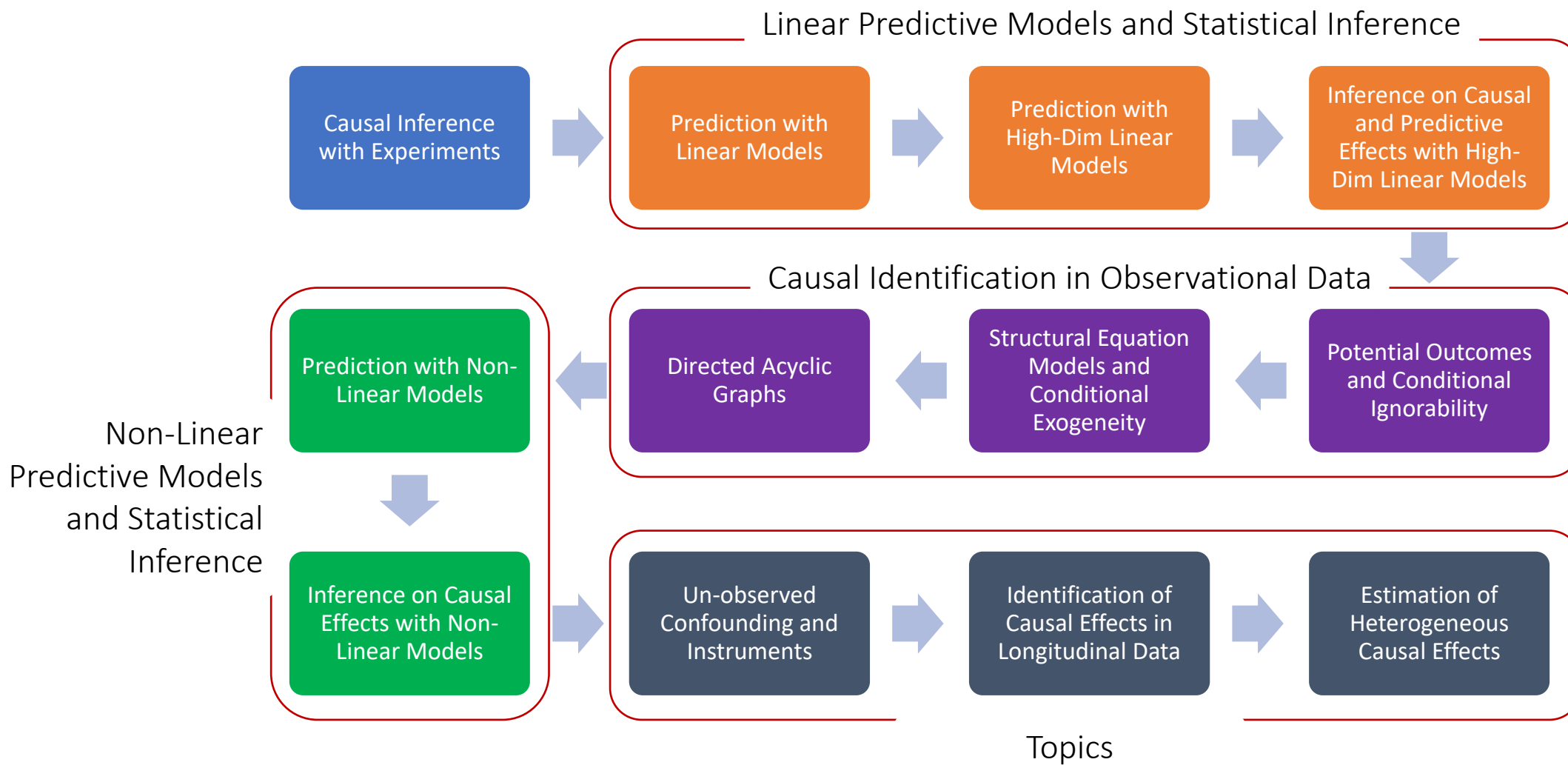
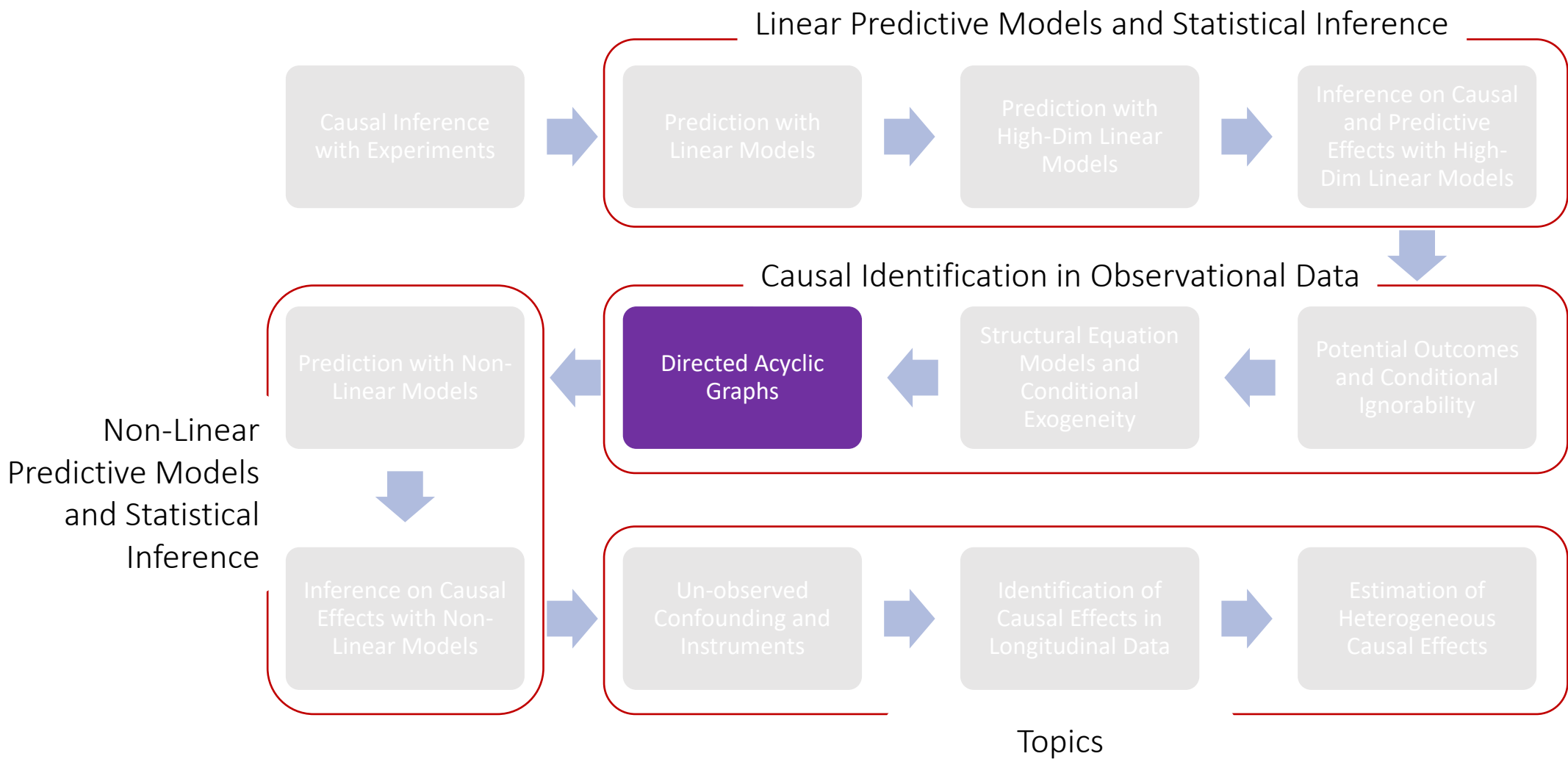


