# CPSC 440: Advanced Machine Learning

## Directed Acyclic Graphical Models

Mark Schmidt

University of British Columbia

Winter 2022

# Last Time: DAG Models

- Directed acyclic graphical (DAG) models write joint probability as

$$p(x_1, x_2, \dots, x_d) = \prod_{j=1}^{d} p(x_j \mid x_{\text{pa}(j)}),$$

  where pa$(j)$ are the "parents" of feature $j$.
  - Assumes independence of non-parents in $1:(j-1)$ given parents.
  - Markov chains are special case where pa$(j)$ is $(j-1)$.

- "Graphical" name comes from visualizing parents/features as a graph:
  - We have a node for each feature $j$.
  - We place an edge into $j$ from each of its parents.

- This graph is not just a visualization tool:
  - Can be used to test arbitrary conditional independences ("d-separation").
  - Graph structure tells us whether message passing is efficient ("treewidth").

# Graph Structure Examples

With product of independent we have

$$p(x) = \prod_{j=1}^{d} p(x_j),$$

so $\text{pa}(j) = \varnothing$ and the graph is:

# Graph Structure Examples

With Markov chain we have

$$p(x) = p(x_1) \prod_{j=2}^{d} p(x_j \mid x_{j-1}),$$

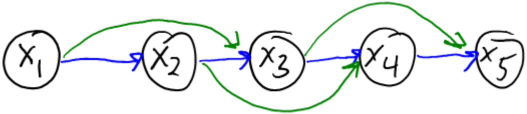
so $\text{pa}(j) = \{j-1\}$ and the graph is:

# Graph Structure Examples

With second-order Markov chain we have

$$p(x) = p(x_1)p(x_2 \mid x_1) \prod_{j=3}^{d} p(x_j \mid x_{j-1}, x_{j-2}),$$

so pa$(j) = \{j-2, j-1\}$ and the graph is:

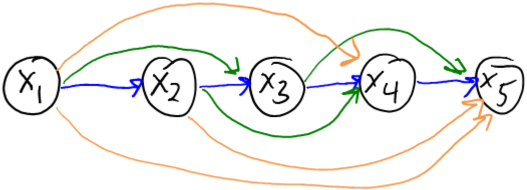


## Graph Structure Examples

With general distribution we have

$$p(x) = \prod_{j=1}^{d} p(x_j \mid x_{1:j-1}).$$

so $\mathrm{pa}(j) = \{1, 2, \ldots, j-1\}$ and the graph is:

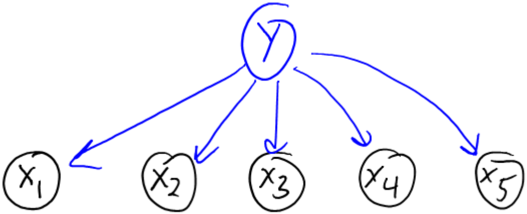


# Graph Structure Examples

In naive Bayes (or GDA with diagonal $\Sigma$) we add an extra variable $y$ and use

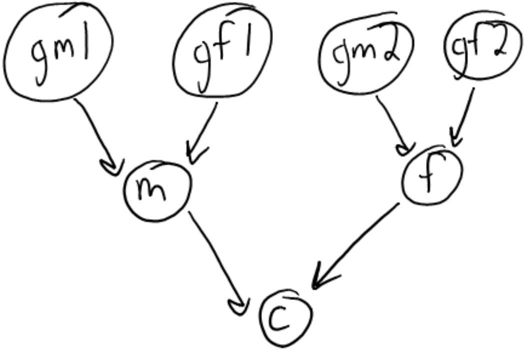
$$p(y, x) = p(y) \prod_{j=1}^{d} p(x_j \mid y),$$

which has $\text{pa}(y) = \emptyset$ and $\text{pa}(x_j) = y$ giving



# Graph Structure Examples

We can consider genetic phylogeny (family trees):




The "parents" in the graph are the actual parents.

- Independence assumption: only depend on grandparent's genes through parents.

