# LEARNING CAUSAL EFFECTS:
## BRIDGING INSTRUMENTS AND BACKDOORS

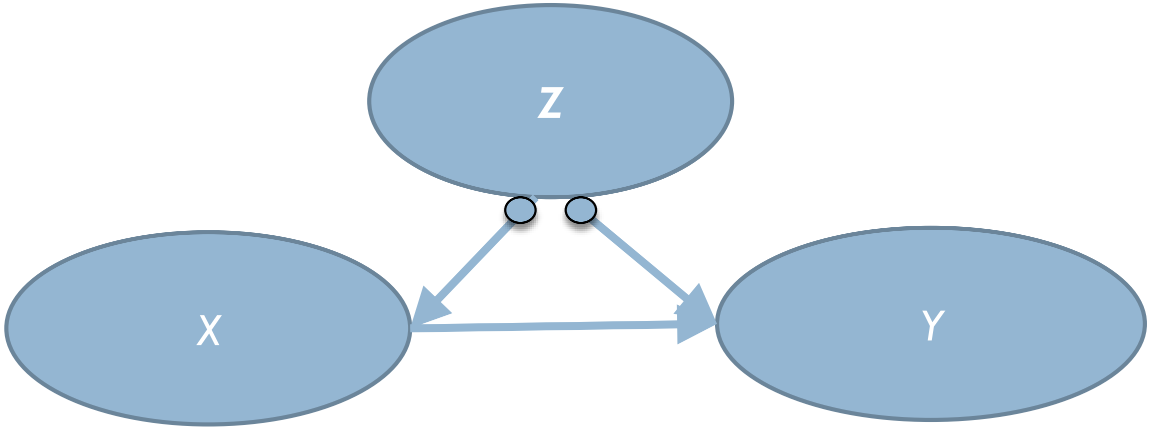
**Ricardo Silva**
Department of Statistical Science and
Centre for Computational Statistics and Machine Learning, UCL
Fellow, Alan Turing Institute
ricardo@stats.ucl.ac.uk

Joint work with Robin Evans (Oxford) and
Shohei Shimizu (Osaka)

Pittsburgh, September 2016

# Goal

- To learn the causal effect of some treatment $X$ on some outcome $Y$ with observational data.
- Assumptions:
  - $Y$ does not precede $X$ causally
  - $X$ and $Y$ do not precede any other covariates measured
  - Variations of faithfulness and parameterizations



# Outline

- We will cover:
  - **The linear case**, where all variables are continuous and all relationships are linear
    - Sets of causal effects can be discovered, sometimes.
    - The role of non-Gaussianity.
  - **The nonlinear discrete case** (binary in particular)
    - The goal is to bound causal effects.
    - The faithfulness continuum.

# Take-home Messages

- The results will rely on different ways of combining backdoor structures and instrumental variables.

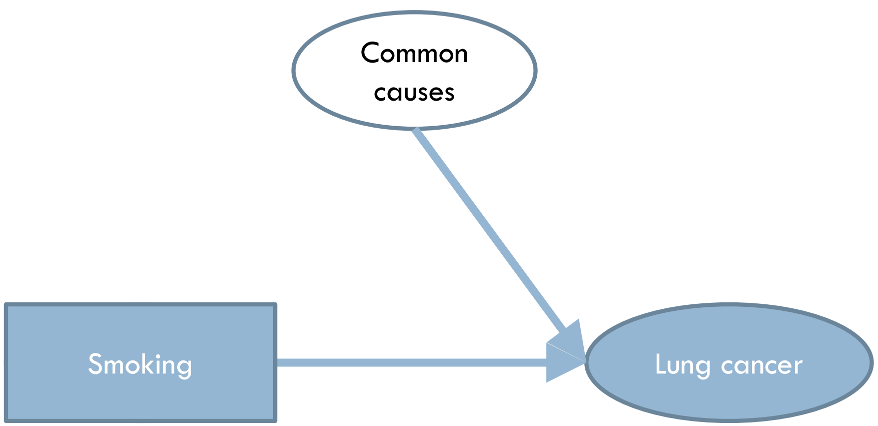
- Discussion points:
  - How to explore redundancies and/or contradictions of assumptions?
  - How to do sensitivity analysis?
  - How to deal with weak associations, both on discovery and control?
  - Please interrupt me at any moment.