MS&E 228: Directed Acyclic Graphs and Non-Linear SEMs

Vasilis Syrgkanis

MS&E, Stanford





Goals for Today

- Learn the “language” of Directed Acyclic Graphs (DAGs) and their associated non-linear structural equation models (SEMs)
- Introduce “intervention” concepts “do” and “fix”
- Introduce d-separation and conditional independence in DAGs
- Proof sketch of fundamental theorem d-separation \Rightarrow conditional ind.

Next lecture

- Graphical criteria for selection of adjustment set
- Crash course on good and bad “controls”

DAGs

Judea Pearl. 'Causal diagrams for empirical research'. In: *Biometrika* 82.4 (1995), pp. 669–688 (cited on page 30).

Trygve Haavelmo. 'The probability approach in econometrics'. In: *Econometrica: Journal of the Econometric Society* 12 (1944), pp. iii–vi+1–115 (cited on pages 30, 32).

James Heckman and Rodrigo Pinto. 'Causal analysis after Haavelmo'. In: *Econometric Theory* 31.1 (2015 (NBER 2013)), pp. 115–151 (cited on pages 30, 35).

James Robins. 'A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect'. In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512 (cited on page 54).

Thomas S. Richardson and James M. Robins. *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality*. Working Paper

Non-Linear SEMs

Non-Linear SEMs

- Last time we looked at linear SEMs
- The language of SEMs does not really rely on the linearity assumption
- For example, the Triangular Structural Equation (TSEM)

$$Y := \delta P + \beta' X + \epsilon_Y$$

$$P := \nu' X + \epsilon_P$$

$$X := \epsilon_X$$

Non-Linear SEMs

- Last time we looked at linear SEMs
- The language of SEMs does not really rely on the linearity assumption
- For example, the Triangular Structural Equation (TSEM)
- Can be made non-linear

$$Y := f_Y(P, X, \epsilon_Y)$$

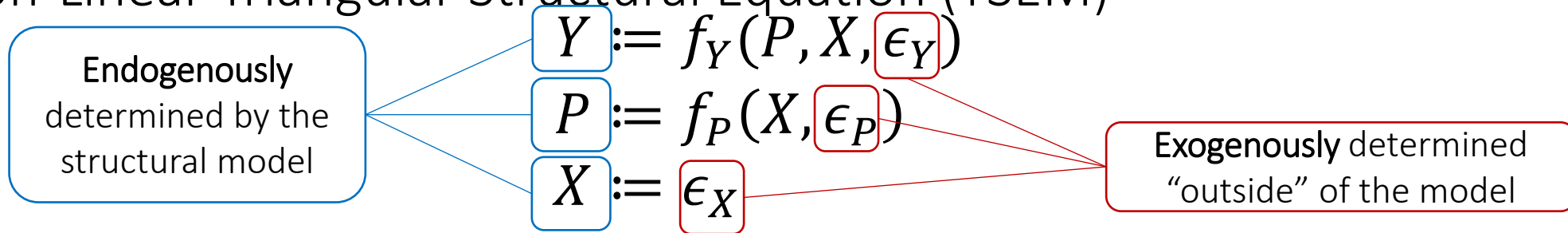
$$P := f_P(X, \epsilon_P)$$

$$X := \epsilon_X$$

- While we still maintain that $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks

Non-Linear SEMs

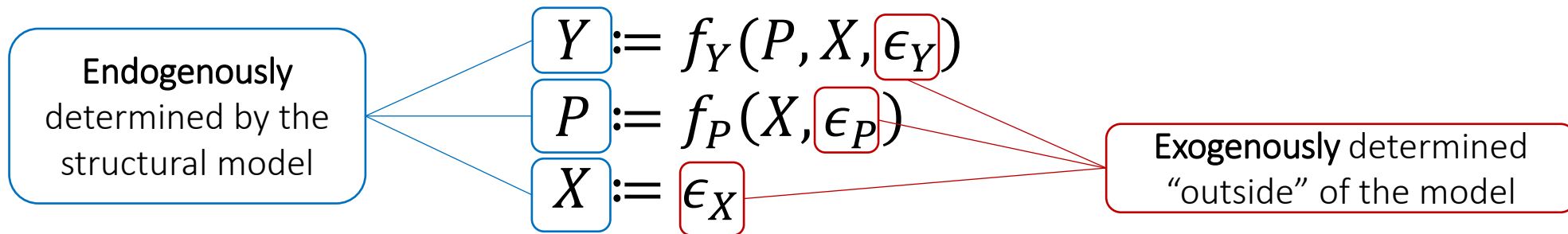
- Non-Linear Triangular Structural Equation (TSEM)



