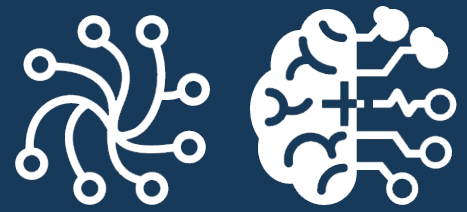
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY



因果图和因果模型

叶南阳

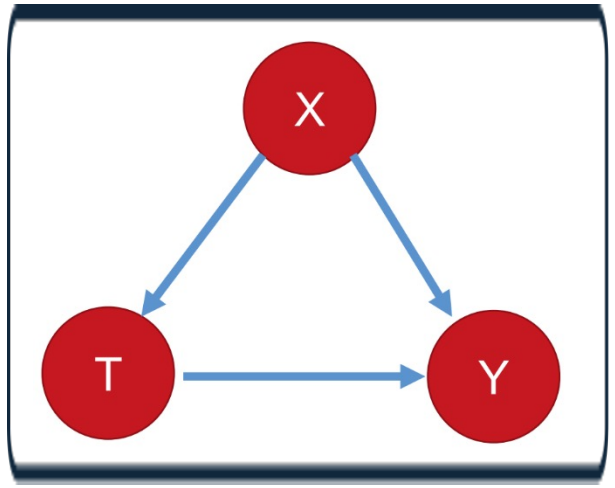
第一节 倾向得分

倾向得分(Propensity score)

- 上节课中，我们提到ATE是计算整体的干预效果

$$ATE = E[Y(1) - Y(0)]$$

- 如果不是随机对照实验，即存在着X同时影响着T和Y



此时，$E[Y|T = 1] \neq E[Y(1)], E[Y|T = 0] \neq E[Y(0)]$

倾向得分(Propensity score)

- 为此，我们引入了倾向得分(Propensity score)
- 定义：个体在一组既定的协变量下，接受某种参与（Treatment）的可能性，即：

$$e(x) \overset{\text{def}}{=} P(T_i = 1 | X_i)$$

- 如果针对多个可观察特征X进行对比匹配非常困难，所以，可将多维协变量X用一个一维变量——倾向得分$P(x)$来代替
- 在随机试验中，我们可以利用propensity score的大小帮助我们找到一种分组使得该分组的效果与treatment的效果独立，不同组之间倾向性评分相近的个体的协变量是基本均衡的
- 在计算出倾向得分后，我们可以为实验组中的每一个个体在对照组中找到一个分数相同或最近的个体进行配对，从而消除协变量的影响，这就是倾向得分匹配(PSM)

逆概率加权(Inverse Probability Weighting)

- IPW是一种消除偏差的方法，可以用于纠正样本分布不均衡导致的辛普森悖论等问题
- IPW中ATE的计算公式：

$$\tau = E[Y(1) - Y(0)] = E\left[\frac{I\{T=1\}Y}{e(X)}\right] - E\left[\frac{I\{T=0\}Y}{1-e(X)}\right]$$

- 可以发现：

$$E\left[\frac{I\{T=1\}Y}{e(X)}\right] = E\left\{\frac{1}{e(X)}E[I\{T=1\}Y(1)|X]\right\} = E\left\{\frac{E(I\{T=1\}|X)}{e(X)}E[Y(1)|X]\right\} = E\{E[Y(1)|X]\}$$

$$= E[Y(1)]$$

- ATE的估计值：

$$\hat{\tau} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\frac{T_i Y_i}{e(X_i)} - \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-e(X_i)}$$

逆概率加权(Inverse Probability Weighting)

- 在实际的使用过程中，IPW可能会不准确或者会出现特别大或者特别小的值，此时$\frac{1}{e(X_i)}$或者$\frac{1}{1-e(X_i)}$就可能过大或者过小，需要采用标准化权重的方法，即对同一个组中的IPW做归一化处理，则会有：

$$\hat{\tau} = \frac{\sum_{i=1}^{N} Y_i T_i w_1(X_i)}{\sum_{i=1}^{N} T_i w_1(X_i)} - \frac{\sum_{i=1}^{N} Y_i (1-T_i) w_0(X_i)}{\sum_{i=1}^{N} (1-T_i) w_0(X_i)}$$

其中：

$$w_1(X_i) = \frac{1}{e(X_i)}$$

$$w_0(X_i) = \frac{1}{1-e(X_i)}$$

逆概率加权(Inverse Probability Weighting)

- 示例：

原始样本分布：

类别	$T=1$	$T=0$	$P(T=1 \mid X)$
$X=1$	8	2	$\frac{8}{10}=0.8$
$X=0$	10	6	$\frac{10}{16}=0.625$

reweighted后：

类别	$T=1$	$T=0$	$P(T=1 \mid X)$
$X=1$	$\frac{8}{0.8}=10$	$\frac{2}{1-0.8}=10$	$\frac{10}{10+10}=0.5$
$X=0$	$\frac{10}{0.625}=16$	$\frac{6}{1-0.625}=16$	$\frac{16}{16+16}=0.5$

逆概率加权(Inverse Probability Weighting)

- 示例：

  原始样本分布 vs reweighted：

类别	$P(T=1 \mid X=1)$	$P(T=1 \mid X=0)$	$P(T=1)$
Original	0.8	0.625	$\frac{8+10}{10+16}=0.69$
reweighted	0.5	0.5	0.5

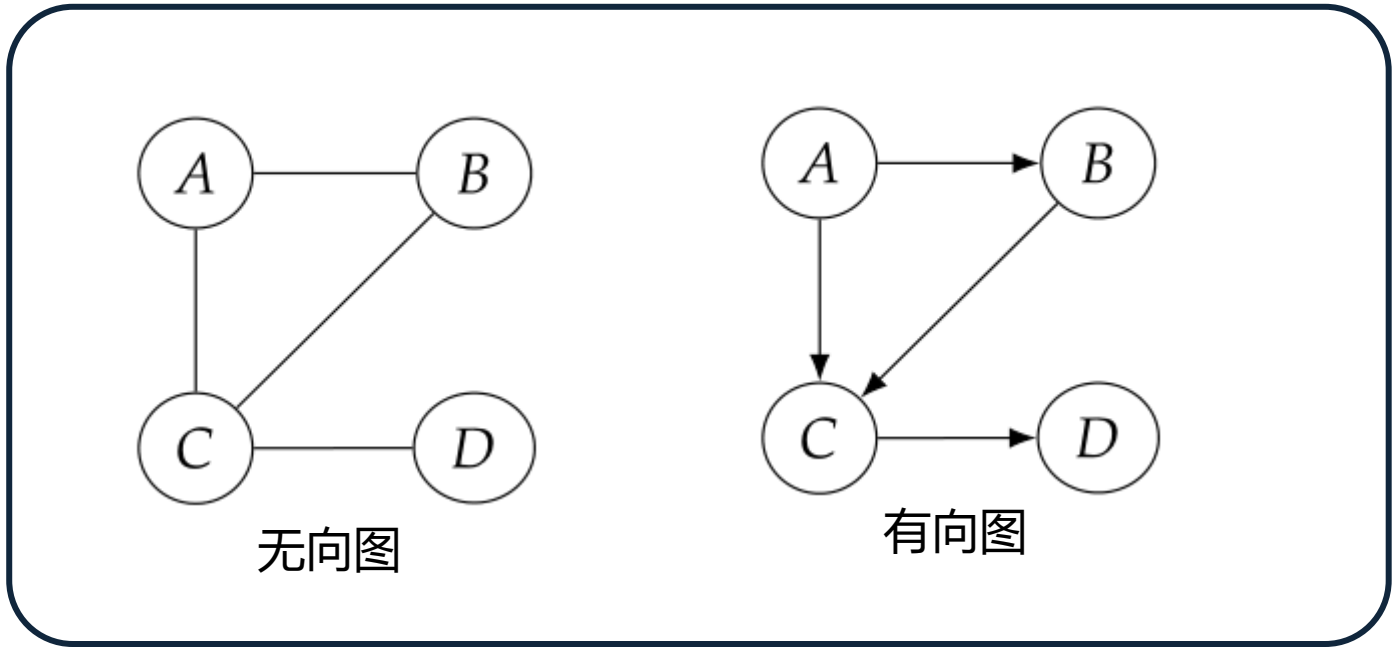
- 可以发现，经过reweighted之后：

$$P(t|x)=P(t)=0.5$$

第二节 图中的关联及因果关系

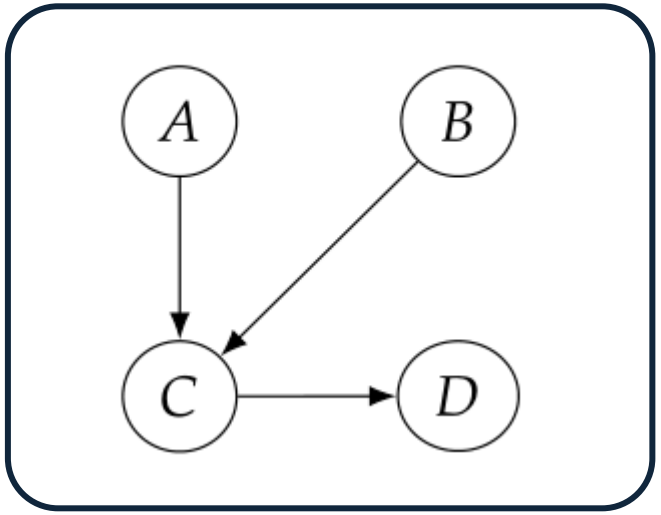
因果图的基本概念

- 图：顶点以及连接顶点的边的集合
- 分类：有向图、无向图
- 相邻节点：由一条边连接的两个节点
- 双亲节点：我们会用pa_i来代表节点X_i的双亲节点



因果图的基本概念

- 路径：相邻节点的任意序列
- 有向路径：指向相同方向的有向边组成的路径
- 环：从某节点出发并指向其自身的有向路径
- 有向无环图(DAG): 图中不包含环的有向图
- 不道德结构(immorality)：若双亲节点X和Y有相同的孩子节点Z，却没有连接X和Y的边，则称$X \rightarrow Z \leftarrow Y$是不道德结构(immorality)