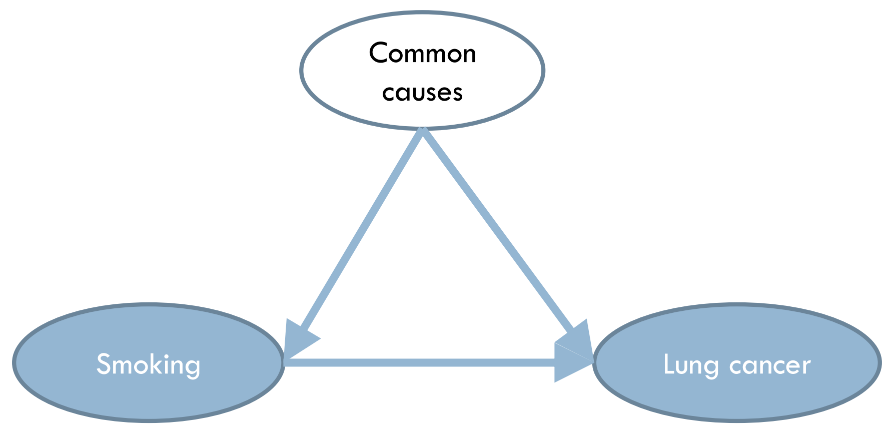
# QUICK BACKGROUND

# Formalizing Observational Studies

We would like to infer P(Outcome | Treatment) in a "world" (regime) like this

All we have is (lousy?) data for P(Outcome | Treatment) in a "world" (regime) like this instead

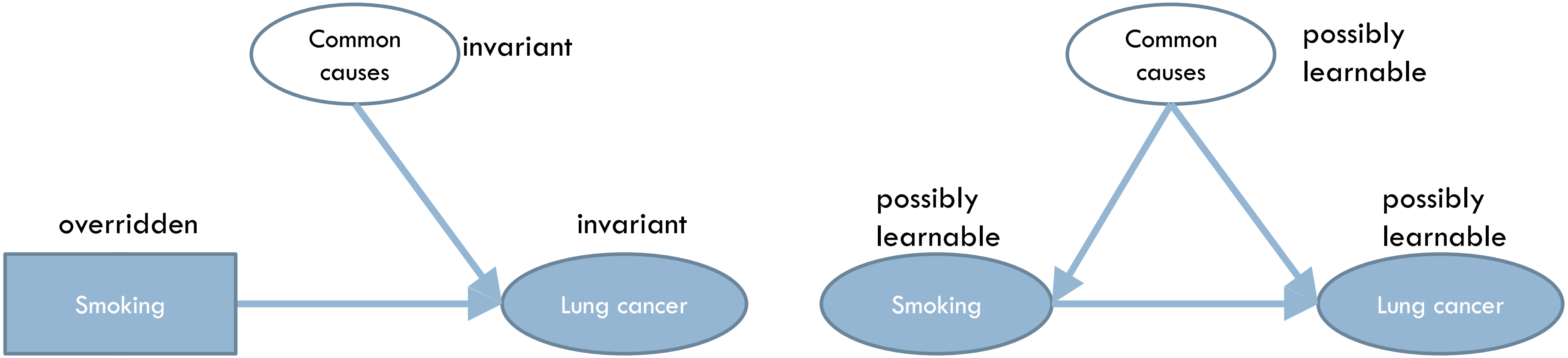



We better make use of an **indexing notation** to distinguish these cases. I will adopt **Pearl's "do" operator.**

# Formalizing Observational Studies

- The jump to causal conclusions from observational data requires assumptions **linking different regimes.**