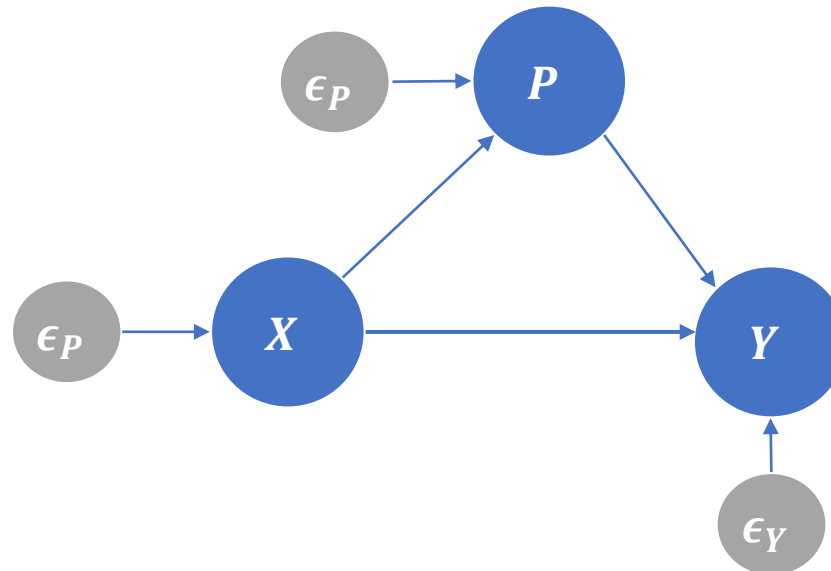
- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks
- The functions f_Y, f_P are now the deterministic “structural functions”
- Instead of “structural parameters” we now have “structural functions”
- Moreover, the dimension of exogenous shocks is un-restricted
- Note that the TSEM implies: $\epsilon_Y \perp\!\!\!\perp P, X$ and $\epsilon_P \perp\!\!\!\perp X$

Non-Linear SEMs

- Non-Linear Triangular Structural Equation (TSEM)

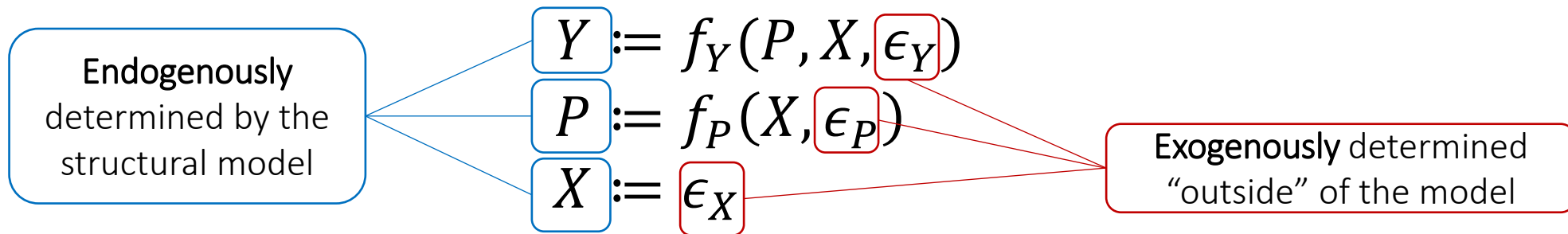


- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent "exogenous" shocks

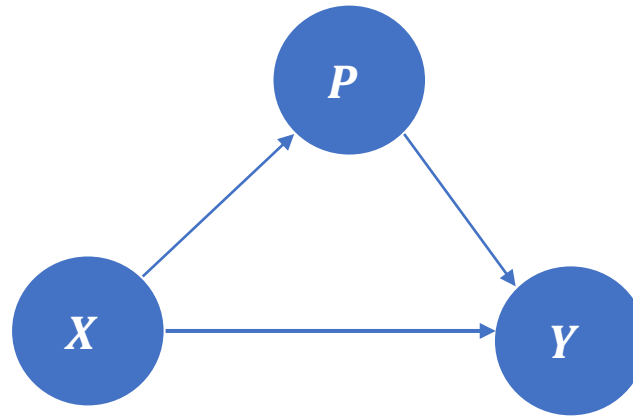


Non-Linear SEMs

- Non-Linear Triangular Structural Equation (TSEM)

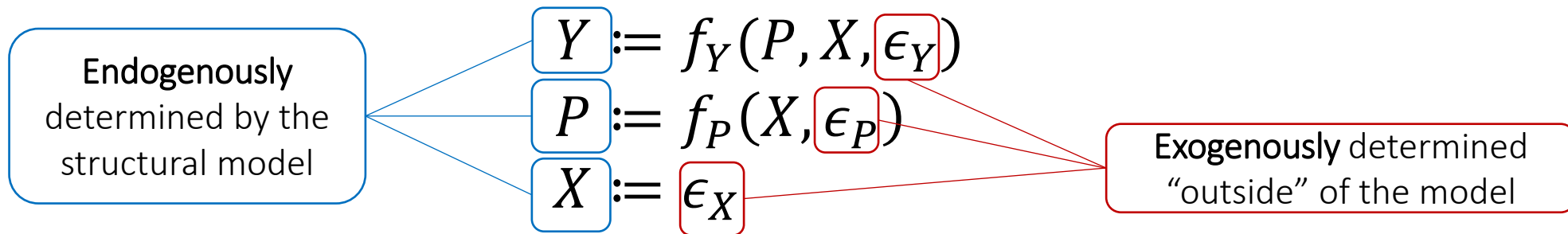


- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks; typically omitted from DAG visualization



Non-Linear SEMs

- Non-Linear Triangular Structural Equation (TSEM)



- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks; typically omitted from DAG visualization
- A TSEM is simply a statistical “generative” model that determines a distribution over observed random variables

Structural Form

- TSEM is “structural” in that it is endowed with the following properties
- Made up of a collection of stochastic potential outcome processes indexed by (p, x)

$$Y(p, x) := f_Y(p, x, \epsilon_Y)$$

$$P(x) := f_P(x, \epsilon_P)$$

$$X := \epsilon_X$$

- **Exogeneity:** $\epsilon_P, \epsilon_Y, \epsilon_X$ are independent “shock” variables generated outside of the model
- **Consistency:** endogenous variables (Y, P, X) generated by recursive substitutions
$$Y := Y(P, X), \quad P := P(X), \quad X := \epsilon_X$$
- **Invariance:** structure remains invariant to changes of distributions of shocks