包含不道德结构的有向图

贝叶斯网络

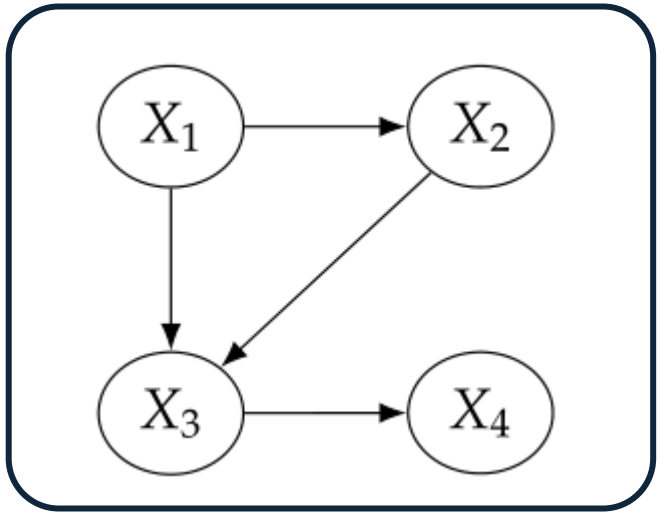
- 一般来说，可以用概率中的链式法则分解任意分布

- 问题：需要指数数量级的参数，随着n的增大，这种参数化方法变得十分困难

- 替代思路：直观来看，针对联合分布中的多个变量，可以通过对局部相关性建模的方式来优化。例如，如果我们有理由相信X_4局部依赖于X_3我们可以用$P(x_4|x_3)$来代替$P(x_4|x_3,x_2,x_1)$

- 事实上，指向X_4的节点只有X_3，这意味着X_4局部依赖于X_3

- 当我们用图G来与概率分布P比较时，G中的节点与P中的随机变量总有一一对应的关系。当我们说G中的节点是独立的，也就是说明P中的随机变量是独立的



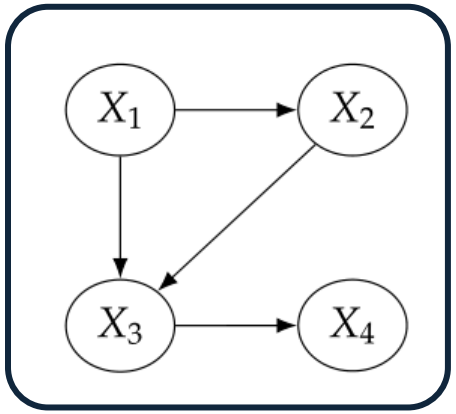
贝叶斯网络

- 给定一个概率分布以及对应的**有向无环图**，可以用局部马尔可夫假设来使独立性规范形式化
- 局部马尔可夫假设：给定有向无环图G中的节点X的双亲节点， X与所有它的非后代节点独立
- 对于如图所示的概率分布P，如果P满足局部马尔可夫假设，我们可以得到：

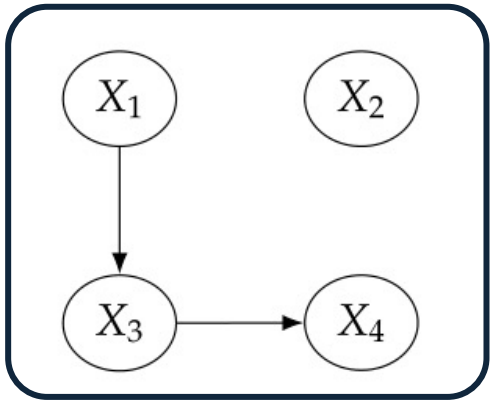
$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)P(x_4|x_3)$$

如果去掉边$X_1 \rightarrow X_2$和边$X_2 \rightarrow X_3$，我们可以进一步简化P的分解：

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_3)$$



X_4局部依赖于X_3



具有更多独立性的DAG图

贝叶斯网络

- 贝叶斯网络分解：

  给定概率分布P和有向无环图G，如果 $P(x_1, \dots, x_n) = \prod_i P(x_i|pa_i)$, 则P可以根据G分解

- 局部马尔可夫假设与贝叶斯网络分解是等价的

- 极小假设（Minimality assumption）：

  给定有向无环图G中的节点X的双亲节点， X与所有它的非后代节点独立(局部马尔可夫假设，Local Markov Assumption)

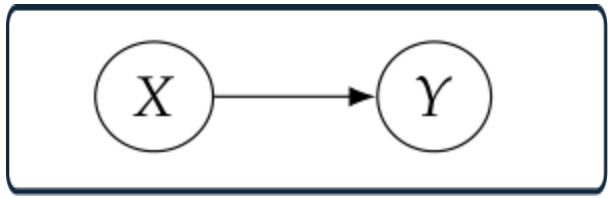
  有向无环图中的相邻节点是相互依赖的

贝叶斯网络

- 极小假设与局部马尔可夫假设的区别：假设有向无环图中只有两个相连的节点X和Y

  局部马尔可夫假设说明$P(x,y)$可分解为$P(x)P(y|x)$，也可以分解为$P(x)P(y)$，也就是说X和Y可以是独立的

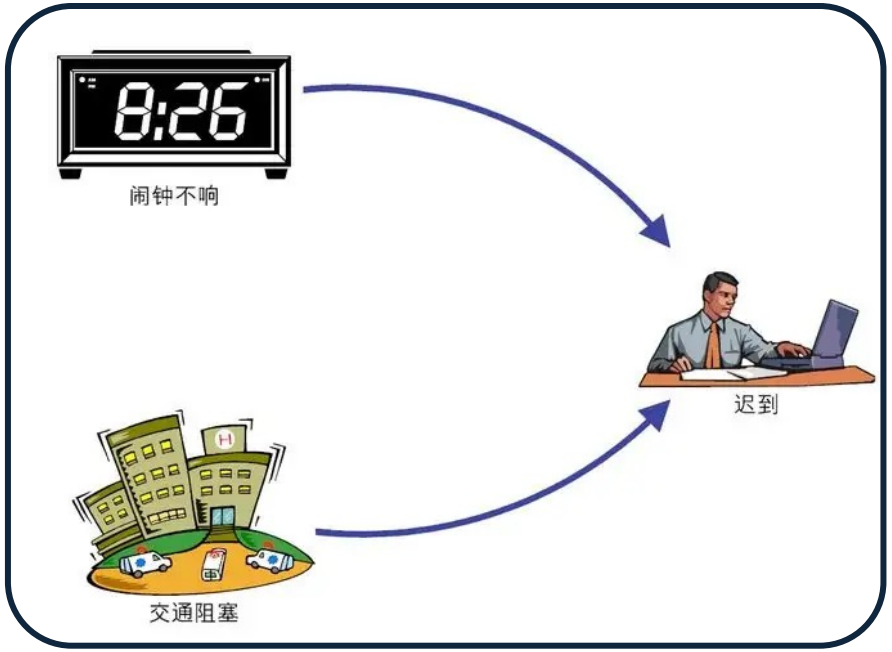
  极小假设说明$P(x,y)$只能分解为$P(x)P(y|x)$，且在与图相关的概率分布P中不存在额外的独立性



- 在贝叶斯网络中，删除某条边等同于增加独立性。而极小假设则等同于我们不能再从图中删除任意条边

因果图和图的基本结构块

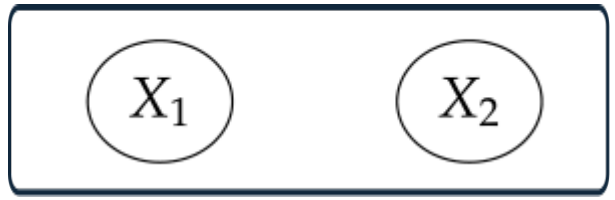
- 什么是因：如果随机变量Y随着随机变量X的变化而变化，那么就可以说X是Y的一个因
- 因果边假设：在有向图中，所有的双亲节点都是它的孩子节点的一个直接原因
- 在添加了因果边假设之后，有向无环图中的有向路径就对应着因果关系
- 图的基本结构块可以分为：链式结构、分叉结构、不道德结构、两个未连接节点以及两个连接节点



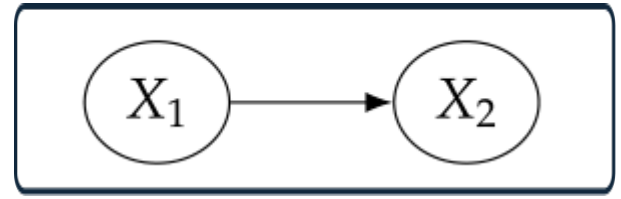
图的基本结构块

- 两个节点的结构：
  - 两个没有连接的节点：
    如图所示，节点X_1与X_2之间没有边相连，因此它们之间没有关联
    证明：对概率图进行贝叶斯网络分解

$$P(x_1, x_2) = P(x_1)P(x_2)$$

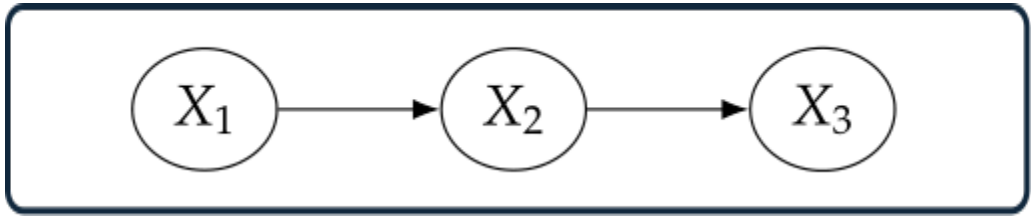


  - 两个相互连接的节点：
    如图所示，节点X_1与X_2之间有一条边相连，因此它们之间有关联
    证明：根据因果边假设， X_1是X_2的因，所以X_2必须随着X_1的改变而改变，因此说X_1，X_2是有关联的



图的基本结构块

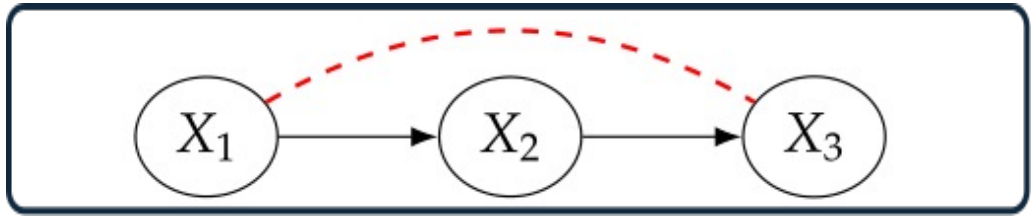
- 三个节点的结构：
  - 链式结构
    - 在如图所示的链式结构中，节点X_1与X_2之间是相互依赖的，节点X_2与X_3之间也是相互依赖的
    - 问题：节点X_1与X_3是否是相互依赖的?
      通常而言，X_1与X_3是相互依赖的(X_2随着X_1的改变而改变，而X_3随着X_2的改变而改变)



