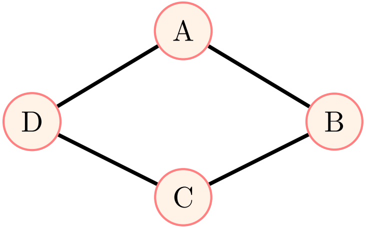
$\phi_1(A,B)$		$\phi_2(B,C)$		$\phi_3(C,D)$		$\phi_4(D,A)$	
a^0	b^0	a^0	b^0	a^0	b^0	a^0	b^0
a^0	b^1	a^0	b^1	a^0	b^1	a^0	b^1
a^1	b^0	a^1	b^0	a^1	b^0	a^1	b^0
a^1	b^1	a^1	b^1	a^1	b^1	a^1	b^1

- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

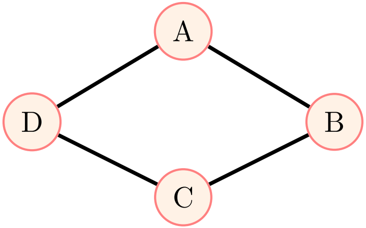
- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But who will give us these values ?

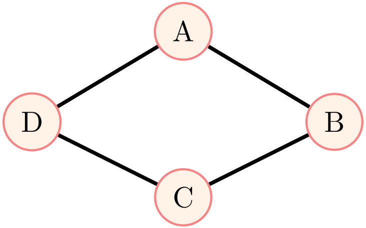
- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors



- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors

$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

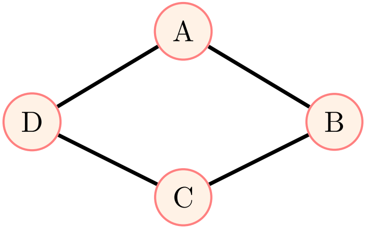
- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)



- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors

$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

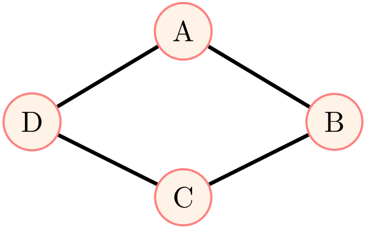
- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)
- If you have access to a lot of past interactions between $A\&B$ then you could learn these values(more on this later)



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)
- If you have access to a lot of past interactions between $A\&B$ then you could learn these values(more on this later)

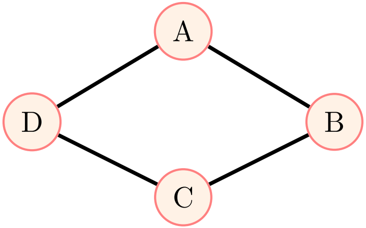
- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors
- Roughly speaking $\phi_1(A,B)$ asserts that it is more likely for A and B to agree [$\because$ weights for $a^0b^0, a^1b^1 > a^0b^1, a^1b^0$]



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)
- If you have access to a lot of past interactions between $A\&B$ then you could learn these values(more on this later)

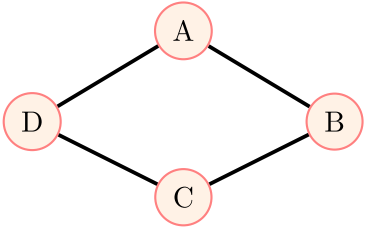
- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors
- Roughly speaking $\phi_1(A,B)$ asserts that it is more likely for A and B to agree [$\because$ weights for $a^0b^0, a^1b^1 > a^0b^1, a^1b^0$]
- $\phi_1(A,B)$ also assigns more weight to the case when both do not have a misconception as compared to the case when both have the misconception $a^0b^0 > a^1b^1$