# First DAG Model

- DAGs were first used to analyze inheritance in guinea pigs:

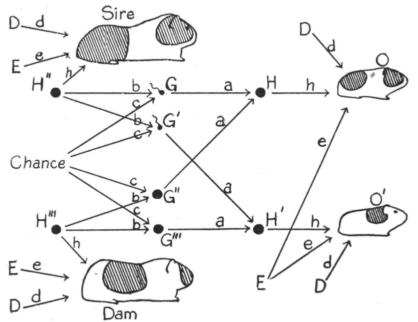


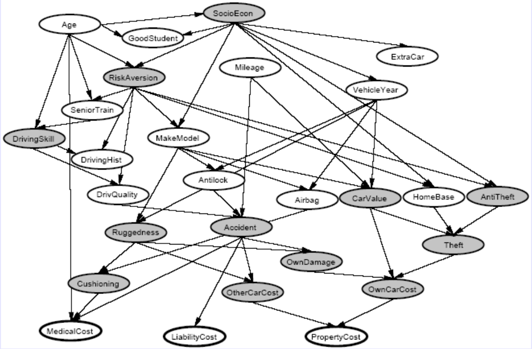

FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

https://www.pnas.org/doi/pdf/10.1073/pnas.6.6.320

# Example: Vehicle Insurance

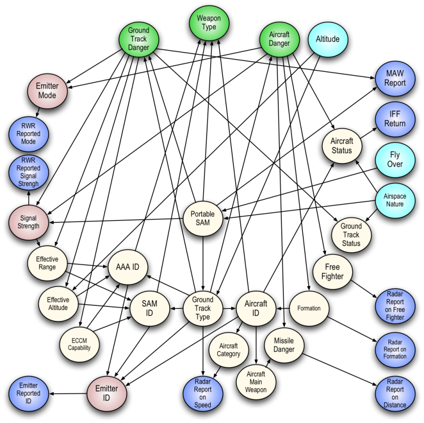
- Want to predict bottom three "cost" variables, given observed and unobserved values:




https://www.cs.princeton.edu/courses/archive/fall10/cos402/assignments/bayes

# Example: Radar and Aircraft Control

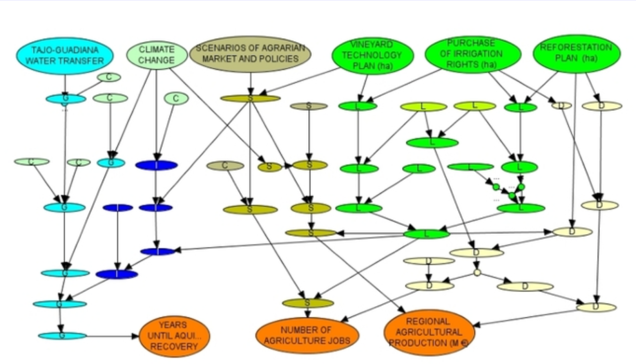
- Modeling multiple planes and radar signals:




https://pr-owl.org/basics/bn.php

# Example: Water Resource Management

- Dependencies in environmental monitor and susatainability issues:



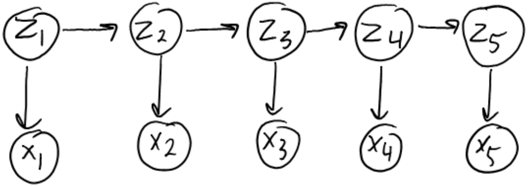

https://www.jstor.org/stable/26268156

# Outline

# Density Estimators vs. Relationship Visualizers?

- In Machine learning, DAGs are often used in two different ways:
  1. As a multivariate density estimation method.
     - We will talk inference and learning in DAGs next time.
  2. As a way to describe the relationships we are modeling.
     - All independence assumptions we have used in 340/440 have DAG representation*.
     - Includes product of Bernoullis and naive Bayes, but also IID and prior vs. hyper-prior.
     - *Except multivariate Gaussians (which can use "undirected" independence).

- For example, later we will talk about hidden Markov models (HMMs):




- The graph and variable names already give you an idea of what this model does:
  - Hidden variables $z_j$ that follow a Markov chain, with feature $x_j$ depend on $z_j$.

# Extra Conditional Independences in Markov Chains

- The Markov assumption in Markov chains is $x_j \perp x_1, x_2, \dots, x_{j-2} \mid x_{j-1}$ for all $j$
- But this implies other independences, like $x_j \perp x_1, x_2, \dots, x_{j-3} \mid x_{j-2}$.
  - We did not assume this directly, it follows from assumptions we made.
  - And we can use this property to easily compute $p(x_j \mid x_{j-2}, x_{j-3}, \dots, x_1)$:

$$
\begin{aligned}
p(x_j \mid x_{j-2}, x_{j-3}, \dots x_1) &= p(x_j \mid x_{j-2}) \\
&= \sum_{x_{j-1}} p(x_j, x_{j-1} \mid x_{j-2}) \\
&= \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2}) p(x_{j-1} \mid x_{j-2}) \\
&= \sum_{x_{j-1}} \underbrace{p(x_j \mid x_{j-1})}_{\text{tran prob}} \underbrace{p(x_{j-1} \mid x_{j-2})}_{\text{tran prob}}.
\end{aligned}
$$

- Mathematically showing extra independence assumptions is tedious (see bonus).
- But all conditional independences implied by a DAG can seen in the graph.