**Interventional Regime:**
**P(Outcome | do(Treatment))**

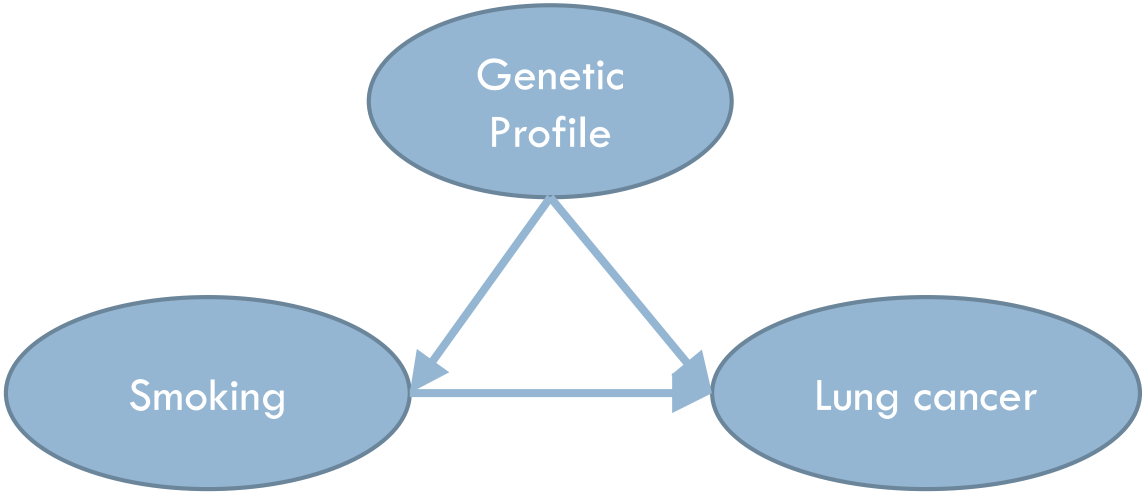
**Observational Regime:**
**P(Outcome | Treatment)**

# General Setup

- In what follows, we will assume we are given a **treatment** variable X, and **outcome** Y, and some **covariates** Z that **precede** X and Y causally.

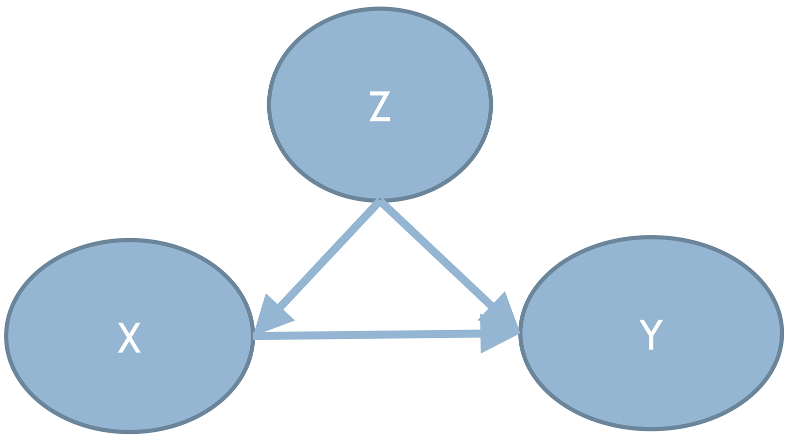
- Unlike the typical graphical model structure learning problem, **we are not interested in reconstructing a full graph. All we care about is** $P(Y \mid do(X = x))$.

# Trick 1: "Adjust" (a.ka., "The Backdoor Adjustment")



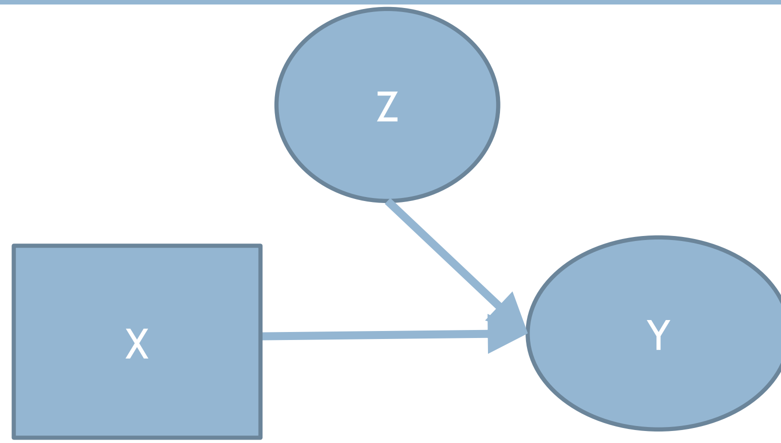
# Why It Works

- Estimand: P(Y | **do(X = x)**), not P(Y | X = x)
- Model:




- Relation to estimand:
  - $P(Y \mid do(x)) = \sum_z P(Y \mid do(x), Z = z)\, P(Z = z \mid do(x))$

# Why It Works

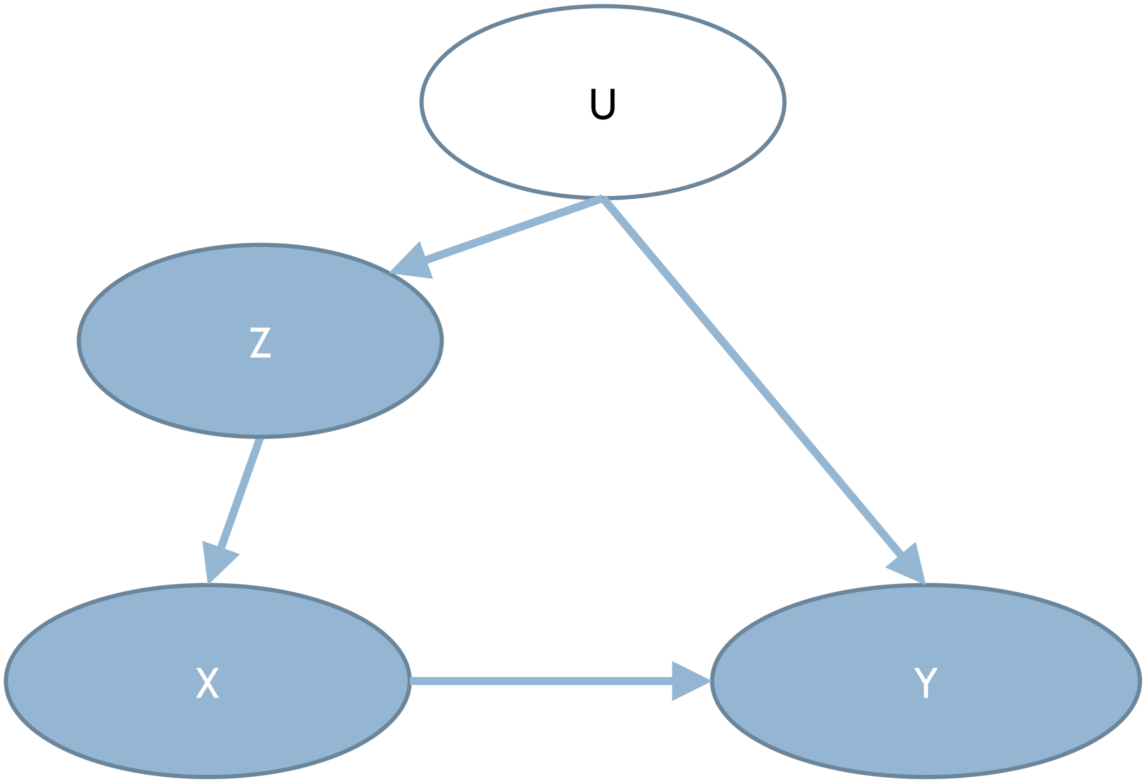



$$P(Y \mid do(x)) = \sum_z P(Y \mid \cancel{do}(x), Z = z)\, P(Z = z \mid \cancel{do}(x))$$