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

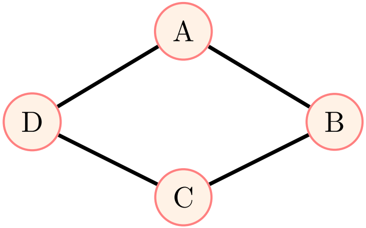
- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)
- If you have access to a lot of past interactions between $A\&B$ then you could learn these values(more on this later)

- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors
- Roughly speaking $\phi_1(A,B)$ asserts that it is more likely for A and B to agree [$\because$ weights for $a^0b^0, a^1b^1 > a^0b^1, a^1b^0$]
- $\phi_1(A,B)$ also assigns more weight to the case when both do not have a misconception as compared to the case when both have the misconception $a^0b^0 > a^1b^1$
- We could have similar assignments for the other factors



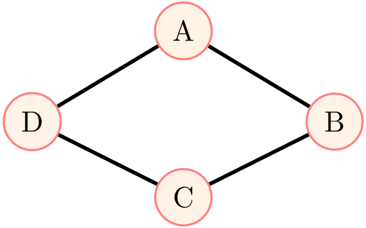
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- Notice a few things



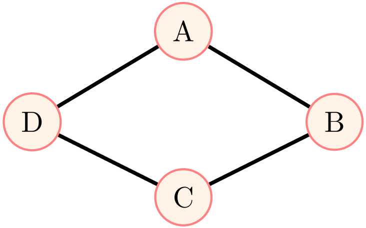
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- Notice a few things
- These tables do not represent probability distributions



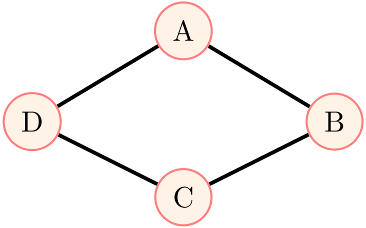
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- Notice a few things
- These tables do not represent probability distributions
- They are just weights which can be interpreted as the relative likelihood of an event



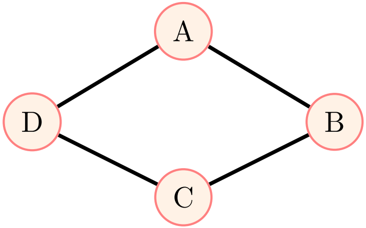
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- Notice a few things
- These tables do not represent probability distributions
- They are just weights which can be interpreted as the relative likelihood of an event
- For example, $a = 0, b = 0$ is more likely than $a = 1, b = 1$



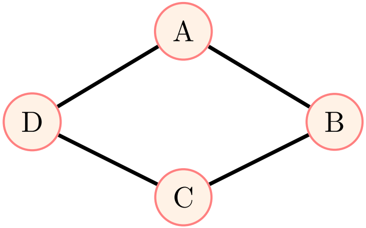
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But eventually we are interested in probability distributions



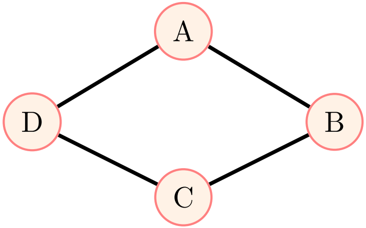
$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But eventually we are interested in probability distributions
- In the directed case going from factors to a joint probability distribution was easy as the factors were themselves conditional probability distributions



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But eventually we are interested in probability distributions
- In the directed case going from factors to a joint probability distribution was easy as the factors were themselves conditional probability distributions
- We could just write the joint probability distribution as the product of the factors (without violating the axioms of probability)



$\phi_1(A,B)$			$\phi_2(B,C)$			$\phi_3(C,D)$			$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

- But eventually we are interested in probability distributions
- In the directed case going from factors to a joint probability distribution was easy as the factors were themselves conditional probability distributions
- We could just write the joint probability distribution as the product of the factors (without violating the axioms of probability)
- What do we do in this case when the factors are not probability distributions