# D-Separation: From Graphs to Conditional Independence

- In DAGs: variables $A$ and $B$ are conditionally independent given $C$ if:
  - "D-separation blocks all undirected paths in the graph from any variable in $A$ to any variable in $B$."

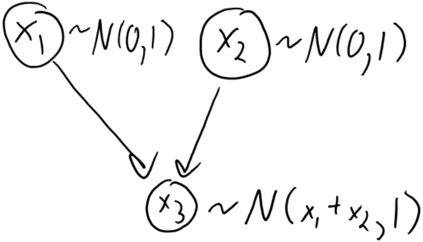
- In the special case of product of independent models our graph is:


- Here there are no paths to block, which implies the variables are independent.

- Checking paths in a graph tends to be faster than tedious calculations.

# D-Separation as Genetic Inheritance

- The rules of d-separation are intuitive in a simple model of gene inheritance:
  - Each node/person has single number, which we'll call a "gene".
  - If you have no parents, your gene is a random number.
  - If you have parents, your gene is a sum of your parents plus noise.
- For example, think of something like this:




- Graph corresponds to the factorization $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$.
  - In this model, does $p(x_1, x_2) = p(x_1)p(x_2)$? (Are $x_1$ and $x_2$ independent ?)

# D-Separation as Genetic Inheritance

- Genes of people are independent if knowing one says nothing about the other.
- Your gene is dependent on your parents:
  - If I know your parent's gene, I know something about yours.
- Your gene is independent of your (unrelated) friends:
  - If you know your friend's gene, it doesn't tell me anything about you.
- Genes of people can be conditionally independent given a third person:
  - Knowing your grandparent's gene tells you something about your gene.
  - But grandparent's gene isn't useful if you know parent's gene.

# D-Separation Case 0 (No Paths and Direct Links)

Are genes in person $x$ independent of the genes in person $y$?

- No path: $x$ and $y$ are not related (independent).


  We have $x \perp y$: there are no paths to be blocked.

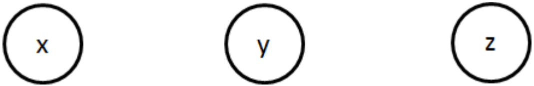
- Direct link: $x$ is the parent of $y$.


  We have $x \not\perp y$: knowing $x$ tells you about $y$ (direct paths aren't blockable).
  - And similarly knowing $y$ tells you about $x$.

# D-Separation Case 0 (No Paths and Direct Links)

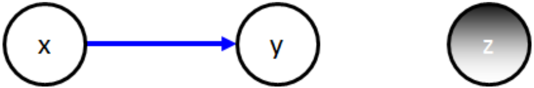
Neither case changes if we have a third independent person $z$:

- No path: If $x$ and $y$ are independent,




  We have $x \perp y$: adding $z$ doesn't make a path.

- Direct link: $x$ is the parent of $y$,

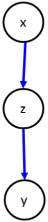



  We have $x \not\perp y \mid z$: adding $z$ doesn't block path.

  - We use **black or shaded** nodes to denote values we condition on (in this case $z$).
    - We sometimes also call the nodes that we condition on the "observations".

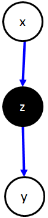
# D-Separation Case 1: Chain

- Case 1: $x$ is the grandparent of $y$.
  - If $z$ is the mother we have:




We have $x \not\perp y$: knowing $x$ would give information about $y$ because of $z$

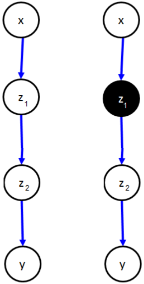
  - But if $z$ is *observed*:




In this case $x \perp y \mid z$: knowing $z$ "breaks" dependence between $x$ and $y$.

# D-Separation Case 1: Chain

- The same logic holds for great-grandparents:

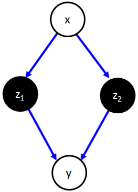



- We have $x \not\perp y$ (left), but $x \perp y \mid z_1$ (right).
  - We also have $x \perp y \mid z_2$ and that $x \perp y \mid z_1, z_2$.