invariance　　　　invariance

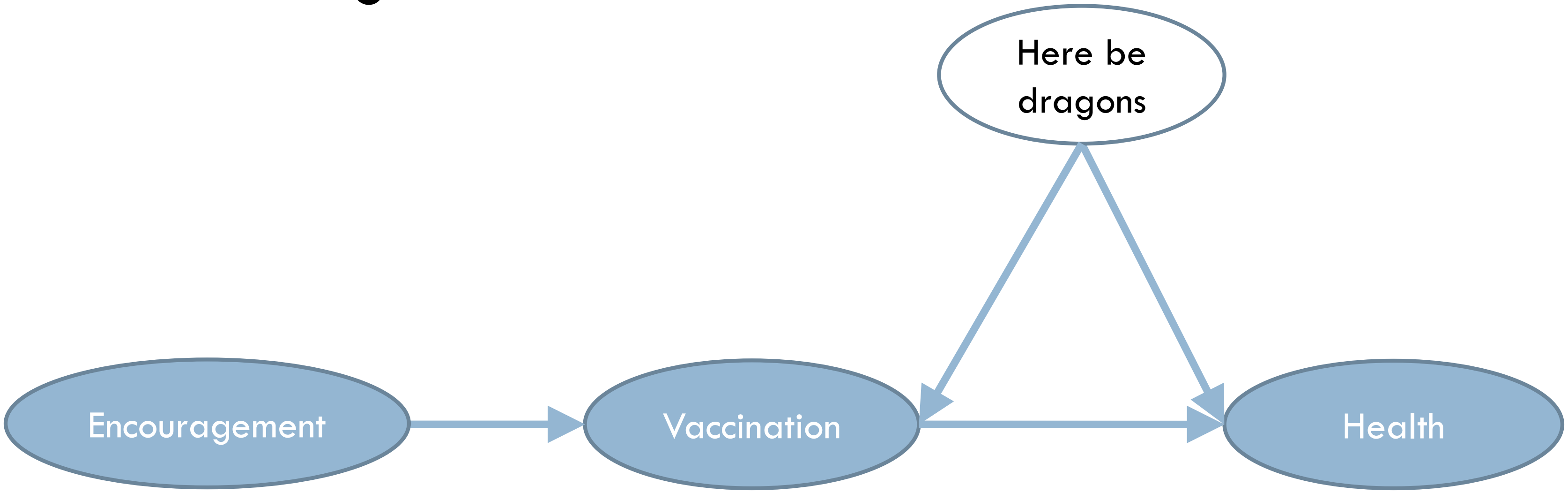
$$= \sum_z P(Y \mid X = x, Z = z)\, P(Z = z)$$

# Note: We don't really need "all" hidden common causes



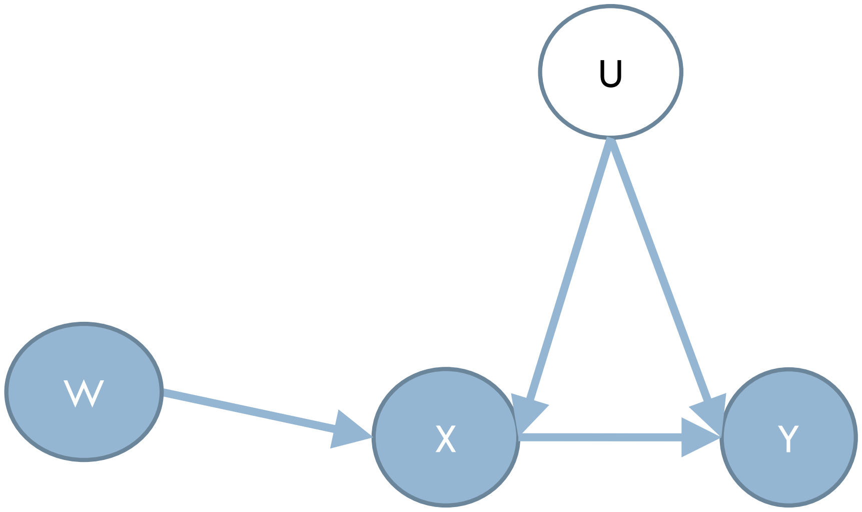
# Trick 2: Instrumental Variables

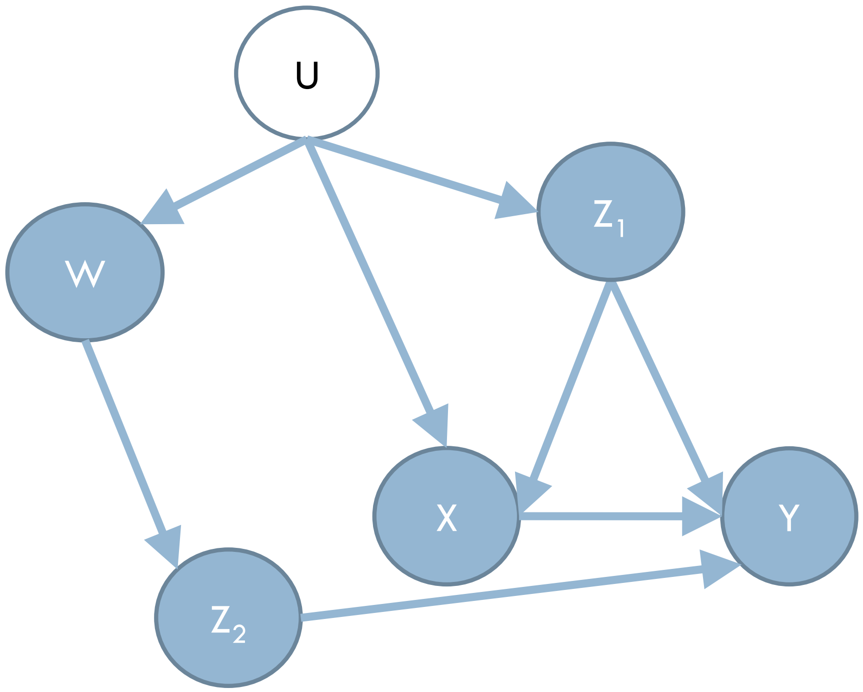
- Variables that can act as "surrogate" experiments.
- Sometimes they *are* surrogate experiments.
- Valuable in the presence of unmeasured confounding.



