



## (Causal) Bayesian Networks & Fairness

Silvia Chiappa

[csilvia@google.com](mailto:csilvia@google.com)

August 29-30, 2019



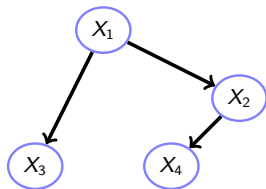
# Graphical Models

Graphical models are **graphs** in which

**Nodes** represent random variables

**Links** represent statistical dependencies between variables

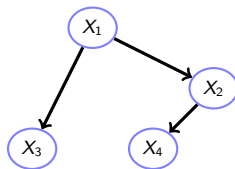
Graphical models provide us with a visual tool for reasoning under uncertainty.



# Graphical Models

Graphical models allow us to

- Answer questions about independence of random variables by just looking at the graph, without the need to perform algebraic manipulations.
- Define general algorithms that perform probabilistic inference efficiently.



As a consequence

- Provided us with a common framework for representing and understanding the properties of different probabilistic models
  - ⇒ Enabled to relate models developed in different communities
- Have accelerated progress in modeling

There exist many different types of graphical models: Bayesian (Belief) Networks, Markov Networks, Factor Graphs

We will focus on Bayesian Networks.

# Probabilistic Inference

Convert prior distribution into posterior distribution by incorporating observations.

## Bayes Rule

$$\overbrace{p(A|B)}^{\text{posterior}} = \frac{p(B|A) \overbrace{p(A)}^{\text{prior}}}{p(B)}$$

**Example of Inference:** Sally throws a die. Tom tells her that she did not score 3. What is the probability that she scored 4?

$S4 = 1$ : Score = 4,  $S3 = 0$ : Score  $\neq 3$

$$\begin{aligned} p(S4 = 1 | S3 = 0) &= \frac{p(S3 = 0 | S4 = 1) p(S4 = 1)}{p(S3 = 0)} \\ &= \frac{1/6}{1 - 1/6} = 1/5 \end{aligned}$$

## Marginal Independence of $A$ and $B$

$$A \perp\!\!\!\perp B \iff p(A|B) = p(A) \text{ or } p(A, B) = p(A)p(B)$$

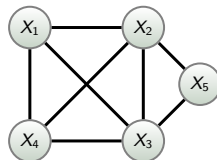
## Conditional Independence of $A$ and $B$ given $C$

$$A \perp\!\!\!\perp B | C \iff p(A|B, C) = p(A|C) \text{ or } p(A, B|C) = p(A|C)p(B|C)$$

## Basic Graph Definitions

- **Directed and Undirected Graph:** Graph with directed and undirected links respectively.
- **Path from  $X_i$  to  $X_j$ :** Sequence of connected nodes starting at  $X_i$  and ending at  $X_j$ .  
A *directed path* is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

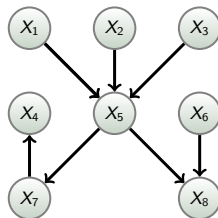
Undirected Graph



### For Directed Graphs

- **Parents and Children:**  $X_i$  is a parent of  $X_j$  if there is a link from  $X_i$  to  $X_j$ .  $X_i$  is a child of  $X_j$  if there is a link from  $X_j$  to  $X_i$ .
- **Ancestors and Descendants:** The ancestors of a node  $X_i$  are the nodes with a directed path ending at  $X_i$ . The descendants of  $X_i$  are the nodes with a directed path beginning at  $X_i$ .
- **Directed Acyclic Graph:** Graph in which by following the direction of the arrows a node will never be visited more than once.

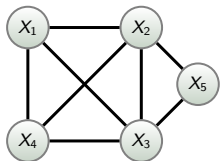
Directed Graph



## Basic Graph Definitions

- **Directed and Undirected Graph:** Graph with directed and undirected links respectively.
- **Path from  $X_i$  to  $X_j$ :** Sequence of connected nodes starting at  $X_i$  and ending at  $X_j$ .  
A *directed path* is a path whose links are directed and pointing from preceding towards following nodes in the sequence.

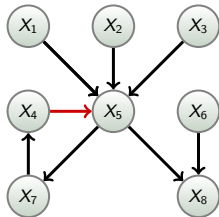
Undirected Graph



### For Directed Graphs

- **Parents and Children:**  $X_i$  is a parent of  $X_j$  if there is a link from  $X_i$  to  $X_j$ .  $X_i$  is a child of  $X_j$  if there is a link from  $X_j$  to  $X_i$ .
- **Ancestors and Descendants:** The ancestors of a node  $X_i$  are the nodes with a directed path ending at  $X_i$ . The descendants of  $X_i$  are the nodes with a directed path beginning at  $X_i$ .
- **Directed Acyclic Graph:** Graph in which by following the direction of the arrows a node will never be visited more than once.

Directed Cyclic Graph

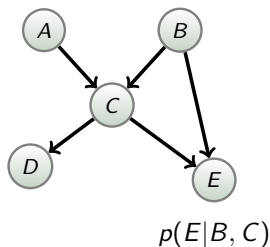


## Bayesian Networks (Belief Networks)

A Bayesian network is a directed acyclic graph in which each node has associated the conditional probability of the node given its parents.

The joint distribution is obtained by taking the product of the conditional probabilities.

$$p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|C)p(E|B, C)$$



## Example – Part I

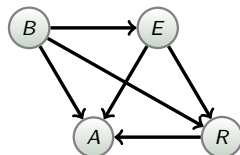
Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

Without loss of generality, we can write

$$\begin{aligned} p(A, R, E, B) &= p(A|R, E, B)p(R, E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E|B)p(B) \end{aligned}$$

However

- The alarm is not directly influenced by any report on the radio, that is  $p(A|R, E, B) = p(A|E, B)$
- $p(R|E, B) = p(R|E)$
- $p(E|B) = p(E)$



Therefore

$$p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$$



## Example – Part I

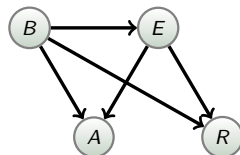
Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

Without loss of generality, we can write

$$\begin{aligned} p(A, R, E, B) &= p(A|R, E, B)p(R, E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E|B)p(B) \end{aligned}$$