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300,000	4.17E-02
a^0	b^0	c^0	d^1	300,000	4.17E-02
a^0	b^0	c^1	d^0	300,000	4.17E-02
a^0	b^0	c^1	d^1	30	4.17E-06
a^0	b^1	c^0	d^0	500	6.94E-05
a^0	b^1	c^0	d^1	500	6.94E-05
a^0	b^1	c^1	d^0	5,000,000	6.94E-01
a^0	b^1	c^1	d^1	500	6.94E-05
a^1	b^0	c^0	d^0	100	1.39E-05
a^1	b^0	c^0	d^1	1,000,000	1.39E-01
a^1	b^0	c^1	d^0	100	1.39E-05
a^1	b^0	c^1	d^1	100	1.39E-05
a^1	b^1	c^0	d^0	10	1.39E-06
a^1	b^1	c^0	d^1	100,000	1.39E-02
a^1	b^1	c^1	d^0	100,000	1.39E-02
a^1	b^1	c^1	d^1	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300,000	4.17E-02
a^0	b^0	c^0	d^1	300,000	4.17E-02
a^0	b^0	c^1	d^0	300,000	4.17E-02
a^0	b^0	c^1	d^1	30	4.17E-06
a^0	b^1	c^0	d^0	500	6.94E-05
a^0	b^1	c^0	d^1	500	6.94E-05
a^0	b^1	c^1	d^0	5,000,000	6.94E-01
a^0	b^1	c^1	d^1	500	6.94E-05
a^1	b^0	c^0	d^0	100	1.39E-05
a^1	b^0	c^0	d^1	1,000,000	1.39E-01
a^1	b^0	c^1	d^0	100	1.39E-05
a^1	b^0	c^1	d^1	100	1.39E-05
a^1	b^1	c^0	d^0	10	1.39E-06
a^1	b^1	c^0	d^1	100,000	1.39E-02
a^1	b^1	c^1	d^0	100,000	1.39E-02
a^1	b^1	c^1	d^1	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300,000	4.17E-02
a^0	b^0	c^0	d^1	300,000	4.17E-02
a^0	b^0	c^1	d^0	300,000	4.17E-02
a^0	b^0	c^1	d^1	30	4.17E-06
a^0	b^1	c^0	d^0	500	6.94E-05
a^0	b^1	c^0	d^1	500	6.94E-05
a^0	b^1	c^1	d^0	5,000,000	6.94E-01
a^0	b^1	c^1	d^1	500	6.94E-05
a^1	b^0	c^0	d^0	100	1.39E-05
a^1	b^0	c^0	d^1	1,000,000	1.39E-01
a^1	b^0	c^1	d^0	100	1.39E-05
a^1	b^0	c^1	d^1	100	1.39E-05
a^1	b^1	c^0	d^0	10	1.39E-06
a^1	b^1	c^0	d^1	100,000	1.39E-02
a^1	b^1	c^1	d^0	100,000	1.39E-02
a^1	b^1	c^1	d^1	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

where

$$Z = \sum_{a,b,c,d} \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300,000	4.17E-02
a^0	b^0	c^0	d^1	300,000	4.17E-02
a^0	b^0	c^1	d^0	300,000	4.17E-02
a^0	b^0	c^1	d^1	30	4.17E-06
a^0	b^1	c^0	d^0	500	6.94E-05
a^0	b^1	c^0	d^1	500	6.94E-05
a^0	b^1	c^1	d^0	5,000,000	6.94E-01
a^0	b^1	c^1	d^1	500	6.94E-05
a^1	b^0	c^0	d^0	100	1.39E-05
a^1	b^0	c^0	d^1	1,000,000	1.39E-01
a^1	b^0	c^1	d^0	100	1.39E-05
a^1	b^0	c^1	d^1	100	1.39E-05
a^1	b^1	c^0	d^0	10	1.39E-06
a^1	b^1	c^0	d^1	100,000	1.39E-02
a^1	b^1	c^1	d^0	100,000	1.39E-02
a^1	b^1	c^1	d^1	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

where

$$Z = \sum_{a,b,c,d} \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

- Based on the values that we had assigned to the factors we can now compute the full joint probability distribution