# (Conditional) Instrumental Variables

- Conditionally, no direct effect, no unblocked confounding with outcome, not affected by treatment.



# Why Do We Care?

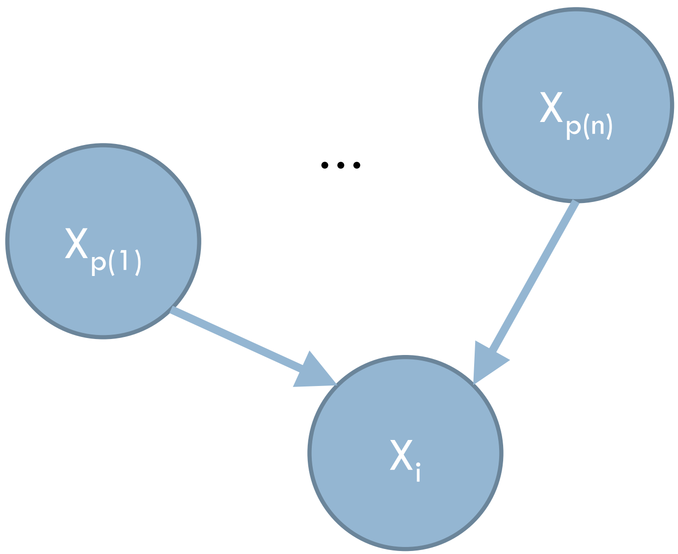
- Instrumental variables **constraint** the distribution of the **hidden common causes.**
- It can be used to infer **bounds** on causal effects or, **under further assumptions, the causal effects** even if hidden common causes are out there.

# THE LINEAR CASE

This is work in progress

# Parametric assumptions

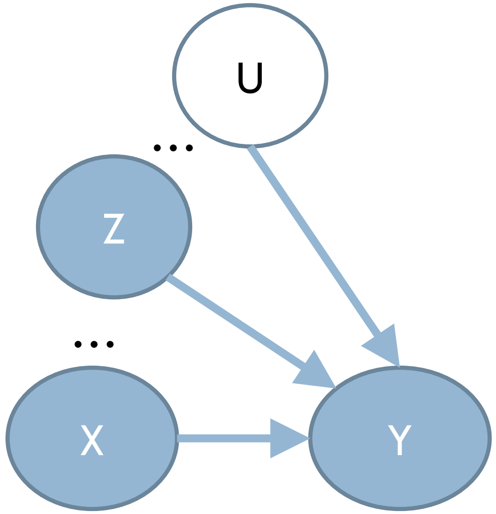
- Assume (causal) acyclic graphical model with linear relationships