However

- The alarm is not directly influenced by any report on the radio, that is  $p(A|R, E, B) = p(A|E, B)$
- $p(R|E, B) = p(R|E)$
- $p(E|B) = p(E)$



Therefore

$$p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$$

## Example – Part I

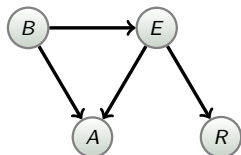
Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

Without loss of generality, we can write

$$\begin{aligned} p(A, R, E, B) &= p(A|R, E, B)p(R, E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E|B)p(B) \end{aligned}$$

However

- The alarm is not directly influenced by any report on the radio, that is  $p(A|R, E, B) = p(A|E, B)$
- $p(R|E, B) = p(R|E)$
- $p(E|B) = p(E)$



Therefore

$$p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$$

## Example – Part I

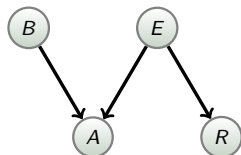
Sally's burglar **A**larm is sounding. Has she been **B**urgled, or was the alarm triggered by an **E**arthquake? She turns the car **R**adio on for news of earthquakes.

Without loss of generality, we can write

$$\begin{aligned} p(A, R, E, B) &= p(A|R, E, B)p(R, E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E, B) \\ &= p(A|R, E, B)p(R|E, B)p(E|B)p(B) \end{aligned}$$

However

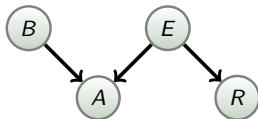
- The alarm is not directly influenced by any report on the radio, that is  $p(A|R, E, B) = p(A|E, B)$
- $p(R|E, B) = p(R|E)$
- $p(E|B) = p(E)$



Therefore

$$p(A, R, E, B) = p(A|E, B)p(R|E)p(E)p(B)$$

## Example – Part II: Specifying the Tables



$$p(A|B, E)$$

Alarm = 1	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$$p(R|E)$$

Radio = 1	Earthquake
1	1
0	0

The remaining tables are  $p(B = 1) = 0.01$  and  $p(E = 1) = 0.000001$ . The tables and graphical structure fully specify the distribution.

## Example Part III: Inference

**Initial Evidence: The alarm is sounding**

$$\begin{aligned} p(B = 1|A = 1) &= \frac{\sum_{E,R} p(B = 1, E, A = 1, R)}{\sum_{B,E,R} p(B, E, A = 1, R)} \\ &= \frac{\sum_{E,R} p(A = 1|B = 1, E)p(B = 1)p(E)p(R|E)}{\sum_{B,E,R} p(A = 1|B, E)p(B)p(E)p(R|E)} \approx 0.99 \end{aligned}$$

**Additional Evidence: The radio broadcasts an earthquake warning:**

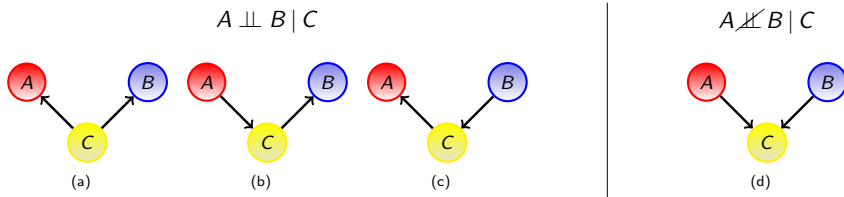
A similar calculation gives  $p(B = 1|A = 1, R = 1) \approx 0.01$ .

Initially, because the alarm sounds, Sally thinks that she's been burgled. However, this probability drops dramatically when she hears that there has been an earthquake.

The earthquake 'explains away' to an extent the fact that the alarm is ringing.

## Independence $\perp\!\!\!\perp$ in Bayesian Networks – Part I

All Bayesian networks with three nodes and two links:



- In (a), (b) and (c),  $A, B$  are conditionally independent given  $C$ .

$$(a) \quad p(A, B|C) = \frac{p(A, B, C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{p(C)} = p(A|C)p(B|C)$$

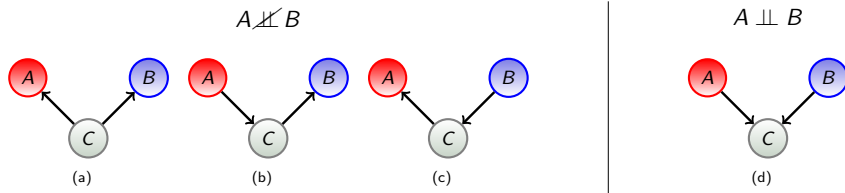
$$(b) \quad p(A, B|C) = \frac{p(A)p(C|A)p(B|C)}{p(C)} = \frac{p(A, C)p(B|C)}{p(C)} = p(A|C)p(B|C)$$

$$(c) \quad p(A, B|C) = \frac{p(A|C)p(C|B)p(B)}{p(C)} = \frac{p(A|C)p(B, C)}{p(C)} = p(A|C)p(B|C)$$

- In (d) the variables  $A, B$  are conditionally dependent given  $C$ .

$$p(A, B|C) = \frac{p(A, B, C)}{p(C)} = \frac{p(C|A, B)p(A)p(B)}{p(C)} \quad \text{in general} \quad \neq p(A|C)p(B|C).$$

## Independence $\perp\!\!\!\perp$ in Bayesian Networks – Part II

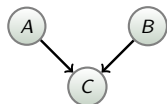


- In (a), (b) and (c), the variables  $A, B$  are marginally dependent.
- In (d) the variables  $A, B$  are marginally independent.

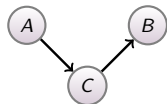
$$p(A, B) = \sum_C p(A, B, C) = \sum_C p(A)p(B)p(C|A, B) = p(A)p(B)$$

# Collider

Summary of two previous slides:



If  $C$  has more than one incoming link, then  $A \perp\!\!\!\perp B$  and  $A \not\perp\!\!\!\perp B \mid C$ . In this case  $C$  is called **collider**.

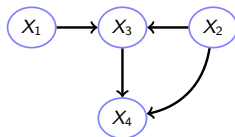


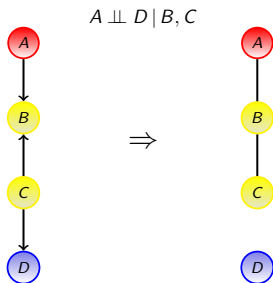
If  $C$  has at most one incoming link, then  $A \perp\!\!\!\perp B \mid C$  and  $A \not\perp\!\!\!\perp B$ . In this case  $C$  is called **non-collider**.