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300,000	4.17E-02
a^0	b^0	c^0	d^1	300,000	4.17E-02
a^0	b^0	c^1	d^0	300,000	4.17E-02
a^0	b^0	c^1	d^1	30	4.17E-06
a^0	b^1	c^0	d^0	500	6.94E-05
a^0	b^1	c^0	d^1	500	6.94E-05
a^0	b^1	c^1	d^0	5,000,000	6.94E-01
a^0	b^1	c^1	d^1	500	6.94E-05
a^1	b^0	c^0	d^0	100	1.39E-05
a^1	b^0	c^0	d^1	1,000,000	1.39E-01
a^1	b^0	c^1	d^0	100	1.39E-05
a^1	b^0	c^1	d^1	100	1.39E-05
a^1	b^1	c^0	d^0	10	1.39E-06
a^1	b^1	c^0	d^1	100,000	1.39E-02
a^1	b^1	c^1	d^0	100,000	1.39E-02
a^1	b^1	c^1	d^1	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

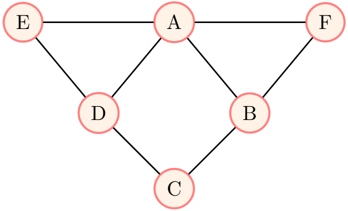
$$P(a,b,c,d) = \frac{1}{Z}\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

where

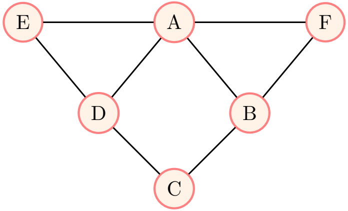
$$Z = \sum_{a,b,c,d} \phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)$$

- Based on the values that we had assigned to the factors we can now compute the full joint probability distribution
- Z is called the partition function.

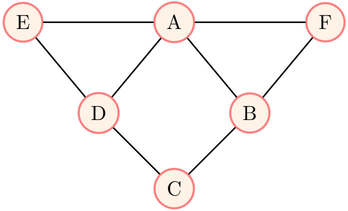
- Let us build on the original example by adding some more students



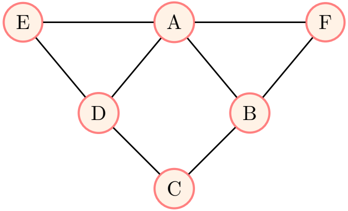
- Let us build on the original example by adding some more students



- Let us build on the original example by adding some more students
- Once again there is an edge between two students if they study together

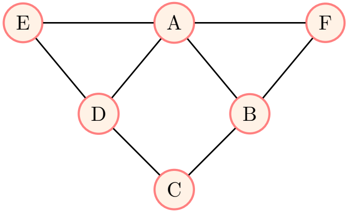


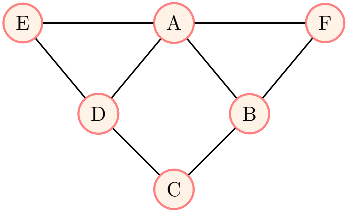
- Let us build on the original example by adding some more students
- Once again there is an edge between two students if they study together
- One way of interpreting these new connections is that $\{A, D, E\}$ from a study group or a clique



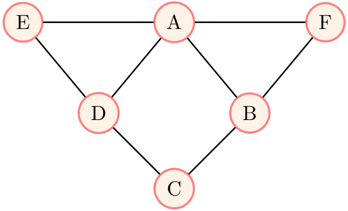
- Let us build on the original example by adding some more students
- Once again there is an edge between two students if they study together
- One way of interpreting these new connections is that $\{A, D, E\}$ from a study group or a clique
- Similarly $\{A, F, B\}$ form a study group and $\{C, D\}$ form a study group and $\{B, C\}$ form a study group

- Now, what should the factors be?