$$X_i = \lambda_{i1} X_{p(1)} + \ldots + \lambda_{in} X_{p(n)} + \varepsilon_i$$

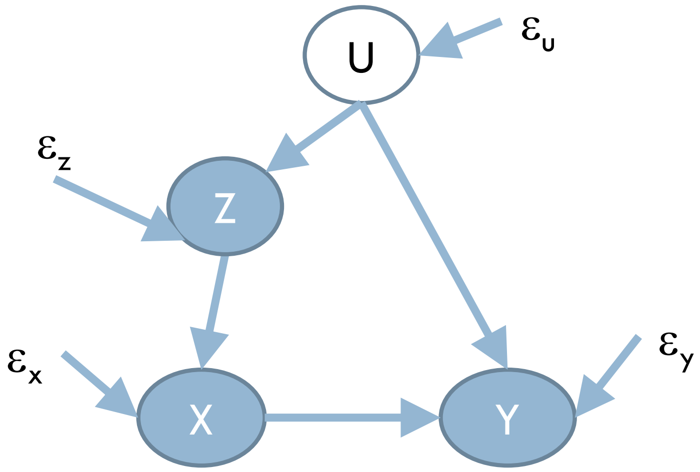
# In Our Setup

- The ultimate goal is to estimate coefficient $\lambda_{yx}$.
- In practice, we will estimate *sets* of plausible values.



# A Test for Back-Door Adjustments

- If error terms are non-Gaussian then **least-square residuals** of treatment and outcome on covariates are independent if and only if there are no unblocked hidden common causes.




$$r_X \equiv X - (X \sim Z)_{l.s.}$$

$$r_Y \equiv Y - (Y \sim X + Z)_{l.s.}$$

$$r_X \perp\!\!\!\perp r_Y$$

This assumes **known** ordering!

Entner et al. (AISTATS, 2012)

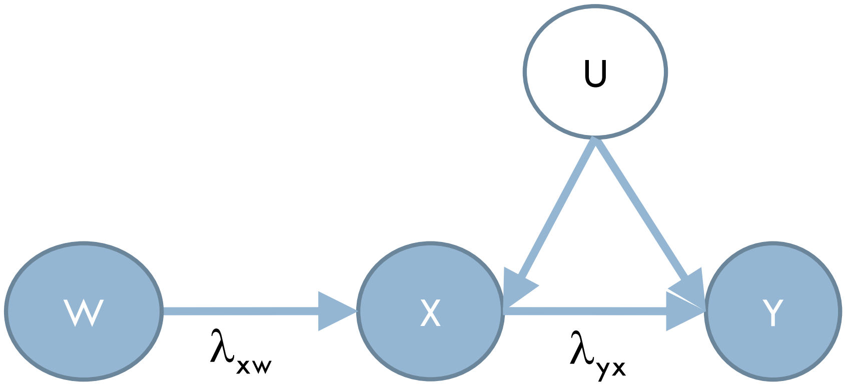
# What If They are Dependent?

- Too bad! Go home empty-handed.

- Instrumental variables, maybe?
  - But how to test them?
  - What if one of my covariates could in principle be an instrumental variable?

# Linear Instrumental Variables

## (or: "All of Econometrics in a Single Slide")




$$\sigma_{wx} = \lambda_{xw}\, \sigma_{ww}$$

$$\sigma_{wy} = \lambda_{xw}\, \lambda_{yx}\, \sigma_{ww}$$

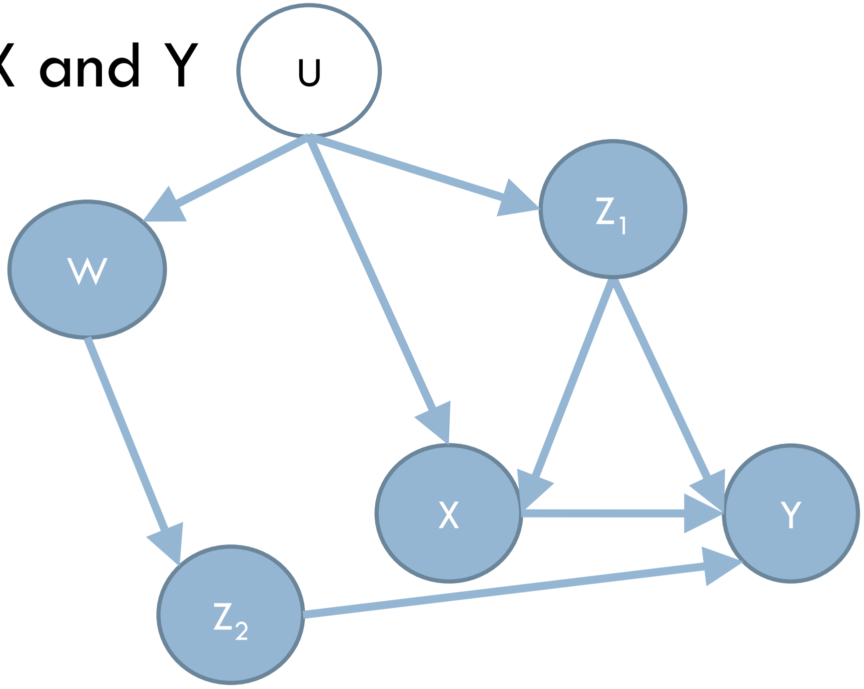
$$\lambda_{yx} = \sigma_{wy} / \sigma_{wx}$$