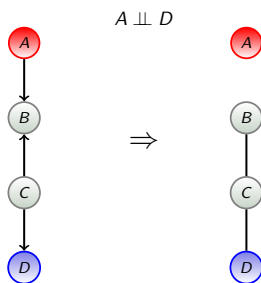
## Collider

A node can be a collider on a path and a non-collider on another path. In the figure,  $X_3$  is a collider on the path  $X_1 \rightarrow X_3 \leftarrow X_2$  and a non-collider on the path  $X_2 \rightarrow X_3 \rightarrow X_4$ .

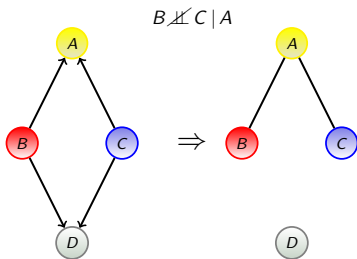
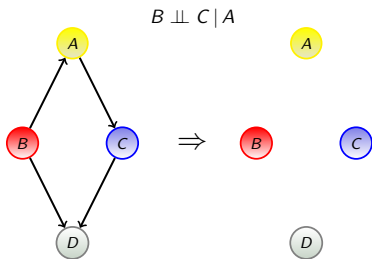




non-collider in the conditioning set blocks a path



collider outside the conditioning set blocks a path




all paths need to be blocked to obtain  $\perp\!\!\!\perp$

## General Rule for Independence in Belief Networks

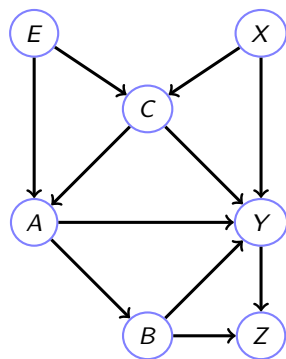
Given three sets of nodes  $\mathcal{X}, \mathcal{Y}, \mathcal{C}$ , if all paths from any element of  $\mathcal{X}$  to any element of  $\mathcal{Y}$  are blocked (closed) by  $\mathcal{C}$ , then  $\mathcal{X}$  and  $\mathcal{Y}$  are conditionally independent given  $\mathcal{C}$ .

We say that  $\mathcal{C}$  d-separates  $\mathcal{X}$  and  $\mathcal{Y}$ .

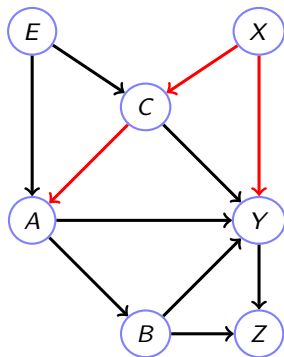
A path  $\mathcal{P}$  is blocked by  $\mathcal{C}$  if one of the following conditions is satisfied:

1. there is a collider  in the path  $\mathcal{P}$  such that neither the collider nor any of its descendants is in the conditioning set  $\mathcal{C}$ .
2. there is a non-collider in the path  $\mathcal{P}$  that is in the conditioning set  $\mathcal{C}$ .

## General Rule for Independence in Bayesian Networks

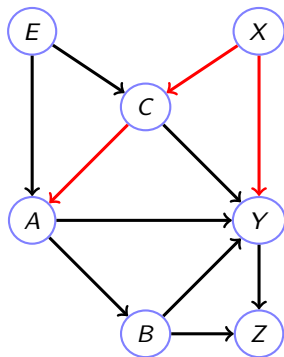


## General Rule for Independence in Bayesian Networks



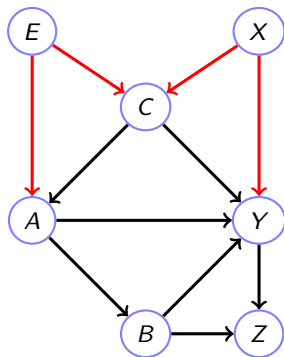
Is the path  $A \leftarrow C \leftarrow X \rightarrow Y$  closed?

## General Rule for Independence in Bayesian Networks



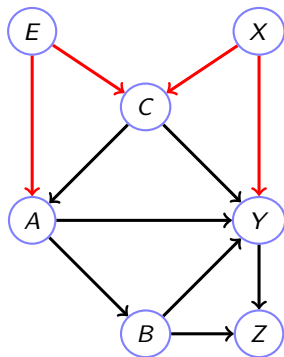
Is the path  $A \leftarrow C \leftarrow X \rightarrow Y$  closed? No

## General Rule for Independence in Bayesian Networks



Is the path  $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$  closed?

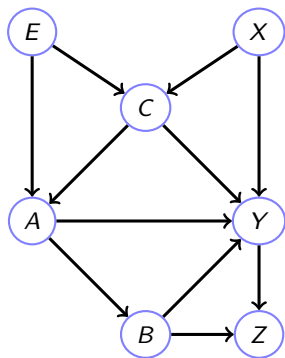
## General Rule for Independence in Bayesian Networks



Is the path  $A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$  closed? Yes

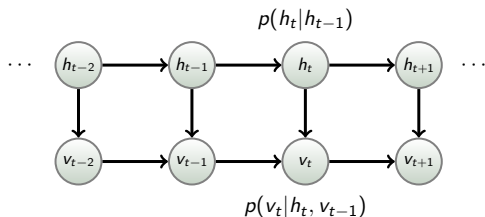


## General Rule for Independence in Bayesian Networks



What happens if we condition on  $C$ ?