Link to Potential Outcomes

- Consider (for simplicity) binary treatments $p \in \{0,1\}$
- Suppose that potential outcomes are generated wlog as:

$$Y(p) := g(p, X, \epsilon_Y(p))$$

- This is equivalent to a SEM where we define $\epsilon_Y = (\epsilon_Y(0), \epsilon_Y(1))$ and

$$Y(p) := f_Y(p, X, \epsilon_Y)$$

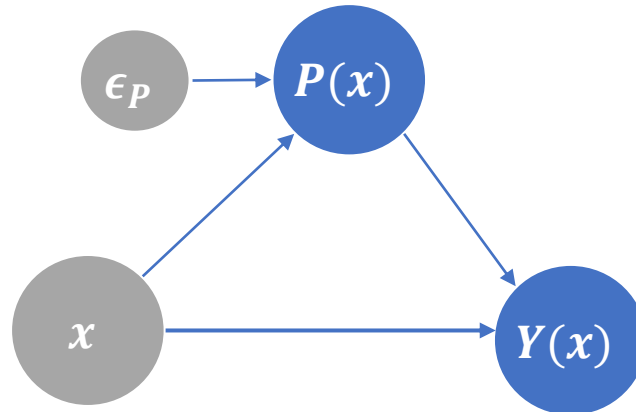
- where

$$f_Y(p, X, e) = p \cdot g(p, X, e(1)) + (1 - p) \cdot g(p, X, e(0))$$

Identification of Structural Responses

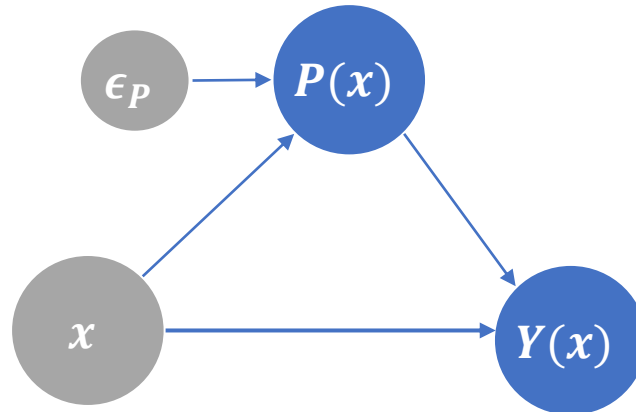
Identification by Regression Revisited

- If we condition on $X = x$ it's as if we're altering the graph and SEM
$$Y = f_Y(P(x), x, \epsilon_Y), \quad \epsilon_Y \perp\!\!\!\perp P(x)$$
- Notice that remnant variation in $P(x)$, is ϵ_P , which is exogenous
- As if driven by a randomized trial process



Identification by Regression Revisited

- If we condition on $X = x$ it's as if we're altering the graph and SEM
$$Y = f_Y(P(x), x, \epsilon_Y), \quad \epsilon_Y \perp\!\!\!\perp P(x)$$
- Notice that remnant variation in $P(x)$, is ϵ_P , which is exogenous
- As if driven by a randomized trial process
- If we further condition on $P(x) = p$, we learn $E[f_Y(p, x, \epsilon_Y)]$



Identification by Regression Revisited

- The conditional expectation function $E[Y|P = p, X = x]$ recovers the conditional average structural response function $E[f_Y(p, x, \epsilon_Y)]$

$$\begin{aligned} E[Y|P = p, X = x] &= E[f_Y(P, X, \epsilon_Y) | P = p, X = x] \\ &= E[f_Y(p, x, \epsilon_Y) | P = p, X = x] \\ &= E[f_Y(p, x, \epsilon_Y)] \end{aligned}$$

- Average structural response = Expected outcome when P, X are exogenously set (outside of the model) to take values (p, x)
- It is useful for generating counterfactual predictions; what “would happen on average if” we intervene and set $(P, X) \leftarrow (p, x)$
- For TSEM: counterfactual predictions \equiv predictions

Identification by Regression Re-stated

- For TSEM, the conditional average structural causal effect coincides with the conditional average predictive effect

$$\begin{aligned} & E[f_Y(p_1, x, \epsilon_Y)] - E[f_Y(p_0, x, \epsilon_Y)] \\ &= E[Y|P = p_1, X = x] - E[Y|P = p_0, X = x] \end{aligned}$$

- Left hand side is a structural hypothetical quantity: what would happen if we intervene and change P from p_0 to p_1 , at $X = x$
- Right hand side is a statistical quantity that can be calculated from observed random variables Y, P, X
- **Identification:** Mapping of “structural hypothetical quantities” to “measurable quantities” from data

Formalizing the Language of Interventions

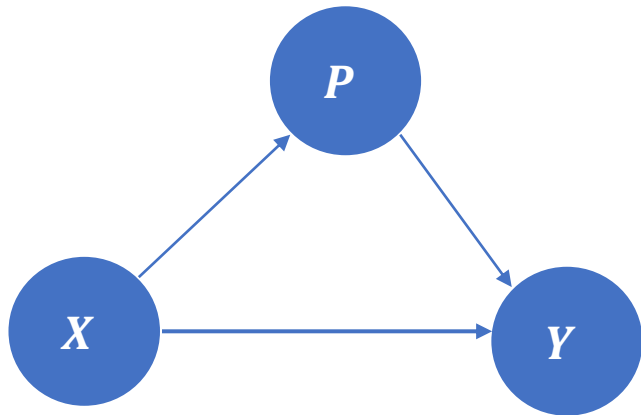
Do Interventions: $do(P = p)$

Original Data Generative Model

$$Y := f_Y(P, X, \epsilon_Y)$$

$$P := f_P(X, \epsilon_P)$$

$$X := \epsilon_X$$

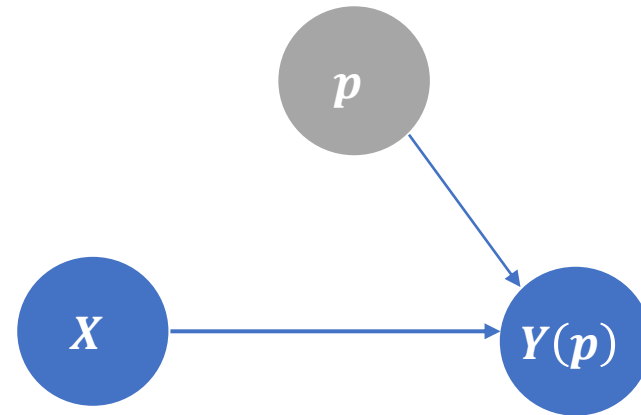


Data Generative Model under $do(P = p)$

$$Y \mid \begin{matrix} P \\ X \end{matrix} \Big|_{do(P = p)} := f_Y(p, X, \epsilon_Y)$$

$$P \mid \begin{matrix} P \\ X \end{matrix} \Big|_{do(P = p)} := p$$

$$X \mid \begin{matrix} P \\ X \end{matrix} \Big|_{do(P = p)} := \epsilon_X$$



Interventions

- Do-interventions is only one way of defining counterfactuals
- We can define any type of counterfactual by simply changing one of the equations to something else
- Wright in his seminal work in '28 defined an intervention where the demand equation was replaced by another one that reflects a tax hike
- We can also define “soft-interventions”: increase price by 10% of its current value
- Another useful variant of do-interventions does not replace the treatment equation are “fix” interventions