链式结构

图的基本结构块

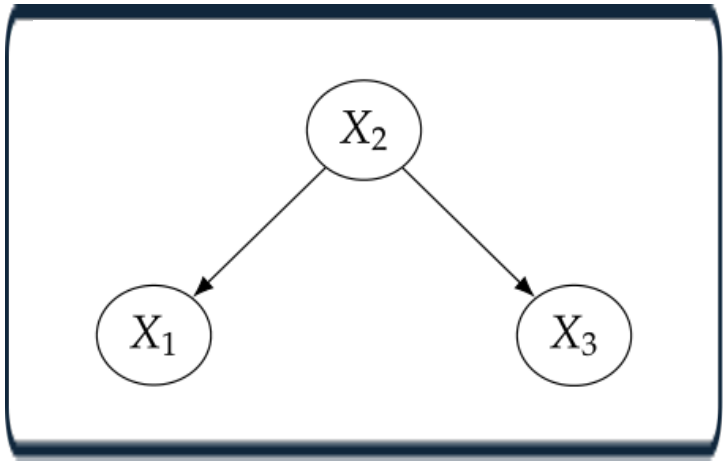
- 三个节点的结构：
  - 链式结构
    - 为了直观地考虑X_1与X_3在链式结构中的关联性，可以可视化关联流
    - 关联流会沿着路径$X_1 \to X_2 \to X_3$从X_1流到X_3
    - 与之对称的，关联流会沿着路径$X_1 \leftarrow X_2 \leftarrow X_3$从$X_3$流到$X_1$



链式结构， 红色虚线弧代表X_1与X_3的关联流

图的基本结构块

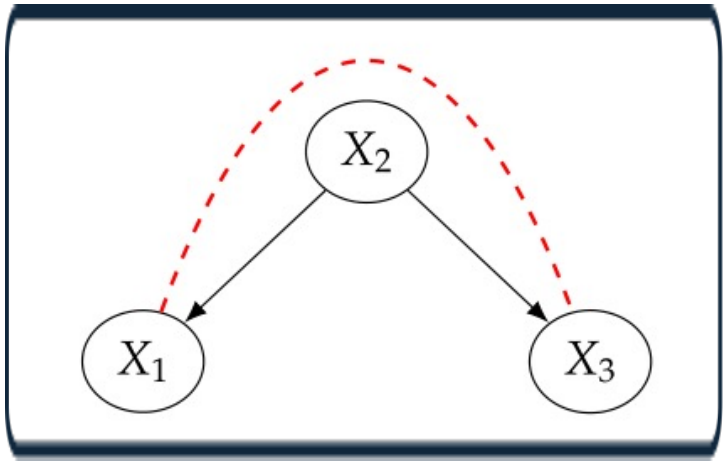
- 三个节点的结构：
  - 分叉结构
    - 在如图所示的分叉结构中，节点X_1与X_2之间是相互依赖的，节点X_2与X_3之间也是相互依赖的
    - 问题：节点X_1与X_3是否是相互依赖的?
      通常而言，X_1与X_3是相互依赖的(X_2的取值同时决定了X_1的取值以及X_3的取值，换言之，X_1与X_3通过他们共同的原因相关联)



分叉结构

图的基本结构块

- 三个节点的结构：
  - 分叉结构
    - 为了直观地考虑X_1与X_3在链式结构中的关联性，可以可视化关联流
    - 关联流会沿着路径$X_1 \leftarrow X_2 \rightarrow X_3$从$X_1$流到$X_3$
    - 同样地，关联流也会沿着路径$X_1 \leftarrow X_2 \rightarrow X_3$从$X_3$流到$X_1$

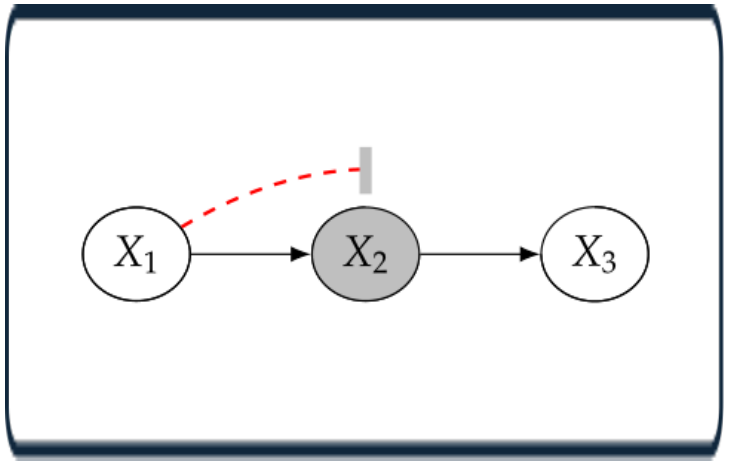


分叉结构，红色虚线弧代表代表X_1与X_3的关联流

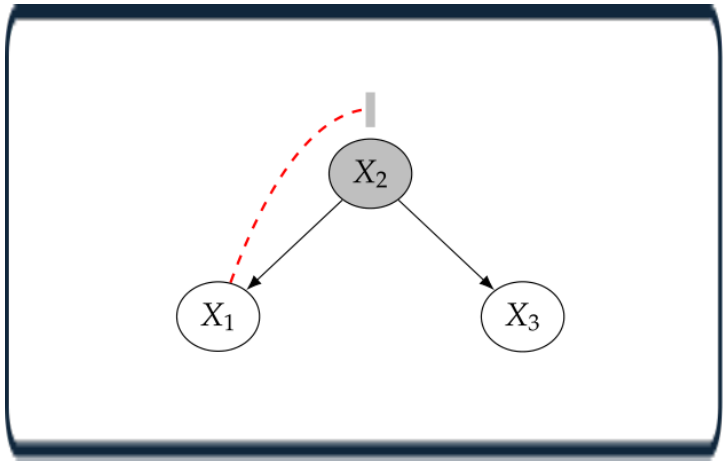
通常而言，关联流都是对称的

图的基本结构块

- 链式结构和分叉结构具有相同的独立性
- 如果我们在链式结构和分叉结构中以节点X_2作为条件，由于每个节点仅仅依赖于它的双亲节点(局部马尔可夫假设)，节点X_1到X_3的关联流就会被阻断，X_3就会独立于X_1，反之亦然
- 如果没有固定条件，链式结构和分叉结构中的关联性是可以自由流动的，这种称作未被阻塞的路径。但若固定了条件，链式结构和分叉结构中的关联性就受到了阻塞，这被称作被阻塞的路径



链式结构中，固定X_2阻断了关联性



分叉结构中，固定X_2阻断了关联性

图的基本结构块

- 通过局部马尔可夫假设证明链式结构中$X_1 \perp\!\!\!\perp X_3 | X_2$

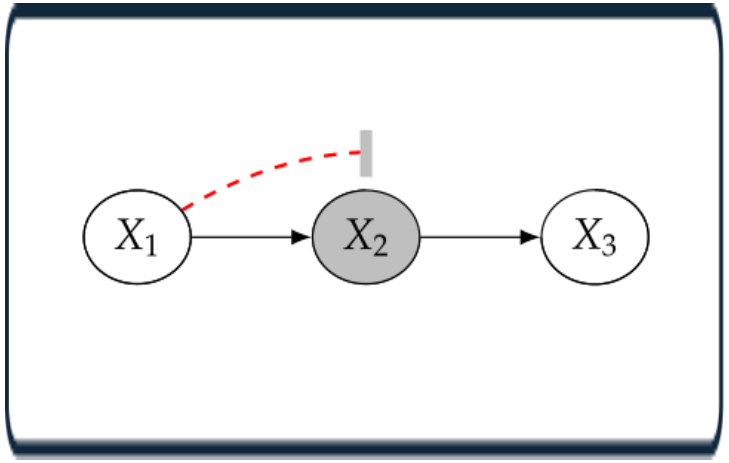
根据贝叶斯网络分解，可以得到 $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_2)$

由贝叶斯定理，可以得到 $P(x_1, x_3|x_2) = \frac{P(x_1,x_2,x_3)}{P(x_2)}$

由此，就会有 $P(x_1, x_3|x_2) = \frac{P(x_1)P(x_2|x_1)P(x_3|x_2)}{P(x_2)} = \frac{P(x_1,x_2)}{P(x_2)}P(x_3|x_2) = P(x_1|x_2)P(x_3|x_2)$

至此，我们就成功证明了$X_1 \perp\!\!\!\perp X_3 | X_2$

【课堂练习】分叉结构中的证明



图的基本结构块

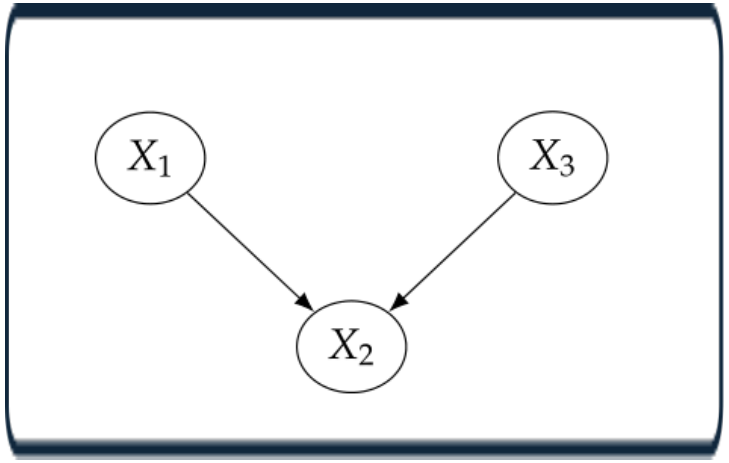
- 三个节点的结构：
  - 对撞结构
    - 对撞结构(不道德结构)如图所示，其中节点X_2被称为Collider
    - 不同于链式和分叉结构，在对撞结构中，$X_1 \perp\!\!\!\perp X_3$