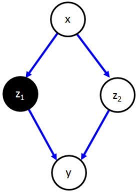
- This case lets you test any independence in Markov chains.
  - "Variables are independent conditioned on any variable inbetweeen".

# D-Separation Case 1: Chain

- Consider weird case where parents $z_1$ and $z_2$ share parent $x$:
  - If $z_1$ and $z_2$ are observed we have:



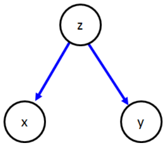

    We have $x \perp y \mid z_1, z_2$: knowing both parents breaks dependency.
  - But if only $z_1$ is *observed*:



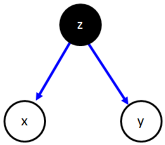

    We have $x \not\perp y \mid z_1$: dependence still "flows" through $z_2$.

# D-Separation Case 2: Common Parent

- Case 2: $x$ and $y$ are sibilings.
  - If $z$ is a common unobserved parent:



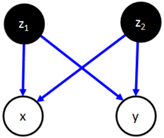

    We have $x \not\perp y$: knowing $x$ would give information about $y$.
  - But if $z$ is *observed*:




    In this case $x \perp y \mid z$: knowing $z$ "breaks" dependence between $x$ and $y$.
- This is the type of independence used in naive Bayes.

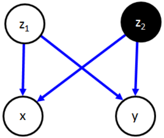
# D-Separation Case 2: Common Parent

- Case 2: $x$ and $y$ are sibilings.
  - If $z_1$ and $z_2$ are common observed parents:




We have $x \perp y \mid z_1, z_2$: knowing $z_1$ and $z_2$ breaks dependence between $x$ and $y$.

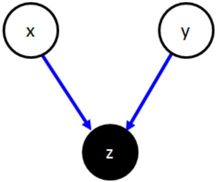
  - But if we only observe $z_2$:



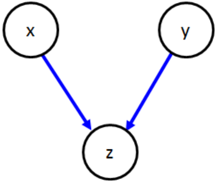

Then we have $x \not\perp y \mid z_2$: dependence still "flows" through $z_1$.

# D-Separation Case 3: Common Child

- Case 3: $x$ and $y$ share a child $z$:
  - If we observe $z$ then we have:



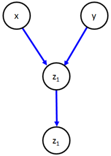

We have $x \not\perp y \mid z$: if we know $z$, then knowing $x$ gives us information about $y$.
  - But if $z$ is not observed:



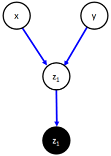

We have $x \perp y$: if you don't observe $z$ then $x$ and $y$ are independent.
- Different from Case 1 and Case 2: not observing the child blocks path.

# D-Separation Case 3: Common Child

- Case 3: $x$ and $y$ share a child $z_1$:
  - If there exists an unobserved grandchild $z_2$:



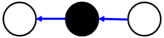

    We have $x \perp y$: the path is still blocked by not knowing $z_1$ or $z_2$.
  - But if $z_2$ is observed:




    We have $x \not\perp y \mid z_2$: grandchild creates dependence even with unobserved child.
- Case 3 needs to consider descendants of child.

# D-Separation Summary (MEMORIZE)

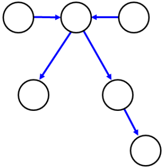
- Checking whether DAG implies $A$ is independent of $B$ given $C$:
  - Consider each undirected path from any node in any $A$ to any node in $B$.
    - Ignoring directions and observations.
  - Use directions/observations, check if any of below hold somewhere along each path:
    1. $P$ includes a "chain" with an observed middle node (e.g., Markov chain):

    2. $P$ includes a "fork" with an observed parent node (e.g., naive Bayes):

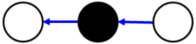
    3. $P$ includes a "v-structure" or "collider" (e.g., genetic inheritance):

       where the "child" and all its descendants are unobserved.
  - If all paths are blocked by one of above, DAG implies the conditional independence.

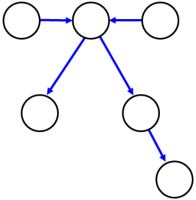
# D-Separation Summary (MEMORIZE)

- We say that $A$ and $B$ are d-separated (conditionally independent) given $C$ if *all undirected paths* from $A$ to $B$ are "blocked" because *one* of the following holds *somewhere* on the path:
  1. $P$ includes a "chain" with an observed middle node (e.g., Markov chain):



  2. $P$ includes a "fork" with an observed parent node (e.g., naive Bayes):

  3. $P$ includes a "v-structure" or "collider" (e.g., genetic inheritance):



where the "child" and all its descendants are unobserved.