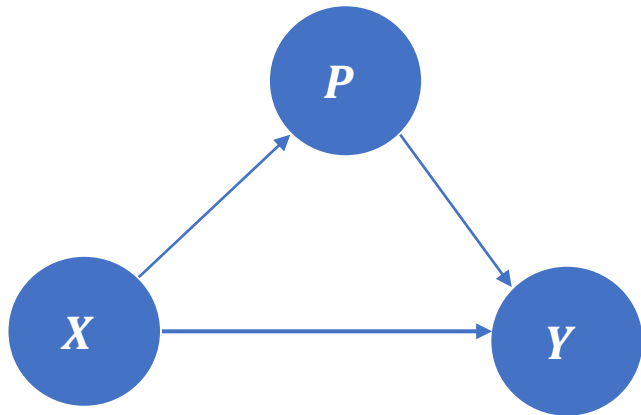
Fix Interventions: $\text{fix}(P = p)$

Original Data Generative Model

$$Y := f_Y(P, X, \epsilon_Y)$$

$$P := f_P(X, \epsilon_P)$$

$$X := \epsilon_X$$

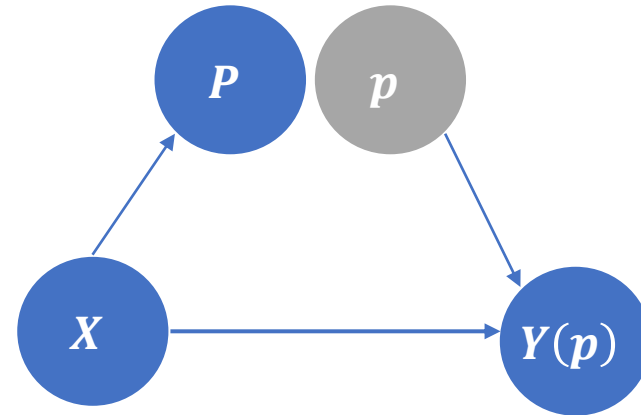


Data Generative Model under $\text{fix}(P = p)$

$$Y \mid \text{fix}(P = p) := f_Y(p, X, \epsilon_Y)$$

$$P \mid \text{fix}(P = p) := f_P(X, \epsilon_P)$$

$$X \mid \text{fix}(P = p) := \epsilon_X$$



Fix Interventions

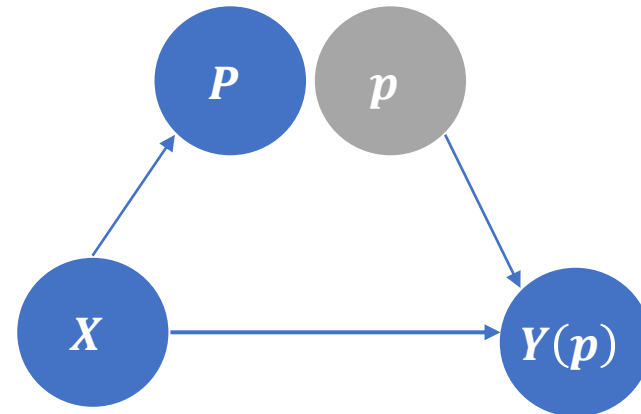
- A fix intervention is a form of “localized” do intervention
- We are only fixing the value of P in the structural equation for Y
- The random variables generated by the fix intervention are the triplets
 $(Y(p), P, X)$
- The intervention does not affect the P, X equations nor the distribution of the exogenous shock ϵ_Y in the outcome equation

Single World Intervention Graphs

- The graphs that represent the generative model under a fix intervention
- Easy to verify visually that
$$Y(p) \perp\!\!\!\perp P \mid X$$
- Then we can do identification based on conditional ignorability

Data Generative Model under $\text{fix}(P = p)$

$$\begin{array}{l|l} Y & := f_Y(p, X, \epsilon_Y) \\ P & \text{fix}(P = p) \quad f_P(X, \epsilon_P) \\ X & := \epsilon_X \end{array}$$



Single World Intervention Graph

Testable Implications of a DAG

D-Separation and Conditional Independence

DAGs Encode Factorization of Probability

- Graph implies factorization of the probability law

$$p(y, d, x, z) = p(y|x, d)p(d|x, z)p(x)p(z)$$

- By repeated application of Bayes rule

$$p(y, d, x, z) = p(y|d, x, z)p(d, x, z)$$

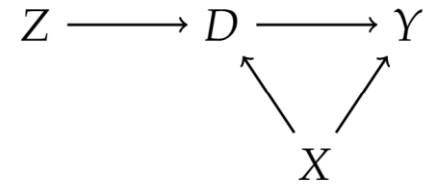
- From graph

$$p(y|d, x, z) = p(y|d, x)$$

- Further Bayes rule

$$p(d, x, z) = p(d|x, z)p(x, z)$$

- From independence: $p(x, z) = p(x)p(z)$



General DAGs

- For any DAG, we can write the ASEM

$$X_j := f_j(\text{Parents}_j, \epsilon_j) = f_j(\text{Pa}_j, \epsilon_j)$$

- Shocks ϵ_j are jointly independent and independent of $\{X_j\}$
- And the corresponding structural response functions

$$X_j(pa_j) := f_j(pa_j, \epsilon_j)$$

- Where pa_j are potential values of the parent nodes that index the stochastic potential outcome processes
- Consistency: variables X_j are generated by generating the shocks and then solving repeatedly the structural response functions

General DAGs and Factorization

- The probability law factorizes as:

$$p(\{x_\ell\}_{\ell \in V}) = \prod_{\ell \in V} p(x_\ell | pa_\ell)$$

DAGs Encode Conditional Independencies

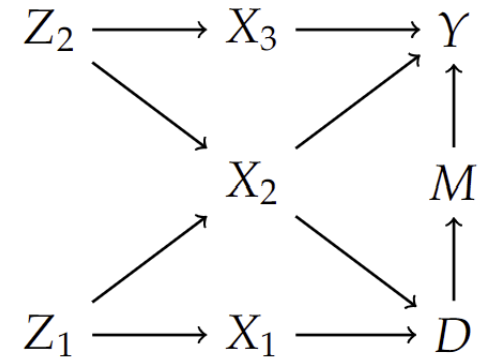
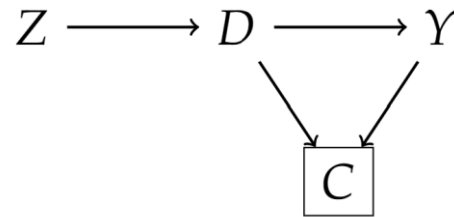
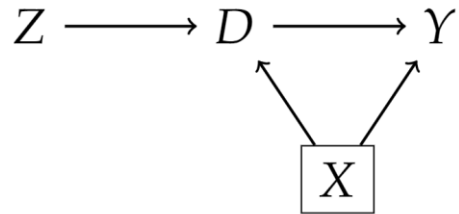
- Any two variables X, Y are independent conditional on a set S if they are D(irected)-separated in the graph

$$(X \perp\!\!\!\perp_d Y \mid S)_G \Rightarrow X \perp\!\!\!\perp Y \mid S$$

- Need to define the concept of D-separation

Some Graph Definitions

- A path π in a graph is blocked by a set of nodes S if
 - Either π contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ and $m \in S$
 - Or π contains a collider $i \rightarrow m \leftarrow j$ and neither m nor its descendants are in S