证明：$P(x_1, x_3) = \sum_{x_2} P(x_1, x_2, x_3)$

$= \sum_{x_2} P(x_1)P(x_3)P(x_2|x_1, x_3)$

$= P(x_1)P(x_3) \sum_{x_2} P(x_2|x_1, x_3)$

$= P(x_1)P(x_3)$

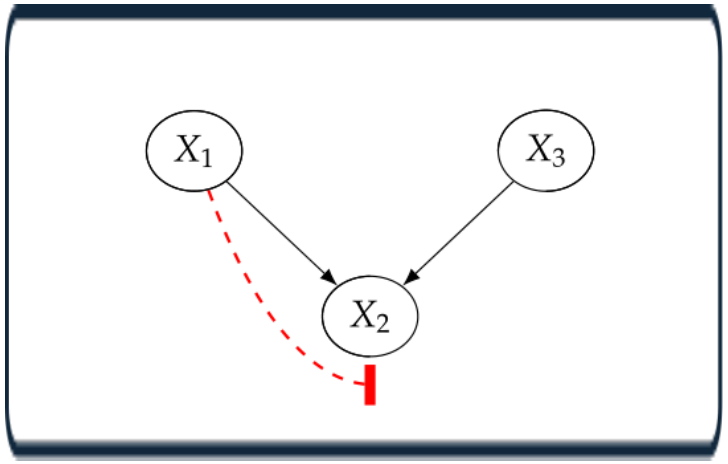


对撞结构

- X_1与X_3既不像链式结构中存在着路径，也不像分叉结构中存在着共同原因

图的基本结构块

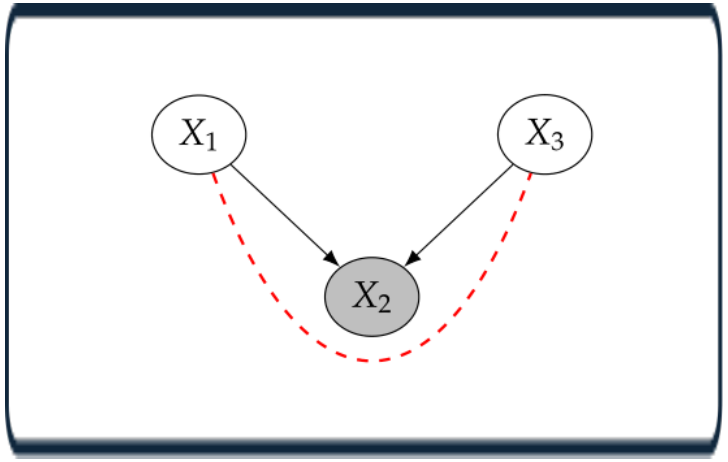
- 三个节点的结构：
  - 对撞结构
    - 为了直观地说明$X_1 \perp\!\!\!\perp X_3$，我们可以画出对撞结构中的关联流，如图所示，我们期望关联流沿着路径$X_1 \rightarrow X_2 \leftarrow X_3$流动，但实际上，关联流却被节点$X_2$阻塞了
    - X_1和X_3之间存在着对撞点(Collider)X_2，因此关联流无法沿着这条路径流动
    - 这也是一条阻塞路径，但路径并不是被条件所阻塞，而是被对撞机阻塞



对撞结构中，关联流被Collider所阻塞

图的基本结构块

- 三个节点的结构：
  - 对撞结构
    - 奇怪的是，如果我们将对撞机X_2作为条件，那么X_1和X_3之间就会变得相互依赖
    - 将对撞机X_2作为条件，就可以将一个阻塞的路径变为未被阻塞的路径



对撞结构中，固定X_2作为条件， X_1和X_3之间就会存在关联流

图的基本结构块

- 三个节点的结构：
  - 对撞结构
    - 表中给出了两枚质地均匀的硬币同时投掷的结果的概率分布。其中，X表示第一枚硬币，Y表示第二枚硬币，Heads代表正面向上，Tails代表反面向上，Z代表铃，如果任一枚硬币落地时正面向上，则铃响

X	Y	Z	$P(X,Y,Z)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	0	0.25

    - 由表中我们可以得知，$P(X = \text{正面}|\text{Y} = \text{正面}) = P(X = \text{反面}|\text{Y} = \text{反面}) = \frac{1}{2}$
      也就是说，X和Y是相互独立的

图的基本结构块

- 三个节点的结构：
  - 对撞结构
    - 现在以$Z=1$(铃声响)和$Z=0$(铃声不响)为条件

X	Y	$P(X,Y\|Z=1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0

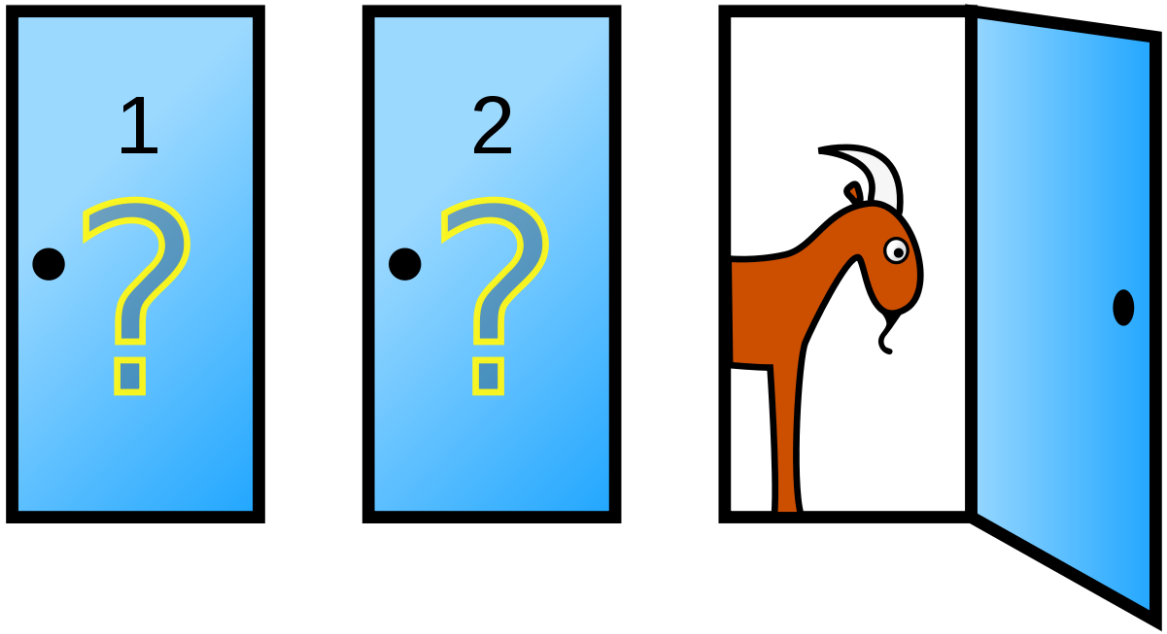
X	Y	$P(X,Y\|Z=0)$
Heads	Heads	0
Heads	Tails	0
Tails	Heads	0
Tails	Tails	1

- 通过计算，我们可以有$P(X=\text{正面}|Z=1)=\frac{1}{3}+\frac{1}{3}=\frac{2}{3}$，$P(X=\text{正面}|Y=\text{正面}，Z=1)=\frac{1}{2}$
  可以看出，在$Z=1$的情况下，当知道$Y=$正面时，$X=$正面的概率从$\frac{2}{3}$变到了$\frac{1}{2}$
  显然，给定$Z=1$后，X和Y就是相互依赖的

图的基本结构块

- 三个节点的结构：
  - 对撞结构
    - 另一个例子：三门问题

假设你参加了一个竞猜游戏类电视节目，在这个游戏中，有三扇门供你选择，其中一扇门后面是一辆车（奖品），另外两扇门后面是山羊。你**挑了一扇门**，比如说1号门，而主持人知道门后面是什么。现在，他打开了**另一扇门**，比如说3号门，你看到这扇门的后面是一只山羊。此时，如果他问你：“你想重新选择，改选2号门吗？“那么，选择换门是否对你赢走奖品更有利？

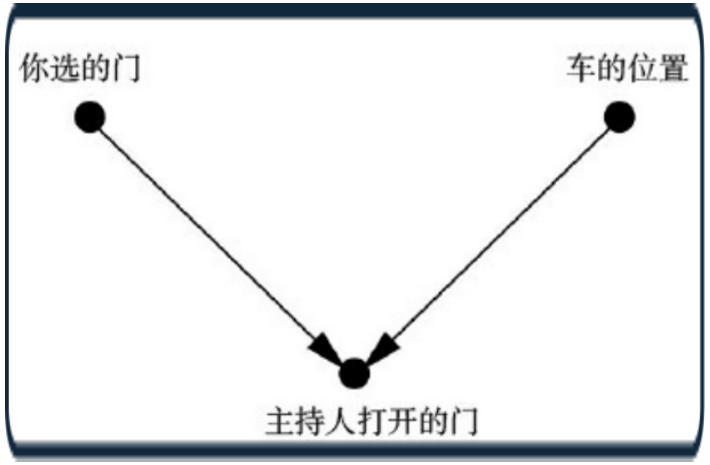


- Prize

图的基本结构块

- 三个节点的结构：
  - 对撞结构
    - 另一个例子：三门问题

“主持人打开的门”是一个对撞机。一旦我们获得了关于这个变量的信息，图示中所有的概率就都变成了关于这一信息的条件概率。但是，当我们以对撞机为条件时，我们就会在两个双亲节点之间制造出一种虚假的依存关系。这种依存可以体现在根据数据得到的概率中：在主持人打开了3号门的前提下，如果你最初选择了1号门，则车在2号门后面的可能性是其在1号门后面的2倍；如果你最初选择了2号门，则车在1号门后面的可能性是其在2号门后面的2倍。



你选的门	主持人选的门	车在的门	不换	换
1	3	1	Win	Lose
1	3	2	Lose	Win
1	2	3	Lose	Win

d-分离

- **d-分离 (d-separation)：**

节点X和Y之间的路径被一组节点Z阻塞，如果下面任意一条成立：

1）在路径中，存在一条链式结构$\ldots \rightarrow W \rightarrow \ldots$或一个分支结构$\ldots \leftarrow W \rightarrow \ldots$，其中固定条件$W(W \in Z)$