# Alarm Example


- Case 1:
  - Earthquake $\not\perp$ Call.
  - Earthquake $\perp$ Call $|$ Alarm.
- Case 2:
  - Alarm $\not\perp$ Stuff Missing.
  - Alarm $\perp$ Stuff Missing $|$ Burglary.

# Alarm Example


- Case 3:
  - Earthquake $\perp$ Burglary.
  - Earthquake $\not\perp$ Burglary | Alarm.
    - "Explaining away": knowing one parent can make the other less/more likely.
- Multiple Cases:
  - Call $\not\perp$ Stuff Missing.
  - Earthquake $\perp$ Stuff Missing.
  - Earthquake $\not\perp$ Stuff Missing | Call.

# Discussion of D-Separation

- D-separation lets you say if conditional independence is implied by assumptions:

$$(A \text{ and } B \text{ are d-separated given } C) \Rightarrow A \perp B \mid C.$$

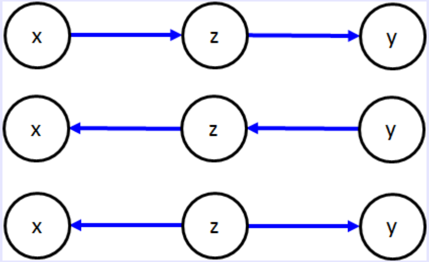
- However, there might be extra conditional independences in the distribution:
  - These would depend on specific choices of the DAG parameters.
    - For example, if we set Markov chain parameters so that $p(x_j \mid x_{j-1}) = p(x_j)$.
  - Or some *orderings* of the chain rule may reveal different independences.
  - So lack of d-separation does not imply dependence.

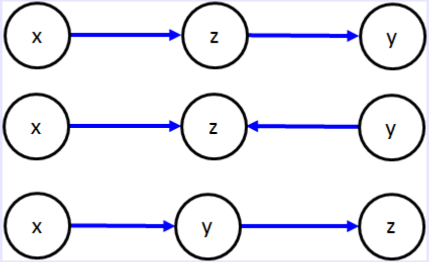
- Instead of restricting to $\{1, 2, \ldots, j-1\}$, consider general parent choices.
  - So $x_2$ could be a parent of $x_1$.
- As long the graph is acyclic, there exists a valid ordering (chain rule makes sense).

(all DAGs have a "topological order" of variables where parents are before children)

# Non-Uniqueness of Graph and Equivalent Graphs

- Note that some graphs imply same conditional independences:
  - Equivalent graphs: same v-structures and other (undirected) edges are the same.
  - Examples of 3 *equivalent* graphs (left) and 3 non-equivalent graphs (right):



# Beware of the "Causal" DAG

- It can be helpful to use the language of causality when reasoning about DAGs.
  - You'll find that they give the correct causal interpretation based on our intuition.
- However, keep in mind that the arrows are not necessarily causal.
  - "$A$ causes $B$" has the same graph as "$B$ causes $A$".
- There is work on causal DAGs which add semantics to deal with "interventions".
  - But these require assuming that the arrow directions are causal.
    - Fitting a DAG to observational data doesn't imply anything about causality.

# Summary

- DAG examples:
  - Most models can be represented as DAGs.
- D-separation allows us to test conditional independences based on graph.
  - Conditional independence follows if all undirected paths are "blocked".
  - Observed values in chain or parent block paths.
  - Unobserved children (with no observed grandchildren) also blocks paths.
- Next time: the IID assumption as a DAG.

# Extra Conditional Independences in Markov Chains

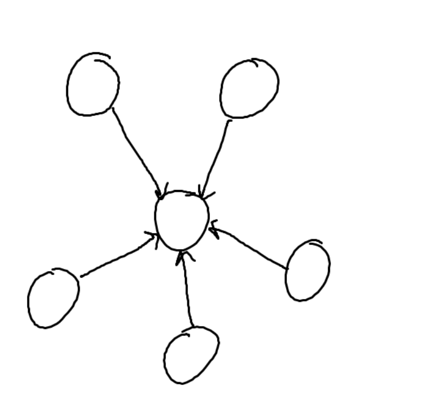
- Proof that $x_j$ is independent of $\{x_1, x_2, \dots, x_{j-3}\}$ given $x_{j-2}$ in Markov chain:

$$
\begin{aligned}
p(x_j \mid x_{j-2}, x_{j-3}, \dots, x_1) &= \frac{p(x_j, x_{j-2}, x_{j-3}, \dots, x_1)}{p(x_{j-2}, x_{j-3}, \dots, x_1)} \quad \text{(def'n cond. prob.)} \\
&= \frac{\sum_{x_{j-1}} p(x_j, x_{j-1}, x_{j-2}, \dots, x_1)}{p(x_{j-2} \mid x_{j-3}, x_{j-4}, \dots, x_1) p(x_{j-3} \mid x_{j-4}, x_{j-5}, \dots, x_1) \cdots p(x_1)} \quad \text{(marg. and chain rule)} \\
&= \frac{\sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2}) p(x_{j-1} \mid x_{j-2}) \dots p(x_2 \mid x_1) p(x_1)}{p(x_{j-2} \mid x_{j-3}) p(x_{j-3} \mid x_{j-4}) \cdots p(x_1)} \quad \text{(chain rule and Markov)} \\
&= \frac{p(x_1) p(x_2 \mid x_1) \cdots p(x_{j-2} \mid x_{j-3}) \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2}) p(x_{j-1} \mid x_{j-2})}{p(x_{j-2} \mid x_{j-3}) p(x_{j-3} \mid x_{j-4}) \cdots p(x_1)} \quad \text{(take terms outside} \\
&= \sum_{x_{j-1}} p(x_j \mid x_{j-1}, x_{j-2}) p(x_{j-1} \mid x_{j-2}) \quad \text{(cancel out in numerator/denominator)} \\
&= \sum_{x_{j-1}} p(x_j, x_{j-1} \mid x_{j-2}) \quad \text{(product rule)} \\
&= p(x_j \mid x_{j-2}) \quad \text{(marg rule).}
\end{aligned}
$$

- Similar steps could be used to show $x_j \perp x_{j+2} \mid x_{j+1}$, and a variety of other conditional independences like $x_1 \perp x_{10} \mid x_5$.

# Conditional Independence in Star Graphs

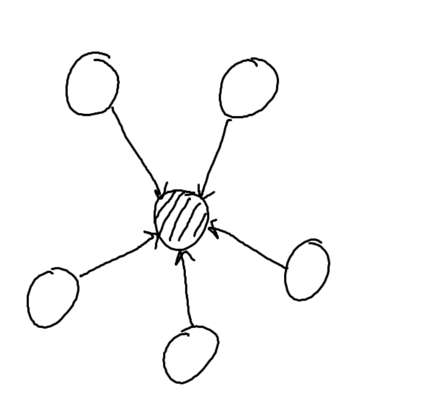
- Consider the following star graph:



- “5 aliens get together and make a baby alien”.
    - Unconditionally, the 5 aliens are independent.

# Conditional Independence in Star Graphs

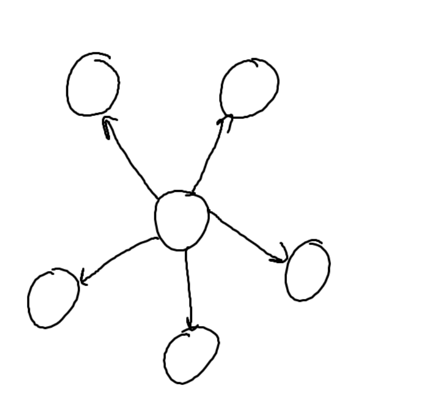
- Consider the following star graph:



- "5 aliens get together and make a baby alien".
  - Conditioned on the baby, the 5 aliens are dependent.

# Conditional Independence in Star Graphs

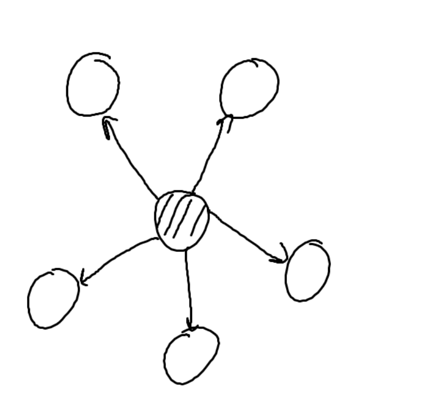
- Consider the following star graph:



- "An organism produces 5 clones".
  - Unconditionally, the 5 clones are dependent.

## Conditional Independence in Star Graphs

- Consider the following star graph:



- "An organism produces 5 clones".
  - Conditioned on the original, the 5 clones are independent.