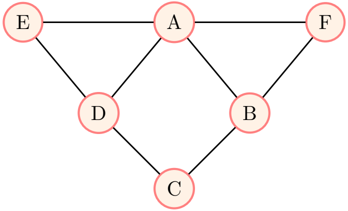


- Now, what should the factors be?
- We could still have factors which capture pairwise interactions



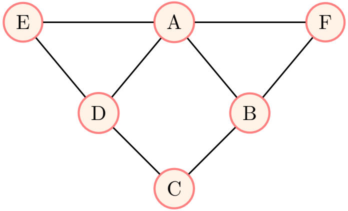
$$\phi_1(A,E)\phi_2(A,F)\phi_3(B,F)\phi_4(A,B)$$
$$\phi_5(A,D)\phi_6(D,E)\phi_7(B,C)\phi_8(C,D)$$

- Now, what should the factors be?
- We could still have factors which capture pairwise interactions



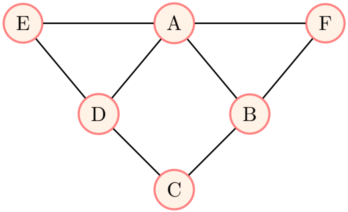
$\phi_1(A,E)\phi_2(A,F)\phi_3(B,F)\phi_4(A,B)$
$\phi_5(A,D)\phi_6(D,E)\phi_7(B,C)\phi_8(C,D)$

- Now, what should the factors be?
- We could still have factors which capture pairwise interactions
- But could we do something smarter (and more efficient)



$\phi_1(A,E)\phi_2(A,F)\phi_3(B,F)\phi_4(A,B)$
$\phi_5(A,D)\phi_6(D,E)\phi_7(B,C)\phi_8(C,D)$

- Now, what should the factors be?
- We could still have factors which capture pairwise interactions
- But could we do something smarter (and more efficient)
- Instead of having a factor for each pair of nodes why not have it for each maximal clique?



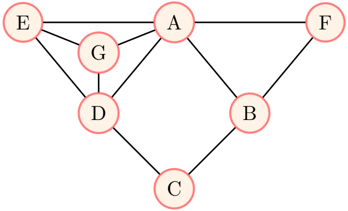
$$\phi_1(A,E)\phi_2(A,F)\phi_3(B,F)\phi_4(A,B)$$
$$\phi_5(A,D)\phi_6(D,E)\phi_7(B,C)\phi_8(C,D)$$

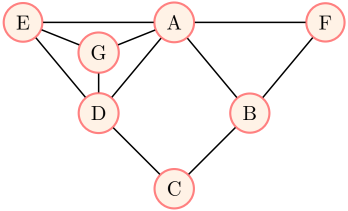
$$\phi_1(A,E,D)\phi_2(A,F,B)\phi_3(B,C)\phi_4(C,D)$$

- Now, what should the factors be?
- We could still have factors which capture pairwise interactions
- But could we do something smarter (and more efficient)
- Instead of having a factor for each pair of nodes why not have it for each maximal clique?

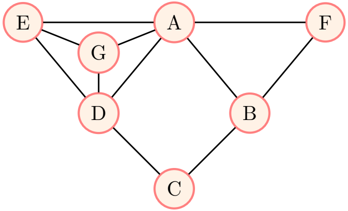
- What if we add one more student?

- What if we add one more student?

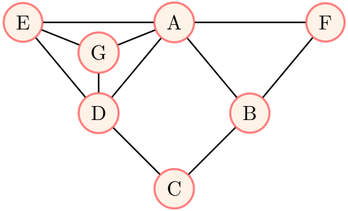




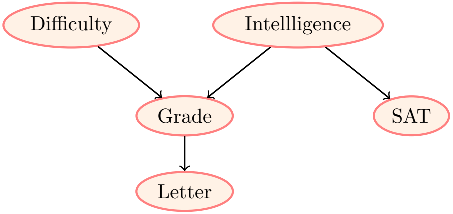
- What if we add one more student?
- What will be the factors in this case?