D-Separation

- In a DAG G , two nodes X, Y are D-separated by a set of nodes S if S blocks all paths between X and Y
- We denote it as:

$$(X \perp\!\!\!\perp_d Y \mid S)_G$$

D-separation
implies
conditional
independency

$$(X \perp\!\!\!\perp_d Y \mid S)_G \Rightarrow X \perp\!\!\!\perp Y \mid S,$$

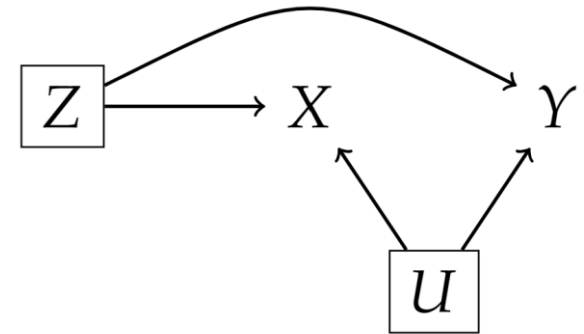
(Verma, Pearl, '88)



Examples

- By factorization property and Bayes rule

$$p(y, x|z, u) = p(y|x, z, u) p(x|z, u) = p(y|z, u) p(x|z, u)$$



Examples

- By factorization property and Bayes rule

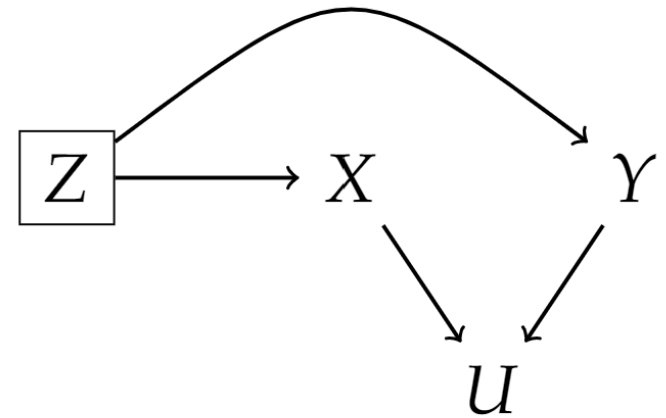
$$p(y, x|z) = p(y|x, z) p(x|z) = p(y|z) p(x|z)$$

- If we had included u , then

$$\begin{aligned} p(y, x|z, u) &= \frac{p(y, x, u|z)}{p(u|z)} = \frac{p(u|x, y, z)p(y|x, z)p(x|z)}{p(u|z)} \\ &= \frac{p(u|x, y)p(y|z)p(x|z)}{p(u|z)} \end{aligned}$$

- We cannot write it as the product of two functions

$$p(y, x|z, u) = f(x, z, u) \cdot g(y, z, u)$$



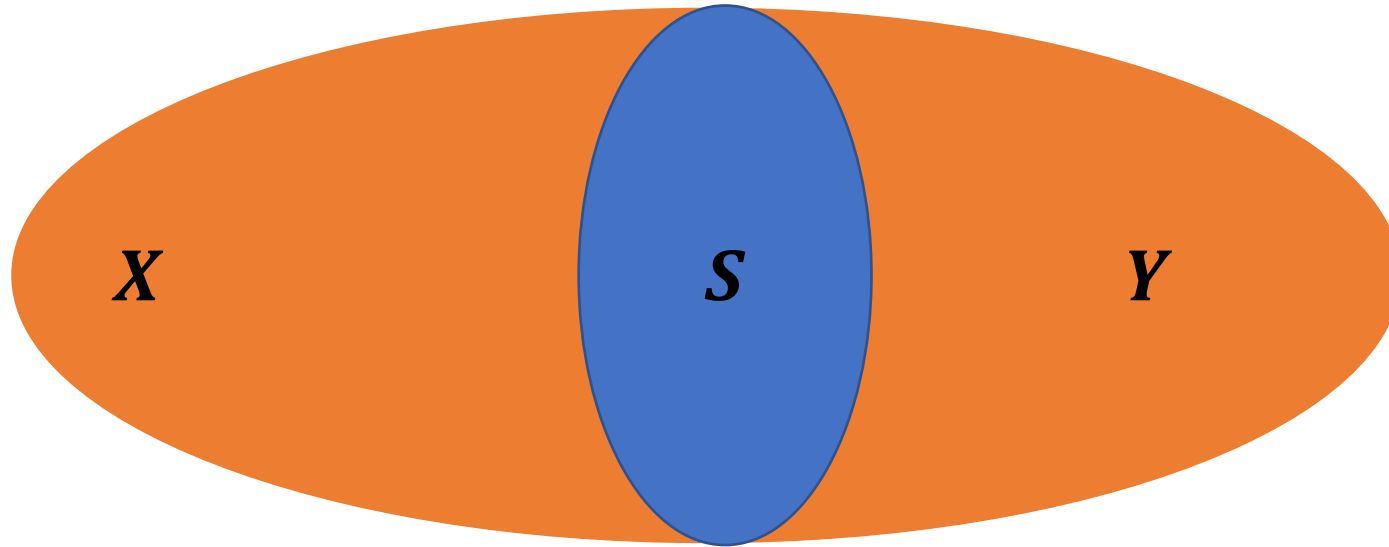
Proving the Main Theorem!