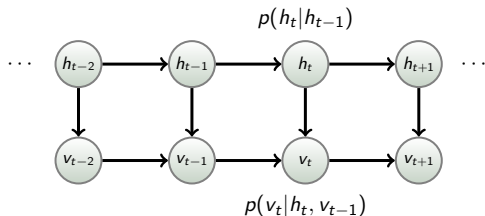
## Example of using the Independence Rule for Time-Series Modeling



- Variables  $v_1, \dots, v_T \equiv v_{1:T}$  represent the observed time-series.
- Discrete hidden variables  $h_{1:T}$  generate the observations.

$$\begin{aligned} p(h_t, v_{1:t}) &= p(v_t | h_t, v_{1:t-1}) p(h_t, v_{1:t-1}) \\ &= p(v_t | h_t, v_{1:t-1}) \sum_{h_{t-1}} p(h_{t-1:t}, v_{1:t-1}) \\ &= p(v_t | h_t, v_{t-1}) \sum_{h_{t-1}} p(h_t | h_{t-1}, v_{1:t-1}) p(h_{t-1}, v_{1:t-1}) \end{aligned}$$

## Example of using the Independence Rule for Time-Series Modeling

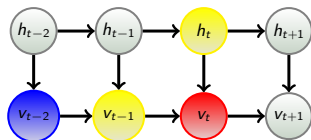


- Variables  $v_1, \dots, v_T \equiv v_{1:T}$  represent the observed time-series.
- Discrete hidden variables  $h_{1:T}$  generate the observations.

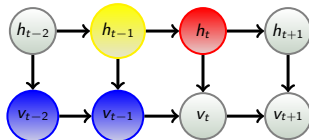
$$\begin{aligned}
 p(h_t, v_{1:t}) &= p(v_t | h_t, \cancel{v_{1:t-2}}, v_{t-1}) p(h_t, v_{1:t-1}) \\
 &= p(v_t | h_t, v_{1:t-1}) \sum_{h_{t-1}} p(h_{t-1:t}, v_{1:t-1}) \\
 &= p(v_t | h_t, v_{t-1}) \sum_{h_{t-1}} p(h_t | h_{t-1}, \cancel{v_{1:t-1}}) p(h_{t-1}, v_{1:t-1})
 \end{aligned}$$

## Example of using the Independence Rule for Time-Series Modeling

$$v_t \perp\!\!\!\perp v_{1:t-2} \mid \{h_t, v_{t-1}\}$$



$$h_t \perp\!\!\!\perp v_{1:t-1} \mid h_{t-1}$$



$$\begin{aligned}
 p(h_t, v_{1:t}) &= p(v_t | h_t, \cancel{v_{1:t-2}}, v_{t-1}) p(h_t, v_{1:t-1}) \\
 &= p(v_t | h_t, v_{1:t-1}) \sum_{h_{t-1}} p(h_{t-1:t}, v_{1:t-1}) \\
 &= p(v_t | h_t, v_{t-1}) \sum_{h_{t-1}} p(h_t | h_{t-1}, \cancel{v_{1:t-1}}) p(h_{t-1}, v_{1:t-1})
 \end{aligned}$$

## Reading

1. Bayesian Reasoning and Machine Learning. D. Barber, 2010 (examples and demos in this talk).
2. Pattern Recognition and Machine Learning. C. M. Bishop, 2009.
3. Probabilistic Networks and Expert Systems. R. G. Cowell and A. P. Dawid, S. L. Lauritzen, D. Spiegelhalter, 2000.
4. Probabilistic graphical Models: Principles and Techniques. D. Koller and N. Friedman, 2009.
5. Bayesian Networks and Decision Graphs. F. V. Jensen, 2001.
6. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. J. Pearl, 1988.
7. Graphical Models. S. Lauritzen, 1996.

## Effects of Causes

Causal inference is used to answer causal (or interventional) questions such as:

Does taking a certain drug induce recovery from a certain disease?

In order to do that we cannot use probabilistic inference in the standard way (conditioning).

# Simpson's Paradox

Consider a medical trial in which the effect of a drug on recovery is investigated. Two trials are conducted, one with 40 males and one with 40 females.

- According to the female data the drug **decreases** recovery. Similarly, according to the male data the drug **decreases** recovery. The conclusion appears that the drug is harmful for the entire population since it is harmful for both subpopulations.
- According to the combined male and female data the drug **increases** recovery.

Females	R=0	R=1		Recovery Rate
No Drug (D=0)	21	9	30	30%
Drug (D=1)	8	2	10	20%
	29	11	40	

Males	R=0	R=1		Recovery Rate
No Drug (D=0)	3	7	10	70%
Drug (D=1)	12	18	30	60%
	15	25	40	

Combined	R=0	R=1		Recovery Rate
No Drug (D=0)	24	16	40	40%
Drug (D=1)	20	20	40	50%
	44	36	80	

Hence, even though the drug doesn't increase recovery for either males or females, it does overall!

## Simpson's Paradox

Our intuition is that since the drug harmful for both males and females, then it should be harmful overall. This is a **causal** intuition that clashes with standard probabilistic inference.

The 'paradox' occurs since we are asking a causal question: "If we give someone the drug, what happens?" However, we answer the question with the conditional distribution  $p(R|D)$ .



# Bayesian Networks for Causal Inference

We would like to answer causal questions by using Bayesian networks.

To achieve that, we need to represent and respond to external or spontaneous changes.

If we assume that each parent-child relationship in the graph represents a stable and autonomous physical mechanism, it is conceivable to change one such relationship without changing the others. This enables to predict the effect of external interventions with minimum extra information.

## Occam's Razor

Two Bayesian networks are (observationally) equivalent (encode the same independence structure) if and only if they have the same skeletons and the same set of collider structures.

Following standard norms of scientific induction, it is reasonable to rule out any theory for which we find a simpler theory that is equally consistent with the data. A theory that survives this selection process is called **minimal**.

**Definition:** A variable  $X$  is said to have a **causal influence** on a variable  $Y$  if a directed path from  $X$  to  $Y$  exists in every minimal structure consistent with the data.