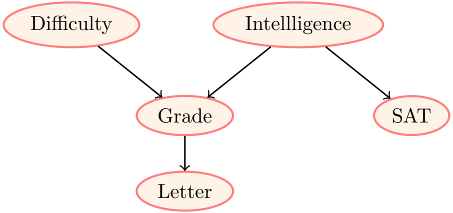


- What if we add one more student?
- What will be the factors in this case?
- Remember, we are interested in maximal cliques



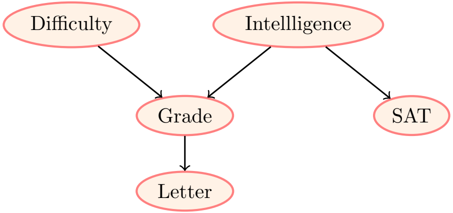
- What if we add one more student?
- What will be the factors in this case?
- Remember, we are interested in maximal cliques
- So instead of having factors $\phi(EAG)$ $\phi(GAD)$ $\phi(EGD)$ we will have a single factor $\phi(AEGD)$ corresponding to the maximal clique

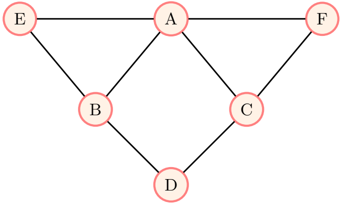

Difficulty
Intellligence
Grade
SAT
Letter



- A distribution P factorizes over a Bayesian Network G if P can be expressed as

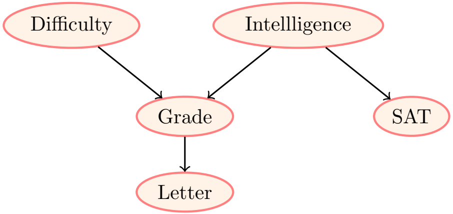
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | P_{a_{X_i}})$$





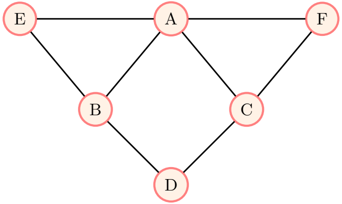
- A distribution P factorizes over a Bayesian Network G if P can be expressed as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | P_{a_{X_i}})$$



- A distribution P factorizes over a Bayesian Network G if P can be expressed as

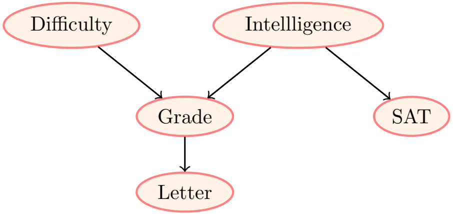
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | P_{a_{X_i}})$$

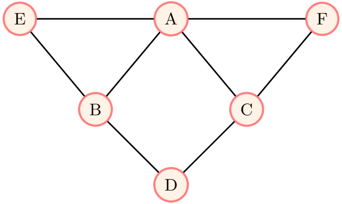


- A distribution factorizes over a Markov Network H if P can be expressed as

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi(D_i)$$

where each D_i is a complete sub-graph (maximal clique) in H





- A distribution P factorizes over a Bayesian Network G if P can be expressed as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | P_{a_{X_i}})$$

- A distribution factorizes over a Markov Network H if P can be expressed as

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi(D_i)$$

where each D_i is a complete sub-graph (maximal clique) in H

A distribution is a Gibbs distribution parametrized by a set of factors $\Phi = \{\phi_1(D_1), \ldots, \phi_m(D_m)\}$ if it is defined as