2）在路径中，存在一个对撞机W，其中W并没有固定$(W \notin Z)$同时W的任意一个子孙都没有固定$(de(W) \notin Z)$

Notes: 未被阻塞的路径就是不满足上述条件的路径
关联流无法沿着被阻塞的路径流动

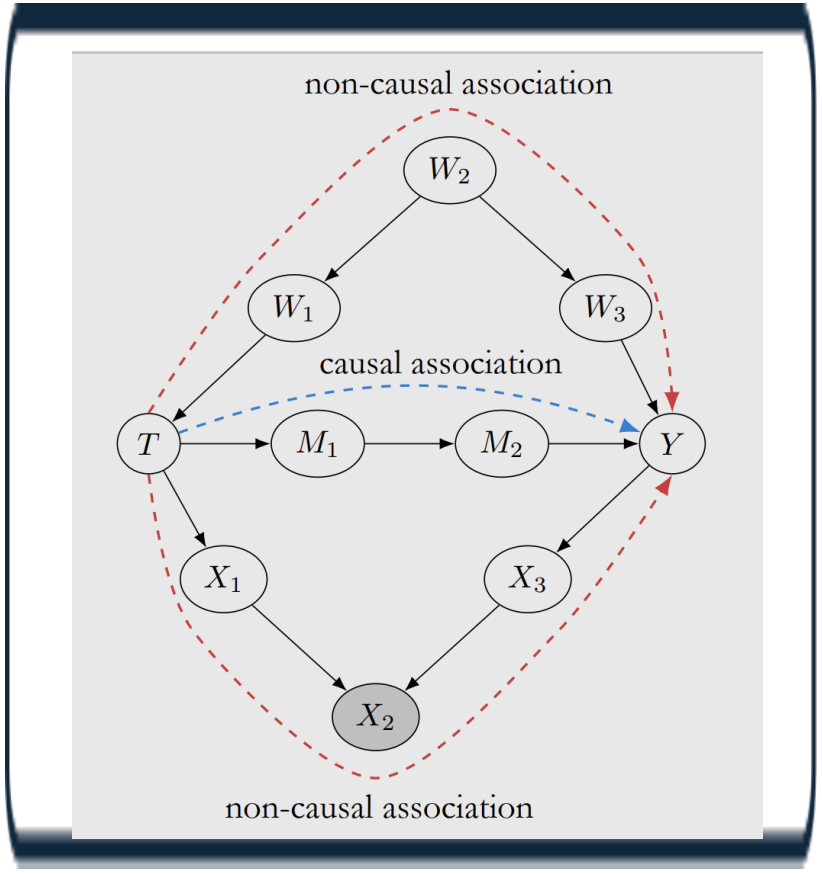
- 如果两个节点X和Y之间存在的任何路径都被一组节点Z阻塞，则它们在Z的条件下是d-分离的。如果X和Y之间存在一条路径没有被阻塞，则它们是d-连通的

- 我们使用符号$X \perp\!\!\!\perp_G Y|Z$来表示在图G中，给定条件Z，X和Y是d-分离的，而用符号$X \perp\!\!\!\perp_P Y|Z$来表示在概率分布P中，给定条件Z，X和Y是相互独立的，等价地，给定概率分布P，关于图G满足局部马尔可夫假设，如果图G中，给定条件Z，X和Y是d-分离的，那么在分布P中，给定条件Z，X和Y是独立的，即

$$X \perp\!\!\!\perp_G Y|Z \Rightarrow X \perp\!\!\!\perp_P Y|Z$$

关联和因果关系

- 关联会在有向图中沿着所有未阻塞的路径流动，但在因果图中，因果关系会沿着有向路径
- 我们将沿着有向路径的关联流称为因果关联
- 关联和因果关系的一个重要的区别：关联是对称的，而因果是不对称的

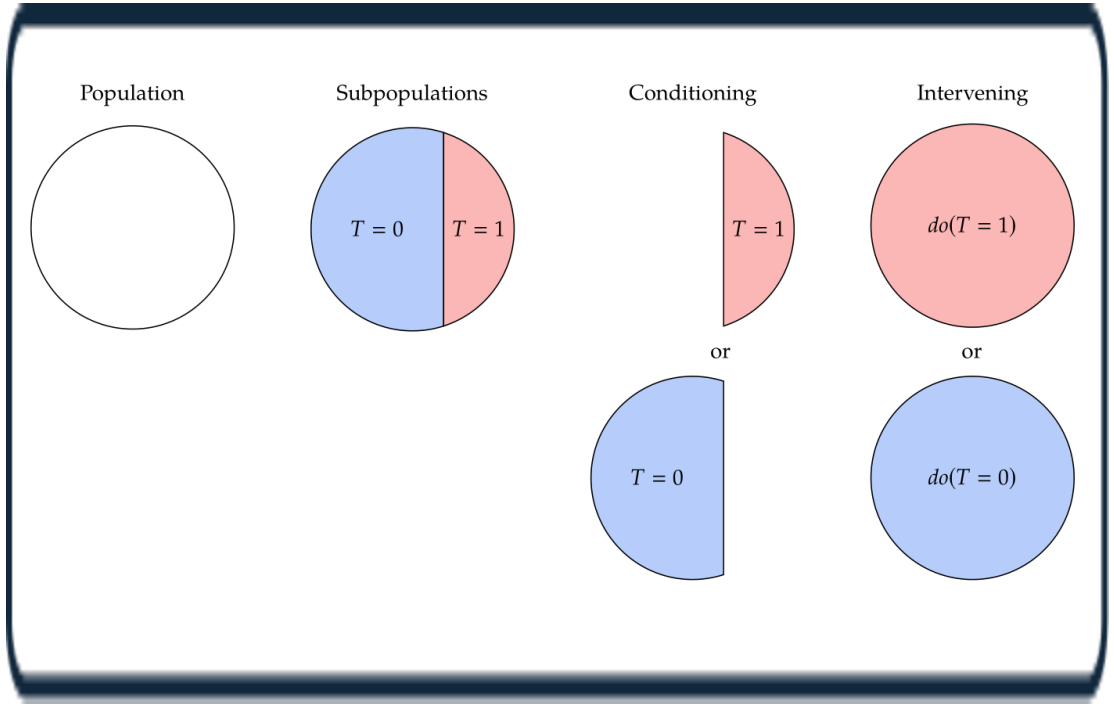


第三节 因果模型

do算子和干预

- 在概率中，我们有以...为条件的概念，但这与干预不同。以$T = t$为条件仅仅意味着我们将关注点限制在整体人群中接受treatment=t的这一部分人群。相比之下，干预是让整体人群都接受treatment=t，而不管观察到的其本身的treatment是否为t

- 通常用do算子表示干预操作，即让整体人群都接受treatment=t 等价于 $do(T = t)$。如图所示，subpopulations表示观察到的数据中蓝色部分是$T = 0$的集合，红色部分是$T = 1$的集合。Conditioning表示我们只关注其中的蓝色部分或红色部分。$do(T = 1)$是指让本身$T = 0$的蓝色部分也变成$T = 1$，即红色



do算子和干预

- 还记得潜在结果(potential outcome)吗？$Y(t)$和$do(Y=t)$是等价的，$Y(t)$的分布可以写成：

$$P(Y(t)=y) \triangleq P(Y=y|do(T=t)) \triangleq P(y|do(t))$$

- 平均因果效应ATE就可以写成如下形式：

$$E[Y|do(T=1)]-E[Y|do(T=0)]$$

- 我们更关心 $P(Y=y \mid do(t))$ 而非其均值，有了概率分布，期望自然就求出来了。我们将 $P(Y=y \mid do(t))$及其他包含do算子的概率分布统称为干预分布

Notes:

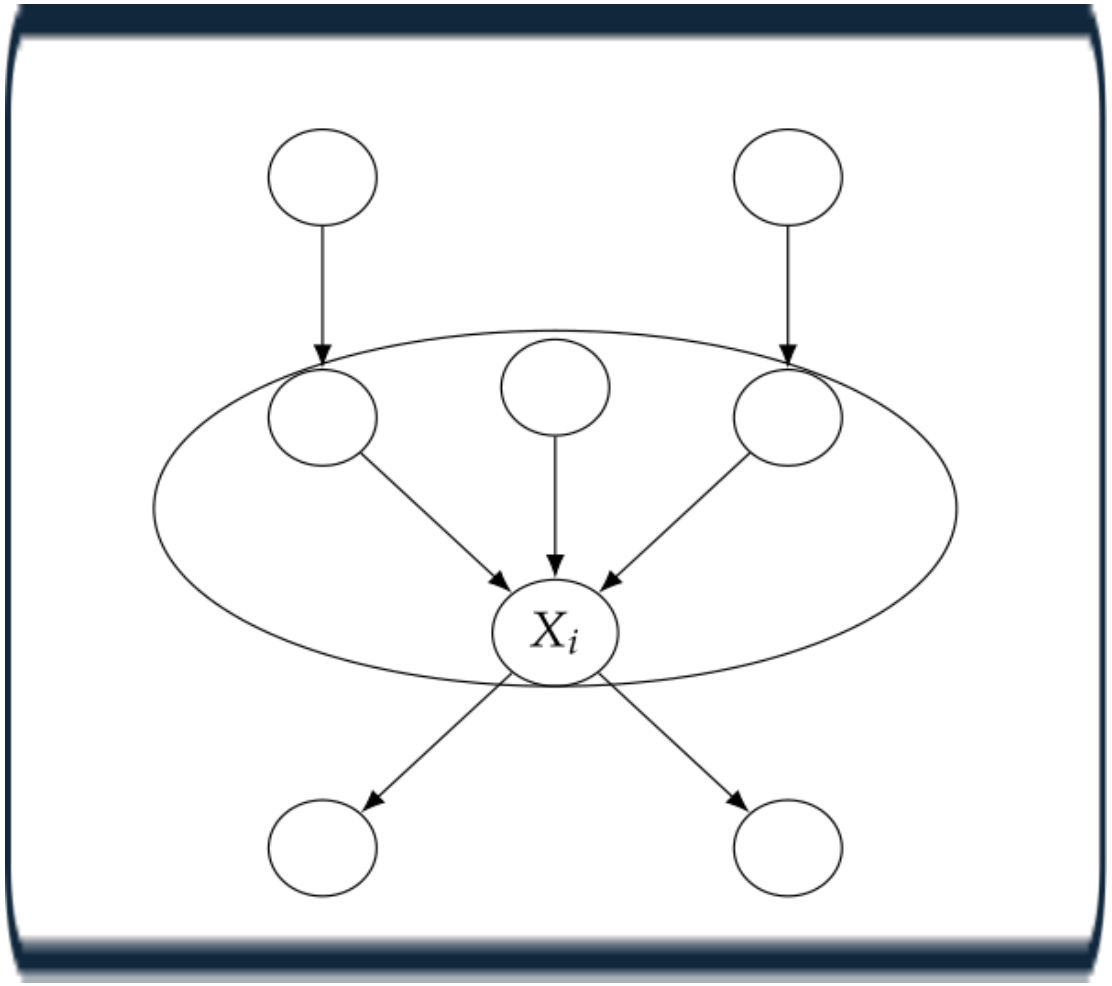
1.干预分布$P(Y=y \mid do(t))$和观察分布$P(Y)$ 有本质的区别。观察分布$P(Y)$或$P(Y,T,X)$中没有．do算子，所以我们可以从观察到的数据中直接求得而不需要做任何额外的实验。如果可以将包含do算子的表达式Q化简成不包含do的形式，那么Q就是可识别的 (Identifiable)。