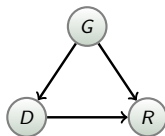
## Simpson's Paradox

The key point is that drug  $D$  is influenced by gender  $G$ .

$$p(G = F|D = 0) = \frac{30}{40} = 0.75, p(G = M|D = 0) = \frac{10}{40} = 0.25$$

$$p(G = F|D = 1) = \frac{10}{40} = 0.25, p(G = M|D = 1) = \frac{30}{40} = 0.75$$

We are giving more often drug to males, who tend to recover more often.



$$p(R = 1|D = 0) = \sum_G p(R = 1|D = 0, G)p(G|D = 0)$$

$$= p(R = 1|D = 0, G = F)p(G = F|D = 0)$$

$$+ p(R = 1|D = 0, G = M)p(G = M|D = 0)$$

$$= \frac{9}{30} * 0.75 + \frac{7}{10} * 0.25$$

$$= 0.3 * 0.75 + 0.7 * 0.25 = 0.4$$

$$p(R = 1|D = 1) = \sum_G p(R = 1|D = 1, G)p(G|D = 1)$$

$$= p(R = 1|D = 1, G = F)p(G = F|D = 1)$$

$$+ p(R = 1|D = 1, G = M)p(G = M|D = 1)$$

$$= 0.2 * 0.25 + 0.6 * 0.75 = 0.5$$

Females	R=0	R=1		Recovery Rate
No Drug (D=0)	21	9	30	30%
Drug (D=1)	8	2	10	20%
	29	11	40	

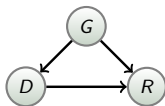
Males	R=0	R=1		Recovery Rate
No Drug (D=0)	3	7	10	70%
Drug (D=1)	12	18	30	60%
	15	25	40	

Combined	R=0	R=1		Recovery Rate
No Drug (D=0)	24	16	40	40%
Drug (D=1)	20	20	40	50%
	44	36	80	

**Hypothetical Intervention  $do(D = 0)$ :** Set the variable  $D$  to value 0.

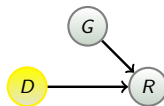
Transform the original DAG  $\mathcal{G}$  with distribution  $p$  into a new DAG  $\mathcal{G}_{\rightarrow D}$  with distribution  $p_{\rightarrow D=0}$  such that:

- Links from  $pa_D$  to  $D$  are removed.
- $p_{\rightarrow D=0}(D = 0) = 1, p_{\rightarrow D=0}(D \neq 0) = 0$ .
- All other conditional probabilities are the same as in  $\mathcal{G}$ .



$$p(D, R, G) = p(R|D, G)p(D|G)p(G)$$

$do(D = 0) \Rightarrow$



$$\begin{aligned}
 p_{\rightarrow D=0}(D = 0, R, G) &= p_{do(D=0)}(R|D = 0, G)p_{\rightarrow D=0}(D = 0)p_{\rightarrow D=0}(G) \\
 &= p(R|D = 0, G)p(G) = p_{\rightarrow D=0}(R, G|D = 0) \\
 p_{\rightarrow D=0}(D \neq 0, R, G) &= 0
 \end{aligned}$$

**Conditional dist.:**  $p(R = 1|D = 0) = \sum_G p(R = 1|D = 0, G)p(G|D = 0) = 0.7 * 0.25 + 0.3 * 0.75 = 0.4$

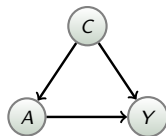
**Causal effect:**  $p_{\rightarrow D=0}(R|D = 0) = \sum_G p(R|D = 0, G)p(G) = 0.7 * 0.5 + 0.3 * 0.5 = 0.5$

## Causal effect

The causal effect of  $A$  on  $Y$  can be seen as the information traveling from  $A$  to  $Y$  through causal paths, or as the conditional distribution of  $Y$  given  $A$  restricted to causal paths. This implies that, to compute the causal effect, we need to disregard the information that travels along non-causal paths, which occurs if such paths are open.

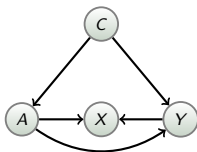
$A$  is a cause of  $Y$

- $p(Y|A)$  measures information traveling along both open paths  $A \leftarrow C \rightarrow Y$  and  $A \rightarrow Y$
- The causal effect is the information traveling only along the causal path  $A \rightarrow Y$

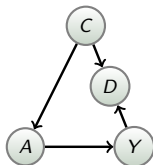


## Causal effect

Since paths with an arrow emerging from  $A$  are either causal or closed by a collider, the problematic paths are only those with an arrow pointing into  $A$ , called back-door paths, which are open if they do not contain a collider.



Non-causal path  $A \rightarrow X \leftarrow Y$  closed by the collider  $X$ .  
Open back-door path  $A \leftarrow C \rightarrow Y$ .



Closed back-door path  $A \leftarrow C \rightarrow D \leftarrow Y$ .

## References

1. Judea Pearl. Causality, Models, Reasoning and Inference, 2000.
2. Alexander Philip Dawid. Fundamentals of Statistical Causality, 2007.
3. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. Causal Inference in Statistics - A Primer 2016.
4. Silvia Chiappa and William Isaac, A Causal Bayesian Networks Viewpoint on Fairness, 2019

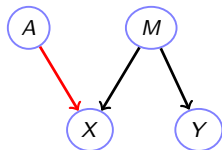
# Bayesian Networks & Fairness through Unawareness

**Fairness through Unawareness:**  $\hat{Y}$  is fair as long as it does not make explicit use of the sensitive attribute  $A$ .

Whilst this fairness criterion is often indicated as problematic because some of the variables used to form  $\hat{Y}$  could be a proxy for  $A$  (such as neighborhood for race), Bayesian networks reveal a more subtle potential issue with it.



## Fairness through Unawareness



Music degree scenario:

$A$  = gender

$M$  = music aptitude

$X$  = score obtained from an ability test taken at the beginning of the degree

$Y$  = score obtained from an ability test taken at the end of the degree.

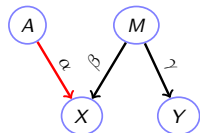
- Individuals with higher music aptitude  $M$  are more likely to obtain higher initial and final scores ( $M \rightarrow X$ ,  $M \rightarrow Y$ ).
- Due to discrimination occurring at the initial testing, women are assigned a lower initial score than men for the same aptitude level ( $A \rightarrow X$ ).

## Fairness through Unawareness

The only path from  $A$  to  $Y$ ,  $A \rightarrow X \leftarrow M \rightarrow Y$ , is closed as  $X$  is a collider on this path.

Therefore the unfair influence of  $A$  on  $X$  does not reach  $Y$  ( $Y \perp\!\!\!\perp A$ ).

Nevertheless, as  $Y \not\perp\!\!\!\perp A | X$ , a prediction  $\hat{Y}$  based on the initial score  $X$  only would contain the unfair influence of  $A$  on  $X$ .