Proof Step 1

- A set of nodes \mathbf{X} is called ancestral if all ancestors of \mathbf{X} are in \mathbf{X}
- Removing all nodes outside of an ancestral set and looking at the resulting graph and ASEM, the probability law is the same as the probability law of \mathbf{X} in the original graph (exercise)

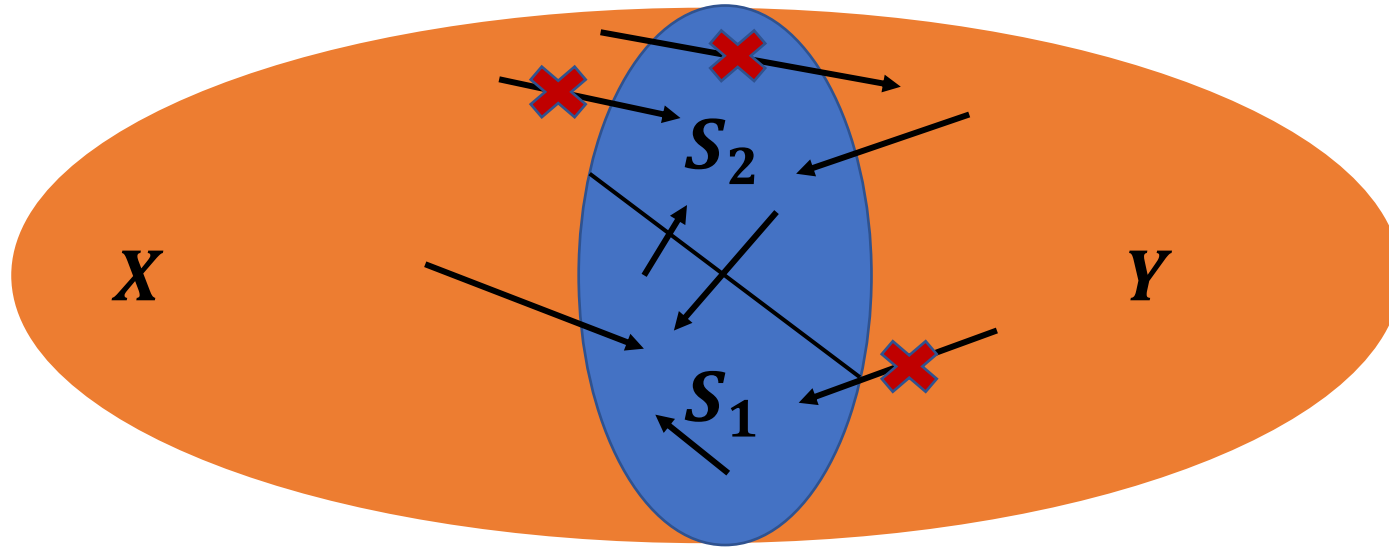
Proof Step 2

- Suppose that a set of nodes \mathbf{X} is D-separated from a set of nodes \mathbf{Y} by a set of nodes \mathbf{S}
- And that $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{S}$ is the set of all nodes



Proof Step 2

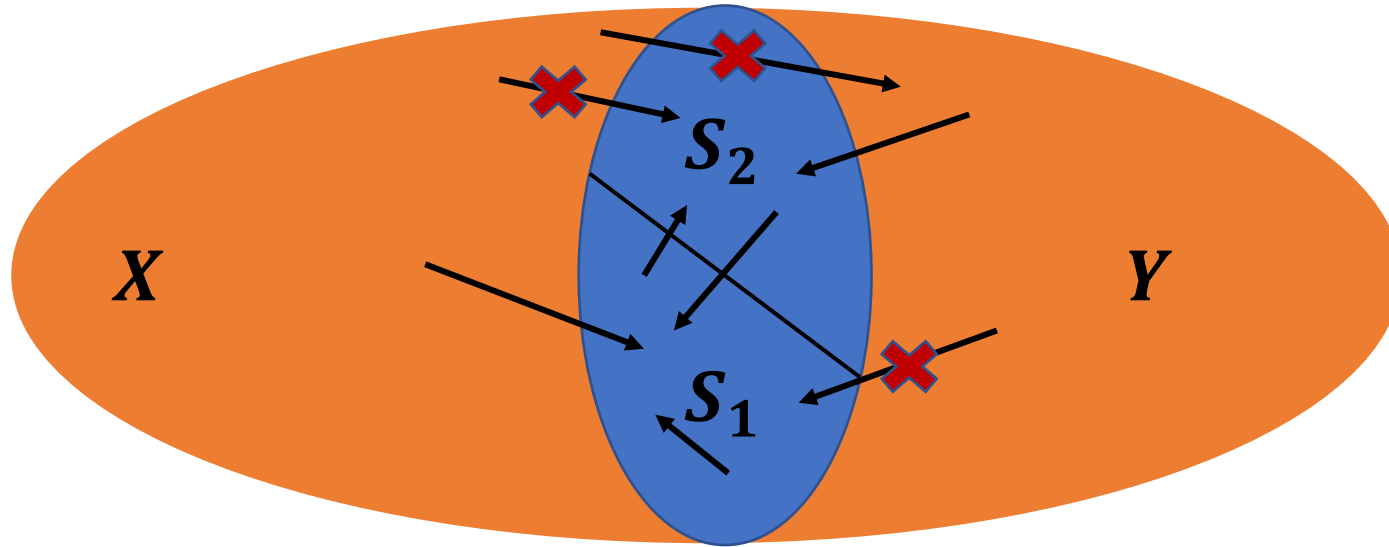
- Let S_1 the subset of S that has a parent in X . Let S_2 the remainder.
- It has to be that $Pa(X \cup S_1) \in X \cup S$
- It has to be that $Pa(Y \cup S_2) \in Y \cup S$



Proof Step 2

- We can factorize:

$$p(x, y, s) = \prod_{W \in X \cup S_1} p(w|pa_w) \prod_{W \in Y \cup S_2} p(w|pa_w) = f(x, s_1)g(y, s_2)$$



Proof Step 2

- We can factorize:

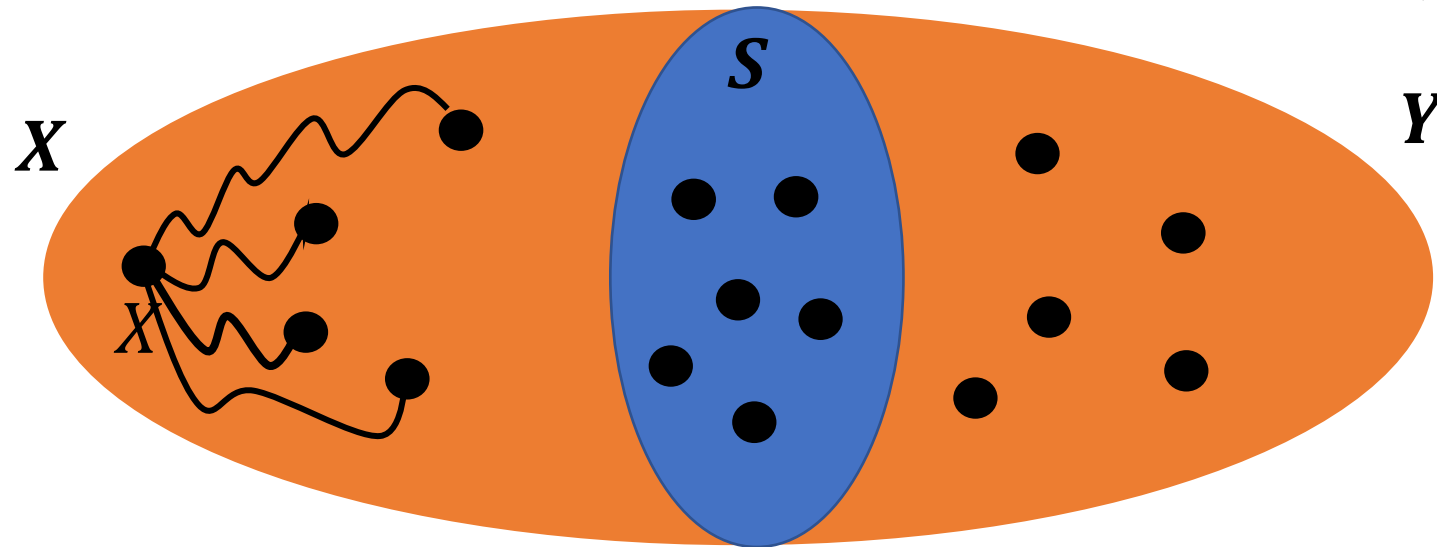
$$p(x, y, s) = \prod_{W \in X \cup S_1} p(w|pa_w) \prod_{W \in Y \cup S_2} p(w|pa_w) = f(x, s_1)g(y, s_2)$$

- Implies that:

$$X \perp\!\!\!\perp Y \mid S$$

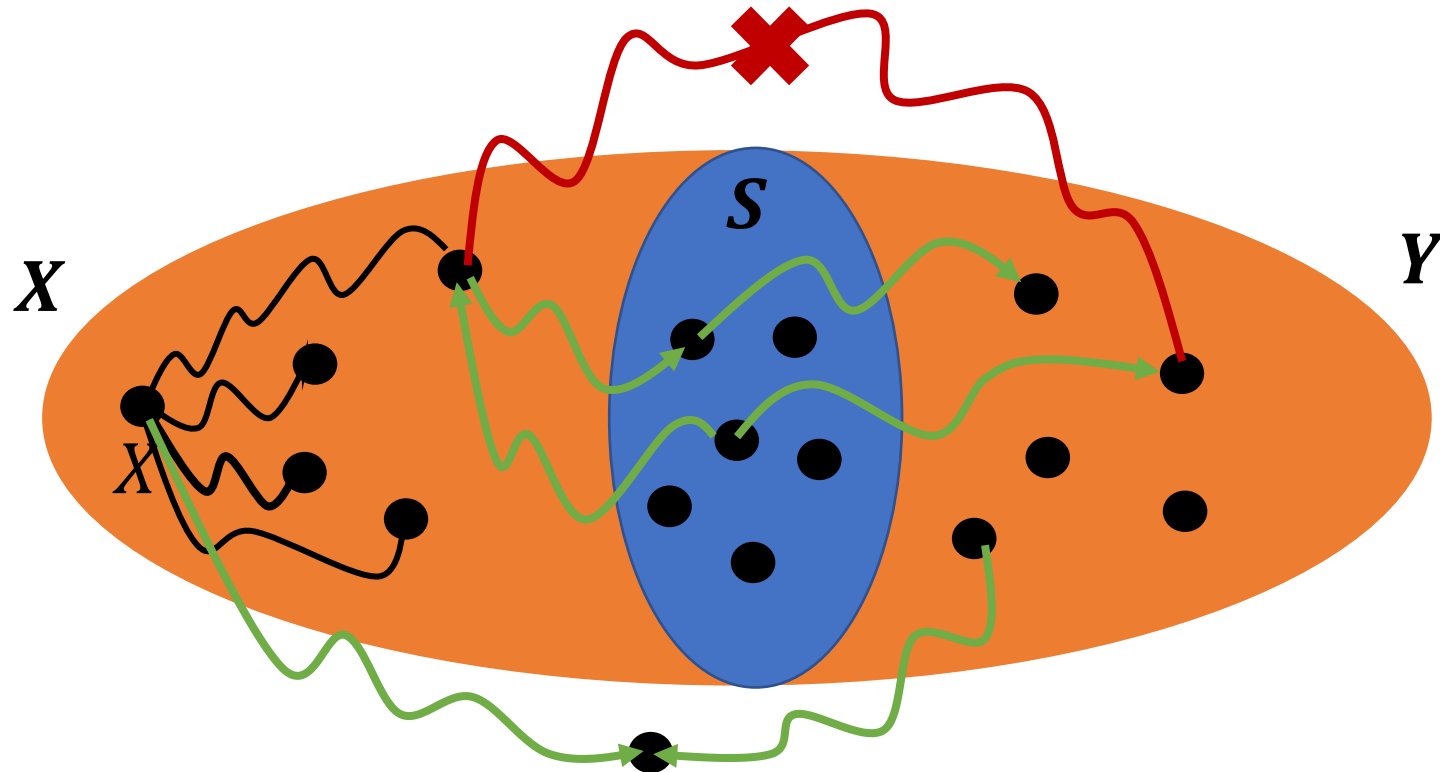
Final Step

- By first step, we can restrict to ancestral set of $X \cup Y \cup S$
- Does not change conditional independence relations (exercise)
- Does not change d-separation relations (exercise)
- Define **X** nodes in ancestral set of $X \cup Y \cup S$ not d-separated from X
- Define **Y** the remainder of nodes in ancestral set not in **X, S** .



Final Step

- By definition of d-separation, S must d-separate X from Y (exercise)
- We can invoke previous critical lemma



Final Step

- By marginalization

$$p(x, y, \mathbf{s}) = \int \int p(x, \mathbf{x}', y, \mathbf{y}', \mathbf{s}) d\mathbf{x}' d\mathbf{y}'$$

- By step 2

$$p(x, y, \mathbf{s}) = \int \int f(x, \mathbf{x}', \mathbf{s}) g(y, \mathbf{y}', \mathbf{s}) d\mathbf{x}' d\mathbf{y}'$$

- We can split integrals

$$p(x, y, \mathbf{s}) = \int f(x, \mathbf{x}', \mathbf{s}) d\mathbf{x}' \int g(y, \mathbf{y}', \mathbf{s}) d\mathbf{y}'$$

- Thus

$$p(x, y, \mathbf{s}) = \bar{f}(x, \mathbf{s}) \bar{g}(y, \mathbf{s}) \Rightarrow X \perp\!\!\!\perp Y \mid S$$