2.我们会将包含do算子的估计称为因果估计，而将不包含do算子的估计称为统计估计

3.不论何时，每当do算子出现在“|”之后，都意味着**该表达式中的一切都在干预措施发生后的情况**下得到的。 例如$E[Y \mid do(t), Z=z]$ 表示在$Z=z$这个子集中让其中所有个体的treatment都等于t后Y的期望。相反， $E[Y \mid Z=z]$ 表示在 $Z=z$ 这个子集中被干预之前的期望。

模块化假设

- 首先，我们必须明确什么是因果机制。我们将产生X_i的因果机制指定为X_i的条件概率分布$P(x_i|pa_i)$。正如图所示，产生X_i的因果机制是所有X_i的双亲节点及其指向X_i的边。



模块化假设

- 模块化假设：如果对节点集合S进行干预，将其中的变量设为常数，对于任意节点i：

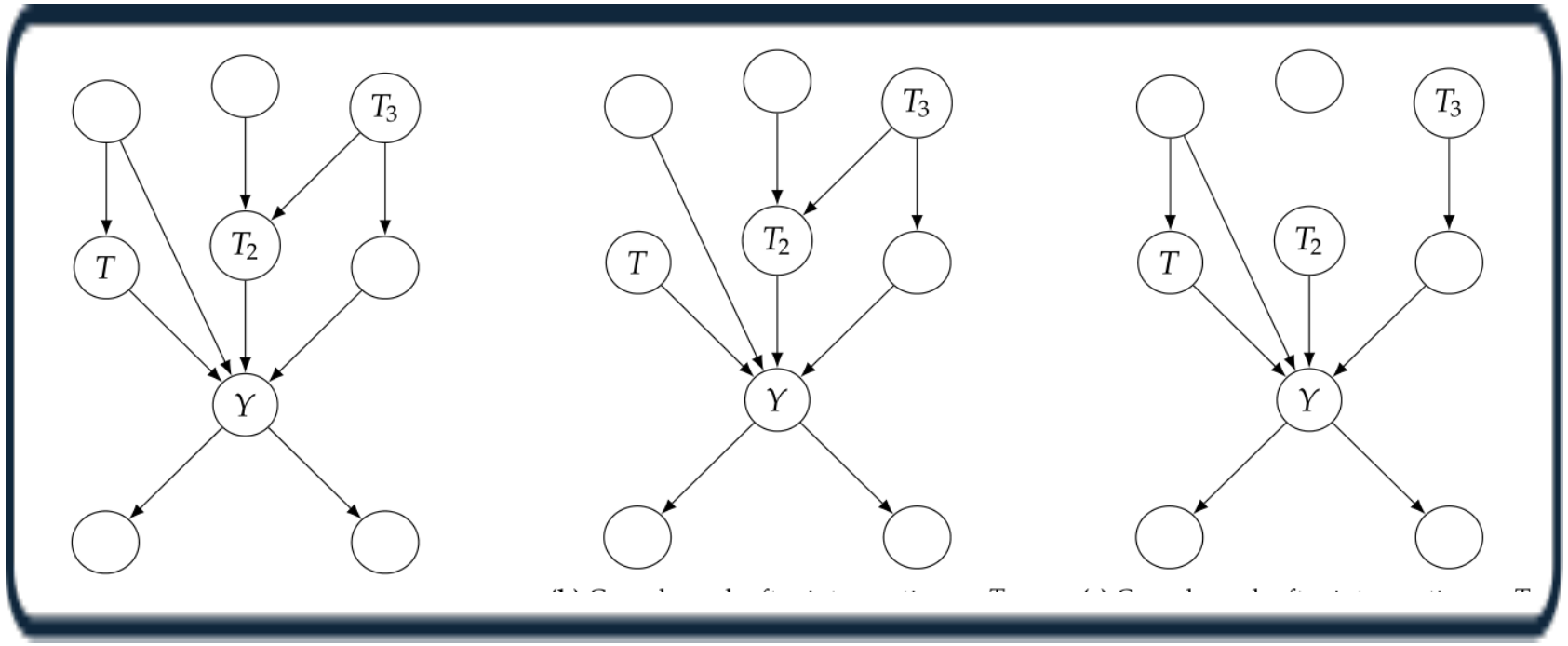
  1）如果节点i不在集合S中，那么其条件概率分布保持不变

  2）如果节点i在集合S中，如果x_i是变量X_i被干预后指定的值，那么$P(x_i|pa_i)$一定为1，否则为0(如果x_i和干预一致，即x_i等于X_i被干预后的值，则$P(x_i|pa_i) = 1$)

- 模块化假设允许我们只在一个图中就可以编码不同的干预分布。例如$P(Y), P(Y|do(T = t))$, $P(Y|do(T = t'))$以及$P(Y|do(T = t_2))$是完全不同的分布，但是我们都可以用表示联合概率分布的图来表示，并且除了被干预的因素，其他的因素在这些图中被共享。

模块化假设

- 干预分布的因果图与用于联合分布的图相同，只不过是移除了指向干预节点的所有边： 这是因为被干预节点的条件概率分布 $P(X_i = x_i|pa_i)$已经是 1了，因此我们可以忽略该因素。



(a)　(b)　(c)

另一种解释是既然干预节点已经设置为常数，那么它必然不会受到双亲节点的影响，因此可以去掉之间的因果关系。 删掉边的图称为manipulated graph。以图为例，对T干预对应(b)，对 T2 干预对应(c).

截断因式分解

- 回顾下贝叶斯网路中联合概率分布的分解形式：

$$P(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}) = \prod_i P(x_i|pa_i)$$

现在对节点集合S进行干预，并假设其符合模块化假设，则对于$\mathrm{x}_i \notin S, P(x_i|pa_i)$和干预前的值保持一样。对于$\mathrm{x}_i \in S, P(x_i|pa_i) = 1$。(值与干预一致时)

- 截断式因式分解：我们假设概率分布P和图G满足马尔可夫假设和模块化假设，给定一组干预节点S，如果x与干预一致，则：

$$P(x_1, \ldots, x_n|do(S = s)) = \prod_{i \notin S} P(x_i|pa_i)$$

否则：

$$P(x_1, \ldots, x_n|do(S = s)) = 0$$

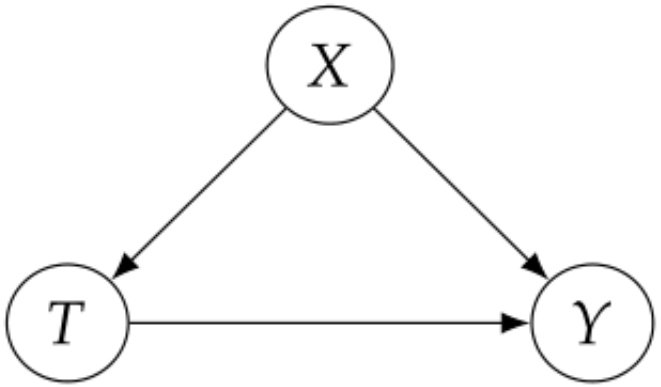
截断因式分解

- 示例：

  如图所示，联合概率分布可以表示为：$P(y,t,x) = P(x)P(t|x)P(y|t,x)$

  对T进行干预后有：$P(y,x\,|do(t)) = P(x)P(y|t,x)$　　干预后：$P(t=t|x)=1$

  y的边缘概率分布为：$P(y|do(t)) = \sum_x P(y\mid t,x)P(x)$



截断因式分解

- 示例：

通过比较干预分布和正常的条件概率分布的差别，可以更深刻地理解为什么“关联不是因果”

$$P(y|do(t)) = \sum_x P(y \mid t,x)P(x)$$

$$P(y|t) = \sum_x P(y,x \mid t) = \sum_x P(y \mid t,x)\, P(x \mid t)$$

可以看到，两个式子的差别在于一个是$P(x)$，一个是$P(x|t)$

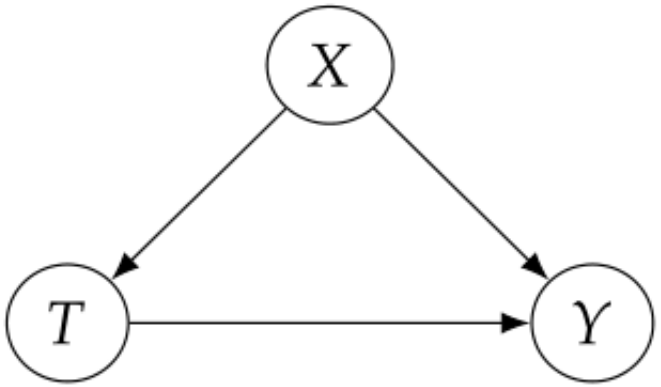
如果简化这个例子，假设T是一个二值的变量，我们想计算ATE，$P(y|do(T=1))$就是潜在结果$Y(1)$的概率分布，因此可以通过求期望得到 $E[Y(1)]$，同理得到 $E[Y(0)]$，因此平均因果效应ATE可以写成：

$$E(Y(1)-Y(0)) = \sum_y yP(y|do(T=1)) - \sum_y yP(y|do(T=0))$$

如果将干预分布代入，则ATE完全可以写成概率的形式，表达式中不含有do，就可以得到可识别的ATE

后门调整

- 如图所示，从T到Y存在两种关联，即$T \rightarrow Y$的因果关联以及$T \leftarrow X \rightarrow Y$的非因果关联。后门路径是指如果一条路径从$T$到$Y$是未被阻塞的，且有指向$T$的边(即$T \leftarrow \cdots$)，则称这条路径是后门路径
- 如果此时对T进行干预，则任何指向T的边都会被去掉，后门路径就被阻塞了，T和Y之间就只剩因果关联了
- 如果固定X作为条件，同样会阻塞后门路径