Linear model:  $Y = \gamma M$ ,  $X = \alpha A + \beta M$ , with  $\mathbb{E}[A^2] = 1$  and  $\mathbb{E}[M^2] = 1$ .

A linear predictor of the form  $\hat{Y} = \theta_X X$  minimizing  $\left\langle (Y - \hat{Y})^2 \right\rangle_{p(A)p(M)}$  would have parameters

$$\theta_X = \frac{\gamma\beta}{\alpha^2 + \beta^2},$$

giving  $\hat{Y} = \gamma\beta(\alpha A + \beta M)/(\alpha^2 + \beta^2)$ , i.e.  $\hat{Y} \not\perp\!\!\!\perp A$ .

Therefore, this predictor would be using the sensitive attribute to form a decision, although implicitly rather than explicitly.

## Fairness through Unawareness

Parameters minimizing  $\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - \theta_X X)^2]$ :

$$\begin{aligned}\frac{\partial \mathbb{E}[(Y - \theta_X X)^2]}{\partial \theta_X} &= \mathbb{E}\left[\frac{\partial (Y - \theta_X X)^2}{\partial \theta_X}\right] \\&= -2\mathbb{E}[(Y - \theta_X X)X] \\&= -2\mathbb{E}[YX] + 2\theta_X \mathbb{E}[X^2] = 0 \\&\iff \theta_X = \frac{\mathbb{E}[YX]}{\mathbb{E}[X^2]}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[YX] &= \mathbb{E}[\gamma M(\alpha A + \beta M)] \\&= \gamma \alpha \mathbb{E}[MA] + \gamma \beta \mathbb{E}[MM] \\&= \gamma \alpha \mathbb{E}[M]\mathbb{E}[A] + \gamma \beta \mathbb{E}[MM] \\&= \gamma \beta\end{aligned}$$

$$\Rightarrow \theta_X = \frac{\gamma \beta}{\alpha^2 + \beta^2}$$

$$\begin{aligned}\mathbb{E}[XX] &= \mathbb{E}[(\alpha A + \beta M)(\alpha A + \beta M)] \\&= \alpha^2 \mathbb{E}[AA] + 2\alpha \beta \mathbb{E}[AM] + \beta^2 \mathbb{E}[MM] \\&= \alpha^2 + \beta^2\end{aligned}$$

## Fairness through Unawareness

A predictor explicitly using the sensitive attribute,  $\hat{Y} = \theta_X X + \theta_A A$ , minimizing  $\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - \theta_X X - \theta_A A)^2] = \mathbb{E}\left[\left(Y - \underbrace{(\theta_X, \theta_A)}_{\theta} \begin{pmatrix} X \\ A \end{pmatrix}\right)^2\right]$ , would have parameters

$$\theta = (\gamma/\beta, -\alpha\gamma/\beta)$$

giving

$$\hat{Y} = \theta_X X + \theta_A A = \frac{\gamma}{\beta}(\alpha A + \beta M) - \frac{\alpha\gamma}{\beta} A = \cancel{\frac{\alpha\gamma}{\beta} A} + \frac{\gamma}{\beta} \beta M - \cancel{\frac{\alpha\gamma}{\beta} A} = \gamma M,$$

i.e.  $\hat{Y} \perp\!\!\!\perp A$ . Therefore, this predictor would be fair.

## Fairness through Unawareness

$$\begin{aligned}\frac{\partial \mathbb{E} \left[ \left( Y - \theta \begin{pmatrix} X \\ A \end{pmatrix} \right)^2 \right]}{\partial \theta} &= -2 \mathbb{E} \left[ \left( Y - \theta \begin{pmatrix} X \\ A \end{pmatrix} \right) \begin{pmatrix} X \\ A \end{pmatrix}^\top \right] \\&= -2 \mathbb{E} \left[ Y \begin{pmatrix} X \\ A \end{pmatrix}^\top \right] + 2 \theta \mathbb{E} \left[ \begin{pmatrix} X \\ A \end{pmatrix} \begin{pmatrix} X \\ A \end{pmatrix}^\top \right] \\&= -2 \mathbb{E}[(YX, YA)] + 2 \theta \mathbb{E} \left[ \begin{pmatrix} X^2 & AX \\ AX & A^2 \end{pmatrix} \right] = 0 \\&\iff \theta = (\mathbb{E}[YX], \mathbb{E}[YA]) \begin{pmatrix} \mathbb{E}[X^2] & \mathbb{E}[AX] \\ \mathbb{E}[AX] & \mathbb{E}[A^2] \end{pmatrix}^{-1}\end{aligned}$$

$$\mathbb{E}[YA] = \gamma \mathbb{E}[MA] = 0$$

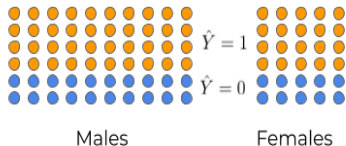
$$\mathbb{E}[AX] = \mathbb{E}[A(\alpha A + \beta M)] = \alpha \mathbb{E}[AA] + \beta \mathbb{E}[AM] = \alpha$$

$$\begin{aligned}\theta &= (\gamma\beta, 0) \begin{pmatrix} \alpha^2 + \beta^2 & \alpha \\ \alpha & 1 \end{pmatrix}^{-1} = (\gamma\beta, 0) \begin{pmatrix} 1/\beta^2 & -\alpha/\beta^2 \\ -\alpha\beta^2 & (\alpha^2 + \beta^2)/\beta^2 \end{pmatrix} \\&= (\gamma/\beta, -\alpha\gamma/\beta)\end{aligned}$$

# Common group-fairness definitions (binary classification setting)

## Demographic Parity

The percentage of individuals assigned to class 1 should be the same for groups  $A=0$  and  $A=1$ .