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i)$$

Module 18.3: Local Independencies in a Markov Network

- Let U be the set of all random variables in our joint distribution

- Let U be the set of all random variables in our joint distribution
- Let X, Y, Z be some distinct subsets of U

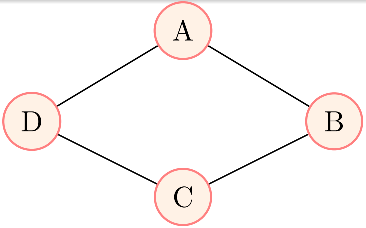
- Let U be the set of all random variables in our joint distribution
- Let X, Y, Z be some distinct subsets of U
- A distribution P over these RVs would imply $X \perp Y | Z$ if and only if we can write

$$P(X) = \phi_1(X, Z)\phi_2(Y, Z)$$

- Let U be the set of all random variables in our joint distribution
- Let X, Y, Z be some distinct subsets of U
- A distribution P over these RVs would imply $X \perp Y | Z$ if and only if we can write

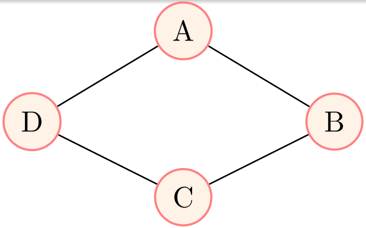
$$P(X) = \phi_1(X, Z)\phi_2(Y, Z)$$

- Let us see this in the context of our original example



- In this example

$$P(A,B,C,D) = \frac{1}{Z}[\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)]$$

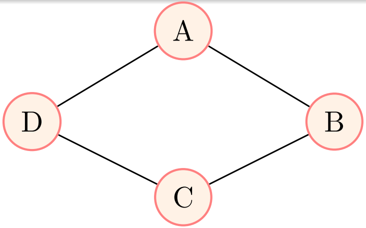


- In this example

$$P(A,B,C,D) = \frac{1}{Z}[\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)]$$

- We can rewrite this as

$$P(A,B,C,D) = \frac{1}{Z}\underbrace{[\phi_1(A,B)\phi_2(B,C)]}_{\phi_5(B,\{A,C\})}\underbrace{[\phi_3(C,D)\phi_4(D,A)]}_{\phi_6(D,\{A,C\})}$$



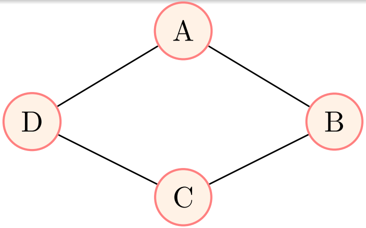
- In this example

$$P(A,B,C,D) = \frac{1}{Z}[\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)]$$

- We can rewrite this as

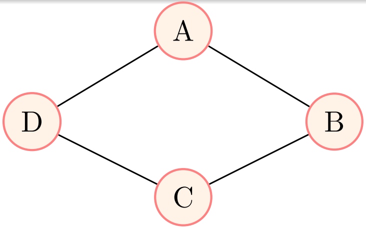
$$P(A,B,C,D) = \frac{1}{Z}\underbrace{[\phi_1(A,B)\phi_2(B,C)]}_{\phi_5(B,\{A,C\})}\underbrace{[\phi_3(C,D)\phi_4(D,A)]}_{\phi_6(D,\{A,C\})}$$

- We can say that $B \perp D | \{A, C\}$ which is indeed true



- In this example

$$P(A, B, C, D) = \frac{1}{Z}[\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)]$$

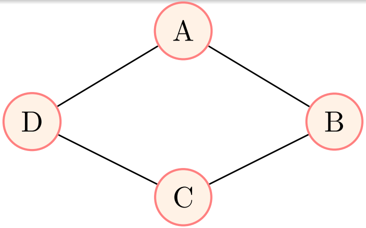


- In this example

$$P(A,B,C,D) = \frac{1}{Z}[\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(D,A)]$$