后门调整

- 对于T和Y，如果以下条件成立的话，则称变量集合W满足后门准则：
  - 固定W作为条件可以阻塞T和Y之间的所有后门路径
  - W不包含T的所有子孙节点
- 将W引入到$P(y|do(t))$中，可以得到：

$$P(y|do(t)) = \sum_w P(y|do(t), w)P(w \mid do(t)) = \sum_w P(y|t,w)P(w|do(t)) = \sum_w P(y|t,w)P(w)$$

- 后门调整：假设变量集合W满足后门准则，那么T对Y的因果效应可由以下公式计算：

$$P(y|do(t)) = \sum_w P(y|t, w)P(w)$$

证明：

$$\begin{aligned} P(y|do(t)) &= \sum_w P(y|do(t), w)P(w|do(t)) \\ &= \sum_w P(y|t, w)P(w|do(t)) \\ &= \sum_w P(y|t, w)P(w) \end{aligned}$$

后门调整

- 后门调整看起来和潜在结果中的调整公式非常相似，它们之间有什么联系吗？

对干预后的Y求期望：

$$
\begin{aligned}
E[Y|do(t)] &= \sum_y yP(Y=y|do(t)) \\
&= \sum_y y \sum_w P(y|t,\ w)P(w) \\
&= \sum_w \sum_y yP(y|t,\ w)\,P(w) \\
&= \sum_w E[Y|t,\ w]P(w) \\
&= E_w[E[Y|t,\ W]]
\end{aligned}
$$

把$T=1$和$T=0$代入得 $E[Y|do(T=1)] - E[Y|do(T=0)] = E_W\left[E[Y|T=1,\ W] - E[Y|T=0,\ W]\right]$

可以发现，$E[Y|do(t)]$是潜在结果$E[Y(t)]$的另一种表示形式

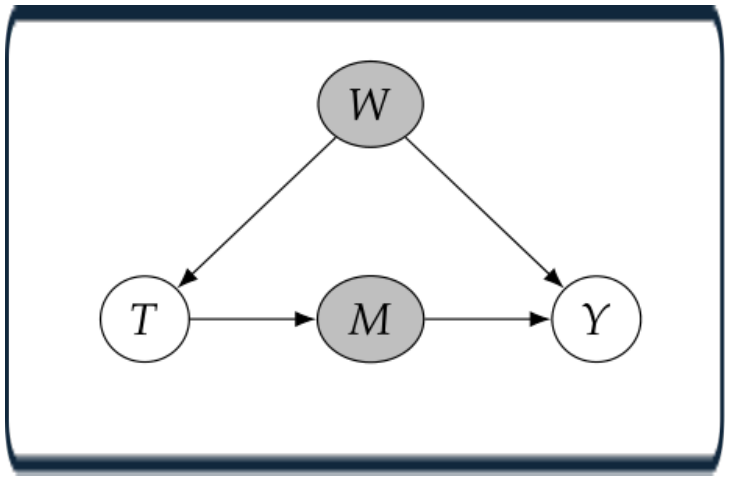
后门调整

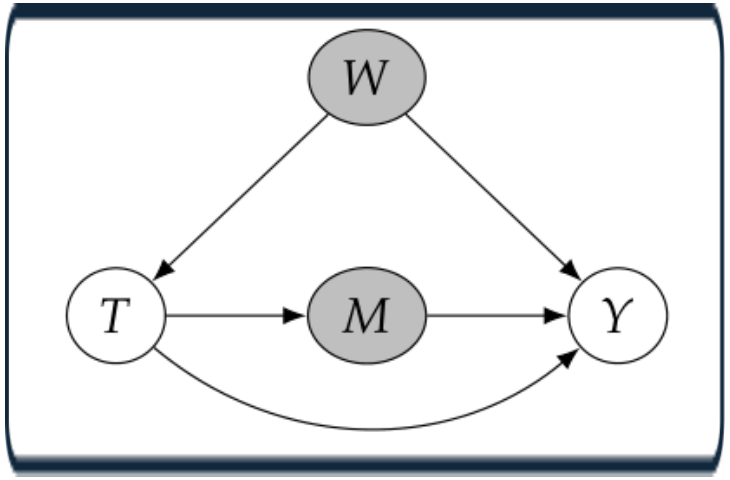
- 在后门准则中，我们要求固定W可以阻塞T和Y之间的所有后门路径，同时W不包含T的所有子孙节点。如果我们固定T的子孙，可能会有两类问题：