Dataset

- $a^n \in \{0, 1\}$  sensitive attribute
- $y^n \in \{0, 1\}$  class label
- $\hat{y}^n \in \{0, 1\}$  prediction of the class
- $\mathbf{x}^n \in \mathbb{R}^d$  features

$$p(\hat{Y} = 1 | A = 0) = p(\hat{Y} = 1 | A = 1)$$

$$\hat{Y} \perp A$$

## Common group-fairness definitions

### Equal False Positive/Negative Rates (EFPRs/EFNRs)

$$p(\hat{Y} = 1 | Y = 0, A = 0) = p(\hat{Y} = 1 | Y = 0, A = 1)$$

$$p(\hat{Y} = 0 | Y = 1, A = 0) = p(\hat{Y} = 0 | Y = 1, A = 1)$$

$$\hat{Y} \perp\!\!\!\perp A | Y$$

### Predictive Parity

$$p(Y = 1 | \hat{Y} = 1, A = 0) = p(Y = 1 | \hat{Y} = 1, A = 1)$$

$$p(Y = 0 | \hat{Y} = 0, A = 0) = p(Y = 0 | \hat{Y} = 0, A = 1)$$

$$Y \perp\!\!\!\perp A | \hat{Y}$$

## COMPAS predictive risk instrument

# Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

May 23, 2016



## COMPAS predictive risk instrument

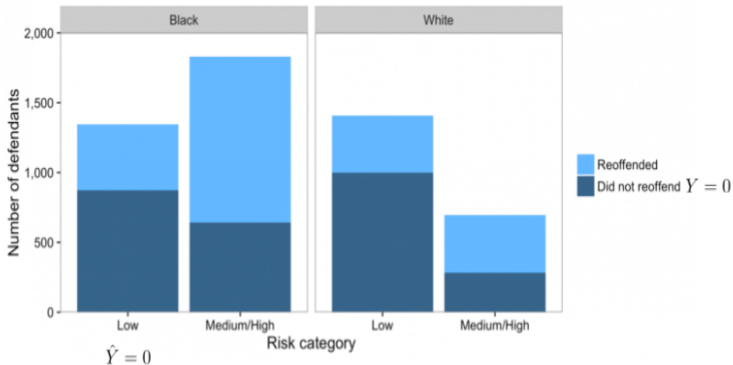
A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel October 17, 2016

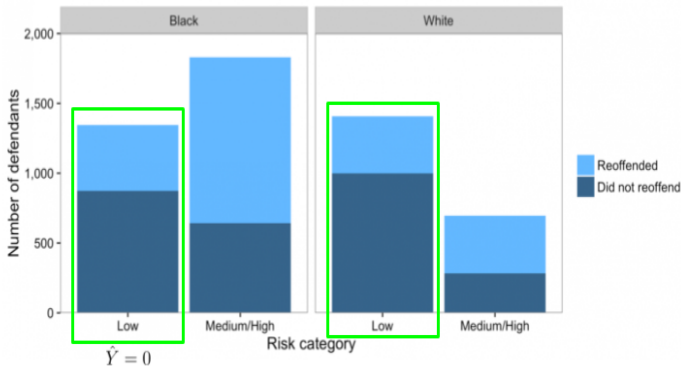
The Washington Post

*Democracy Dies in Darkness*

## COMPAS predictive risk instrument

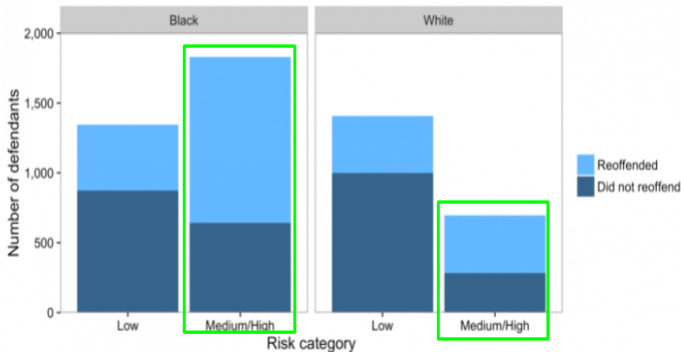


## COMPAS predictive risk instrument



Low risk  
~70% did not reoffend  
for both the black and  
white groups.

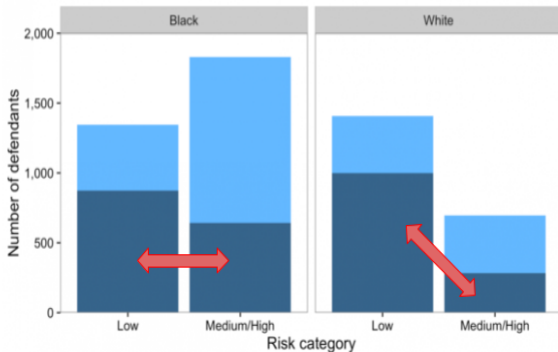
## COMPAS predictive risk instrument



Medium-high risk  
The same percentage of  
individuals did not  
reoffend in both groups.

$$Y \perp A | \hat{Y}$$

# COMPAS predictive risk instrument

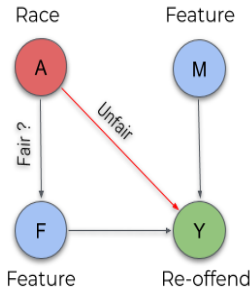


Black defendants who did not reoffend were more often labeled "high risk"

Did not reoffend  
False Positive Rates  
differ

$$\hat{Y} \not\propto A | Y$$

## Patterns of unfairness in the data not considered



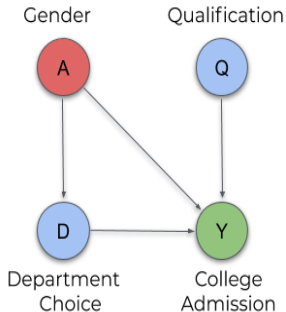
Modern policing tactics center around targeting a small number of neighborhoods --- often disproportionately populated by non-whites.

We can rephrase this as indicating the presence of a direct path  $A \rightarrow Y$  (through unobserved neighborhood).

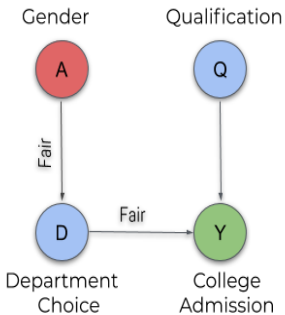
Such tactics also imply an influence of  $A$  on  $Y$  through  $F$  containing number of prior arrests.

EFPRs/EFNRs and Predictive Parity require the rate of (dis)agreement between the correct and predicted label (e.g. incorrect-classification rates) to be the same for black and white defendants, and are therefore not concerned with dependence of  $Y$  on  $A$ .