- Alternatively we can rewrite this as

$$P(A,B,C,D) = \frac{1}{Z}\underbrace{[\phi_1(A,B)\phi_2(D,A)]}_{\phi_5(A,\{B,D\})}\underbrace{[\phi_3(C,D)\phi_4(B,C)]}_{\phi_6(C,\{B,D\})}$$



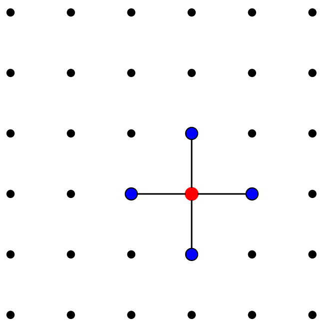
- In this example

$$P(A, B, C, D) = \frac{1}{Z}[\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)]$$

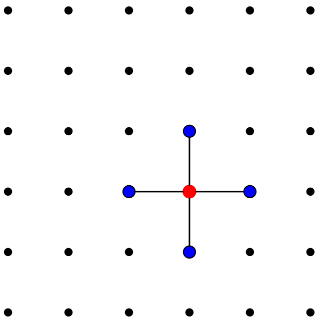
- Alternatively we can rewrite this as

$$P(A, B, C, D) = \frac{1}{Z}\underbrace{[\phi_1(A, B)\phi_2(D, A)]}_{\phi_5(A,\{B,D\})}\underbrace{[\phi_3(C, D)\phi_4(B, C)]}_{\phi_6(C,\{B,D\})}$$

- We can say that $A \perp C | \{B, D\}$ which is indeed true



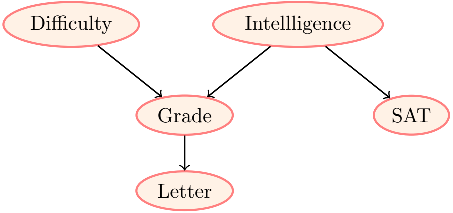
- For a given Markov network H we define Markov Blanket of a RV X to be the neighbors of X in H



- For a given Markov network H we define Markov Blanket of a RV X to be the neighbors of X in H
- Analogous to the case of Bayesian Networks we can define the local independences associated with H to be

$$X \perp (U - \{X\} - MB_H) | MB_H(X)$$

Bayesian network

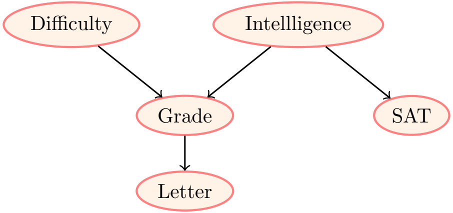


Markov network

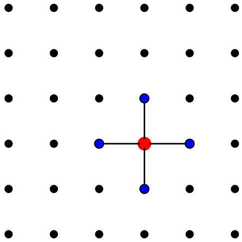
Local Independencies

$$X_i \perp NonDescendents_{X_i} | Parent^G_{X_i}$$

Bayesian network



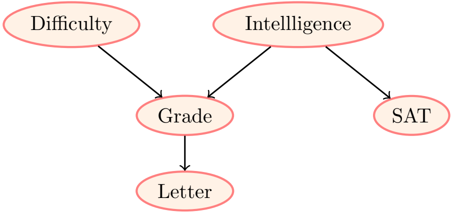
Markov network



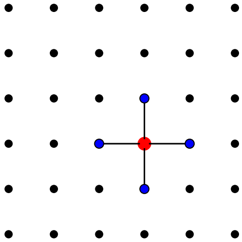
Local Independencies

$$X_i \perp NonDescendents_{X_i} | Parent_{X_i}^{G}$$

Bayesian network



Markov network



Local Independencies

$$X_i \perp NonDescendents_{X_i} | Parent^G_{X_i}$$

Local Independencies

$$X_i \perp NonNeighbors_{X_i} | Neighbors^G_{X_i}$$