  1）阻塞了从T到Y的因果关系流动

  2）引起了T和Y之间的非因果关联

- 如果我们以从 T 到 Y 的有向路径上的节点为条件，那么我们将阻止因果关系沿该因果路径流动。例如，在下面两个图中，控制M分别阻塞了全部和部分因果路径





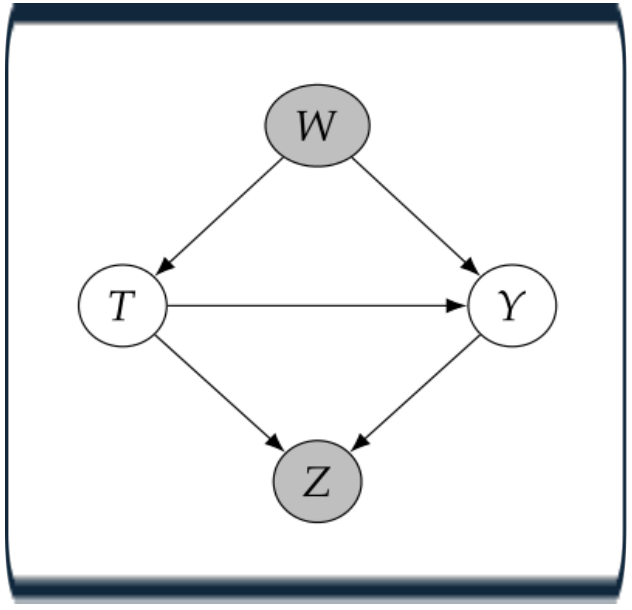
后门调整

- 在后门准则中，我们要求固定W可以阻塞T和Y之间的所有后门路径，同时W不包含T的所有子孙节点。如果我们固定T的子孙，可能会有两类问题：

  1）阻塞了从T到Y的因果关系流动

  2）引起了T和Y之间的非因果关联

- 如果以不在从T到Y的有向路径上的T的后代节点为条件，则它可能会打通被对撞节点阻塞的关联路径，如下图中以Z为条件



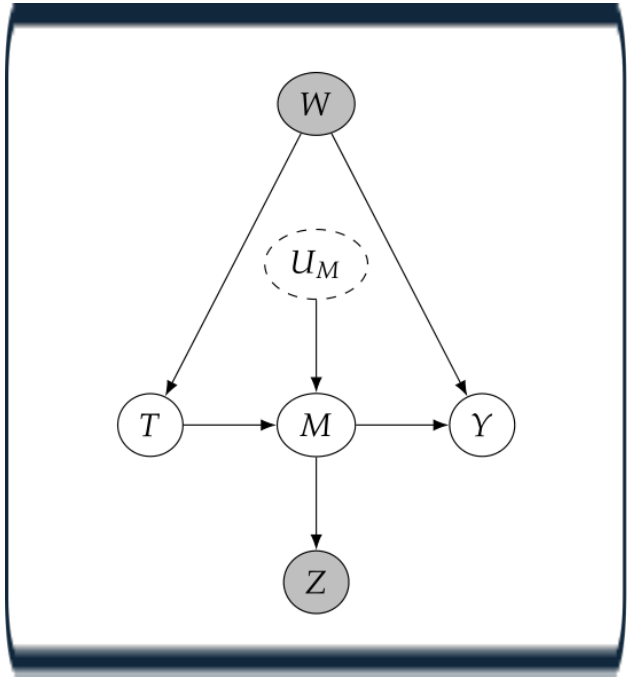
后门调整

- 在后门准则中，我们要求固定W可以阻塞T和Y之间的所有后门路径，同时W不包含T的 所有子孙节点。如果我们固定T的子孙，可能会有两类问题：

  1）阻塞了从T到Y的因果关系流动

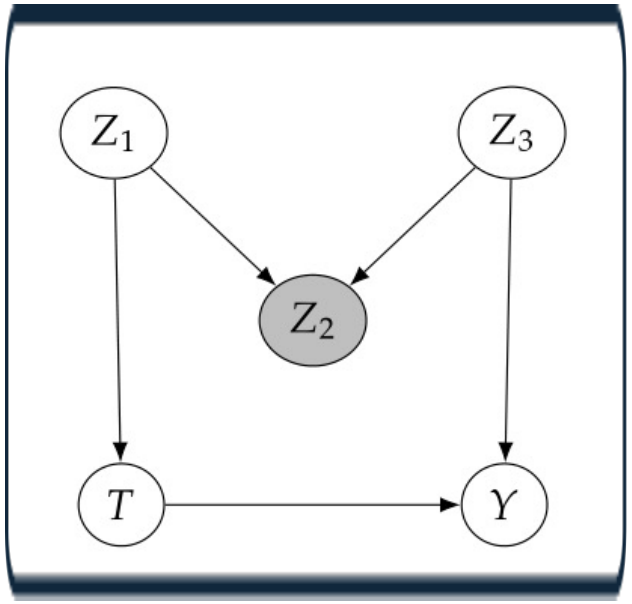
  2）引起了T和Y之间的非因果关联

- 此外，对于下图，如果以Z为条件，会打通T与U_M之间的关联路径，从而对T到Y的因果路径产生影响



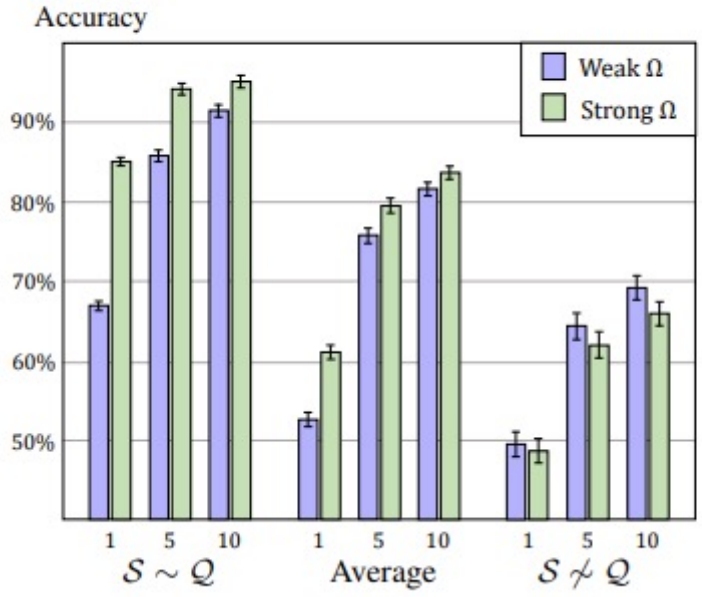
后门调整

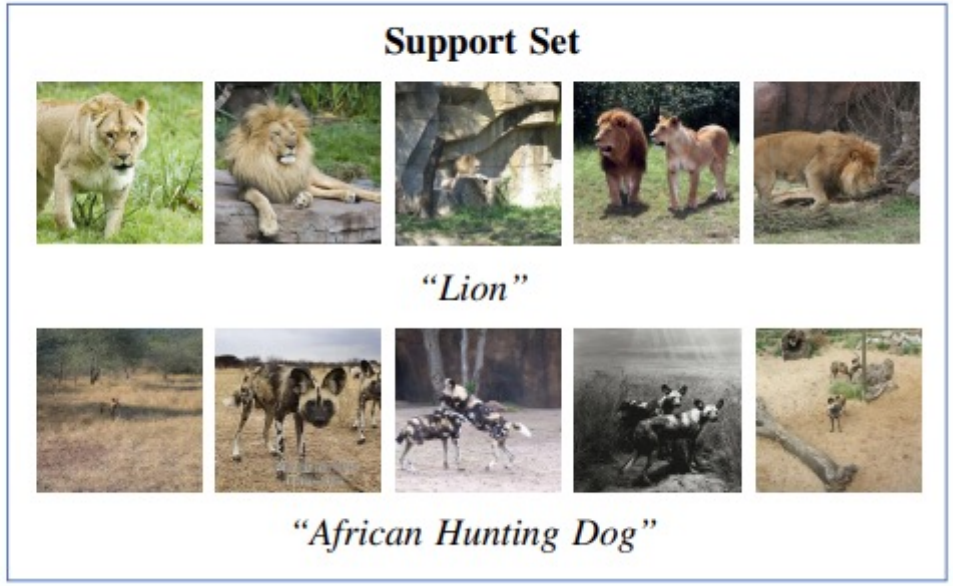
- 有时，仅仅以干预前变量为条件也会引入偏倚，例如M偏倚
- M偏倚：如图，如果固定对撞机Z_2，那么我们就打开了一条后门路径，非因果关联可以沿着这条路径流动，这被称为M偏倚。因为非因果关联会沿着M形状流动



后门调整

- 后门调整的应用： **Interventional Few-Shot Learning**(NeurIPS 2020)
  - 为了在少量样本上快速泛化，需要借助先验知识，而预训练就是有效获取先验知识的方法。但是预训练也同时成为了学习过程中的一个混杂因子

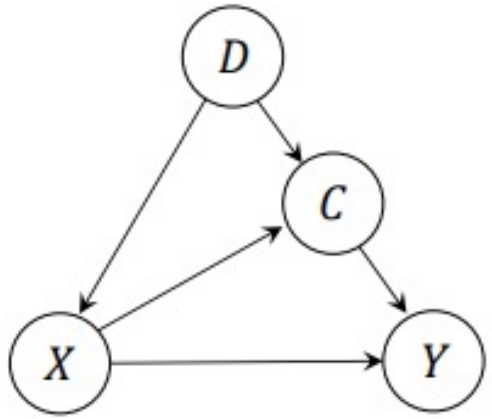




- 例如，在图示中，预训练中的草的颜色成为了一个误导因素，分类器很容易以草的颜色而非动物本身作为分类依据

后门调整

- 后门调整的应用： **Interventional Few-Shot Learning**(NeurIPS 2020)
  - 可以使用右图所示的因果图来表示，其中：
    - D代表预训练的先验知识
    - X代表图片的特征表示
    - C代表一个样本X在预训练数据流形上面的投影
    - Y代表分类器所预测的标签



  - 可以看出，D是X和Y的共因，也就是两者的混杂因子，为了去除D的影响，需要对X进行干预
  - 论文中通过后门调整来实现对X的干预：$p(y|do(x)) = \sum_d p(y|x,d,c)p(d)$。具体来说，就是对预训练的知识 D 进行分层$(d_1, \ldots, d_n)$，每一层有自己的分类器 $P(Y|X,D,C)$，然后把每层分类器的结果通过先验概率 $P(D)$ 平均起来

结构等式

- Judea Pearl说过，数学里的“=”不包含任何的因果信息，$A = B$和$B = A$表示的都是同一个意思，“=”是对称的。但是为了表达因果，需要有一个非对称的符号。如果A是B的原因，那么改变A一定会改变B，但是反之不成立，我们可以用结构等式来表示：

$$B := f(A)$$

- 我们将“ = ”替换成“: = ”。但是，B和A之间的映射关系是确定性的。理想情况下，我们希望它是概率性的，为一些未知因素留出空间。因此可以写成下面这样：

$$B := f(A,\ U)$$

其中，U是观察到的随机变量，在图中用虚线表示，未观察到的 U 类似于我们通过抽样个体看到的随机性；它表示确定 B 的所有相关（嘈杂）背景条件。f 的函数形式无需指定，当不指定时，我们处于非参数状态，因为我们没有对参数形式做出任何假设。 虽然映射是确定性的，但由于它以随机变量 U （“噪声”或“背景条件”变量）作为输入，它可以表示任何随机映射，因此结构方程是$P(x_i|pa_i)$的推广形式。 因此，当我们引入结构方程后，截断因式分解和后门调整仍然成立。

结构等式

- 有了结构等式后，我们可以更详细的定义原因和因果机制。生成变量的因果机制是与该变量相对应的结构方程。例如，生成B的因果机制是：

$$B := f(A, U)$$

- 类似的，如果X出现在结构等式的右边，则X是Y的直接原因。

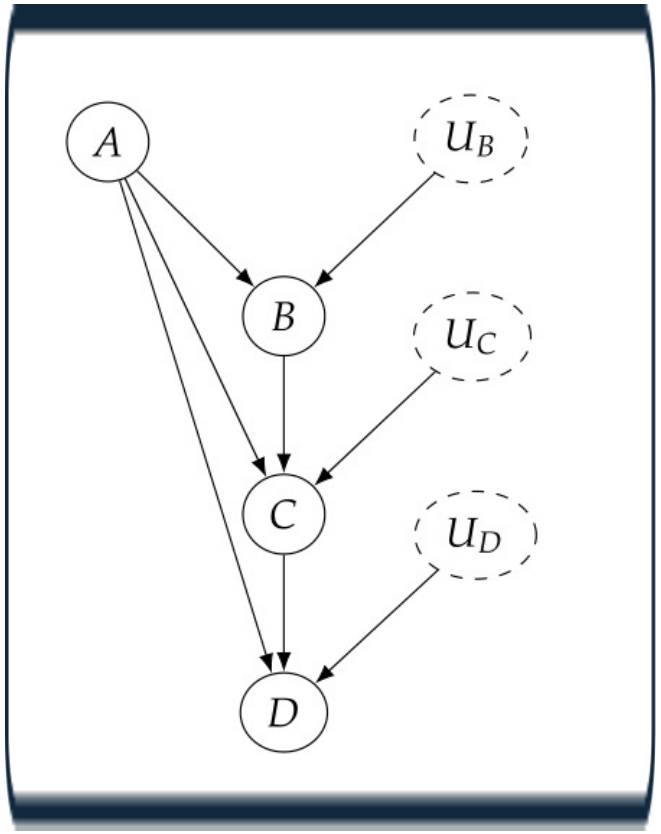
- 我们会称X是Y的原因，如果X是Y的任意一个原因的直接原因或者X是Y的一个直接原因

结构等式

- 考虑更复杂的结构等式

$$
\begin{aligned}
B &:= f_B(A, U_B) \\
C &:= f_C(A, B, U_C) \\
D &:= f_D(A, C, U_D)
\end{aligned}
$$

在因果图中，噪声变量通常是隐式的（虚线），而不是明确绘制的。我们写结构方程时已知的变量 称为内生变量， 这些是我们正在建模因果机制的变量 - 在因果图中具有父母的变量。 相反，外生变量是因果图中没有任何父母的变量。 这些变量是我们因果模型外部的，因为我们没有为其建模因果机制。 例如，在右图描述的因果模型中，内生变量为$\{B, C, D\}$。 外源变量为$\{A, U_B, U_C, U_D\}$



结构等式

- 结构因果模型(SCM)：结构因果模型是包含以下集合的元组

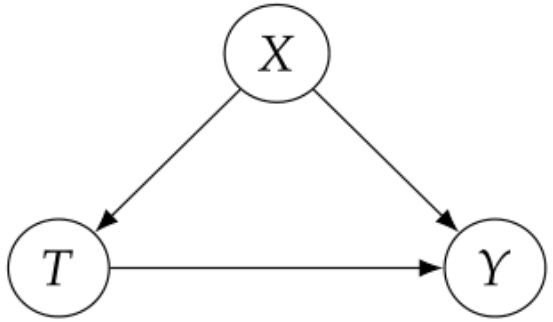
  1）一组内生变量V

  2）一组外生变量U

  3）一组生成内生变量的函数f

干预

- 从SCM的角度来描述干预会非常简单，对T进行干预$do(T = t)$相当于将T的结构等式替换成$T := t$



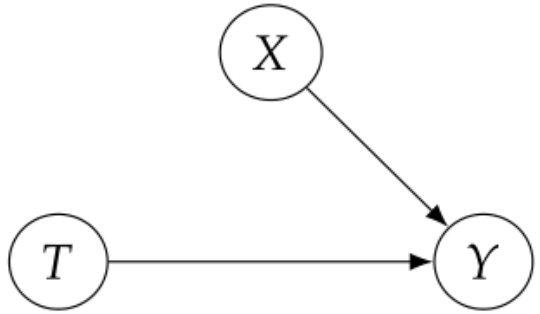
(a)

- 例如：图(a)对应的SCM为：

$$T := f_T(X, U_T)$$
$$Y := f_Y(X, T, U_Y)$$

如果对T进行干预，让其等于t，那么干预后的SCM则为：

$$T := t$$
$$Y := f_Y(X, T, U_Y)$$



(b)

- 可以发现，除了T本身的结构等式，其他的等式都保持不变。这也是由模块化假设决定的