## Patterns of unfairness: college admission example



## Three main scenarios

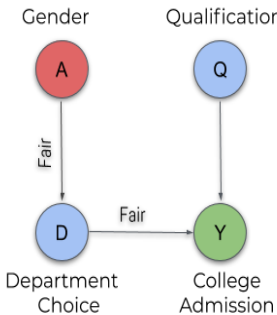


Influence of A on Y is all fair

Predictive Parity  
Equal FPRs/FNRs

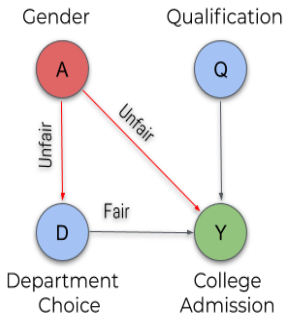


## Three main scenarios



Influence of A on Y is all fair

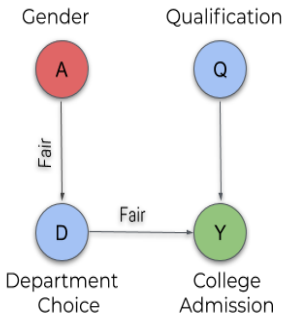
Predictive Parity  
Equal FPRs/FNRs



Influence of A on Y is all unfair

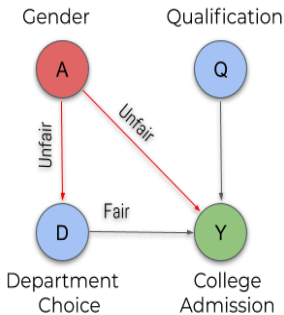
Demographic Parity

## Three main scenarios



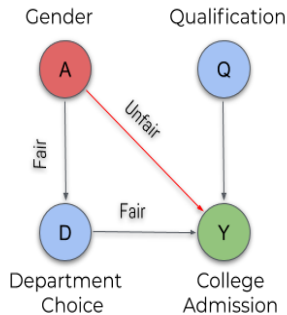
Influence of A on Y is all fair

Predictive Parity  
Equal FPRs/FNRs



Influence of A on Y is all unfair

Demographic Parity



Influence of A on Y is both fair  
and unfair

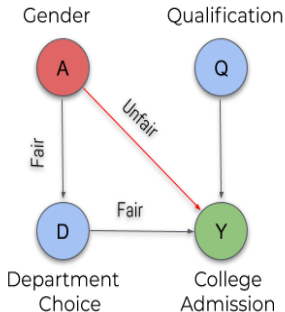
?

## Path-specific fairness

$A=a$  and  $A=\bar{a}$  indicate female and male applicants respectively

$Y_{\bar{a}}(D_a)$  Random variable with distribution equal to the conditional distribution of  $Y$  given  $A$  restricted to causal paths, with  $A=\bar{a}$  along  $A \rightarrow Y$  and  $A=a$  along  $A \rightarrow D \rightarrow Y$ .

$\hat{Y}_{\bar{a}}(D_a)$  **Path-specific Fairness**  
 $p(\hat{Y}_{\bar{a}}(D_a) = 1) = p(\hat{Y}_a = 1)$



# Accounting for full shape of distribution

Wasserstein fair classification.

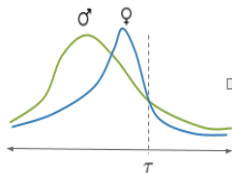
R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa (2019)

Binary classifier outputs a continuous value that represents the probability that individual  $n$  belong to class 1,  $s^n = p(Y = 1|A = a^n, X = x^n)$ . A decision is taken by thresholding  $\hat{y}^n = \mathbb{1}_{s^n > \tau}$

General expression including regression  $s^n = \mathbb{E}_{p(Y|A=a^n, X=x^n)}[Y]$   $\hat{y}^n = s^n$  regression  
 $\hat{y}^n = \mathbb{1}_{s^n > \tau}$  classification

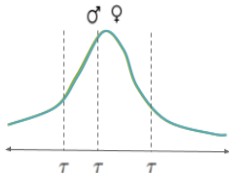
Demographic Parity

$$\mathbb{E}_{p(\hat{Y}|A=\bar{a})}[\hat{Y}] = \mathbb{E}_{p(\hat{Y}|A=a)}[\hat{Y}]$$



Strong Demographic Parity

$$p(S|\bar{a}) = p(S|a)$$



Strong Path-specific Fairness

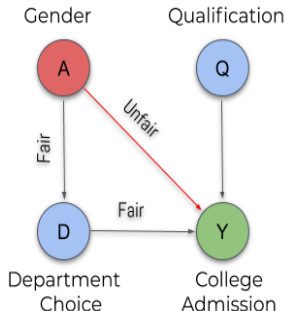
$$p(S_{\bar{a}}(D_a)) = p(S_a)$$

## Individual fairness

Similar individuals should be treated similarly.

Fairness through awareness. C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2011)

A female applicant should get the same decision as a male applicant with the same qualification and applying to the same department.



## Path-specific counterfactual fairness: linear model example

$$A \sim \text{Bern}(\pi), Q = \theta^q + \epsilon_q,$$

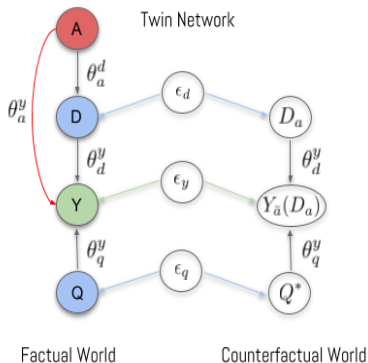
$$D = \theta^d + \theta_a^d A + \epsilon_d,$$

$$Y = \theta^y + \theta_a^y A + \theta_q^y Q + \theta_d^y D + \epsilon_y,$$

$$\mathbb{E}_{p(Y_{\bar{a}}(D_a) | A=a, Q=q^n, D=d^n)}[Y_{\bar{a}}(D_a)]$$

As  $Q$  is non-descendant of  $A$ , and  $D$  is descendant of  $A$  along a fair path, this coincides with

$$\mathbb{E}_{p(Y | A=\bar{a}, Q=q^n, D=d^n)}[Y]$$



In more complex scenarios we would need to use corrected versions of the features.