# IV Discovery

- We would like to discover IVs in the true graph that generated the data, so we could exploit them.
- For that we will focus on a particular graphical characterization of what it means to be an IV.
- We then illustrate why this won't be easy without further assumptions even in linear systems.

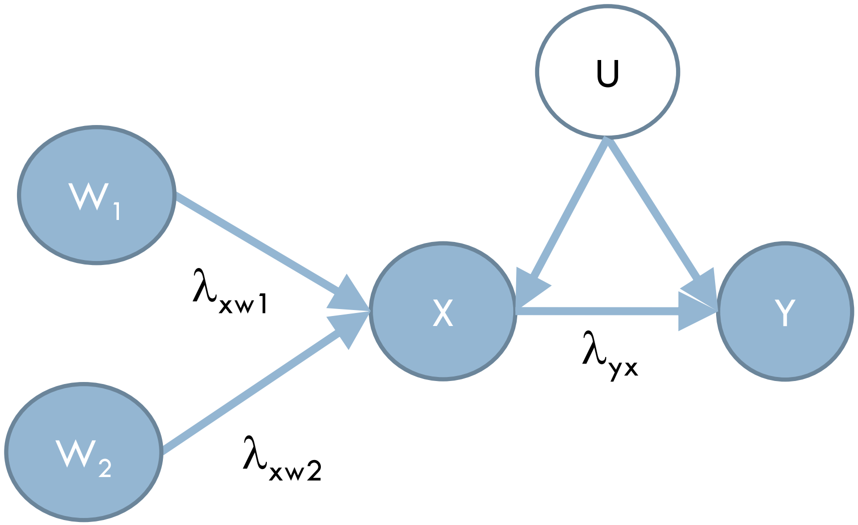
# A Graphical Criteria for Defining IVs

- W is an IV, conditioned on Z, for X ➔ Y if
  1. Z does not d-separate W from X
  2. Z d-separates W from Y in the graph where we remove X ➔ Y
  3. Z are non-descendants of X and Y




Notice how 1 and 3 are "easy to test".

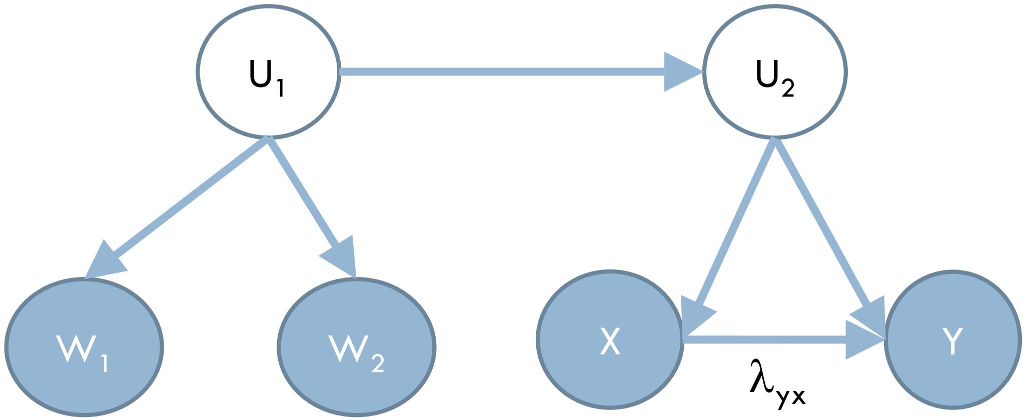
# Falsifying Instrumental Variables




$$\lambda_{yx} = \sigma_{w1y} / \sigma_{w1x} = \sigma_{w2y} / \sigma_{w2x}$$

A tetrad constraint.

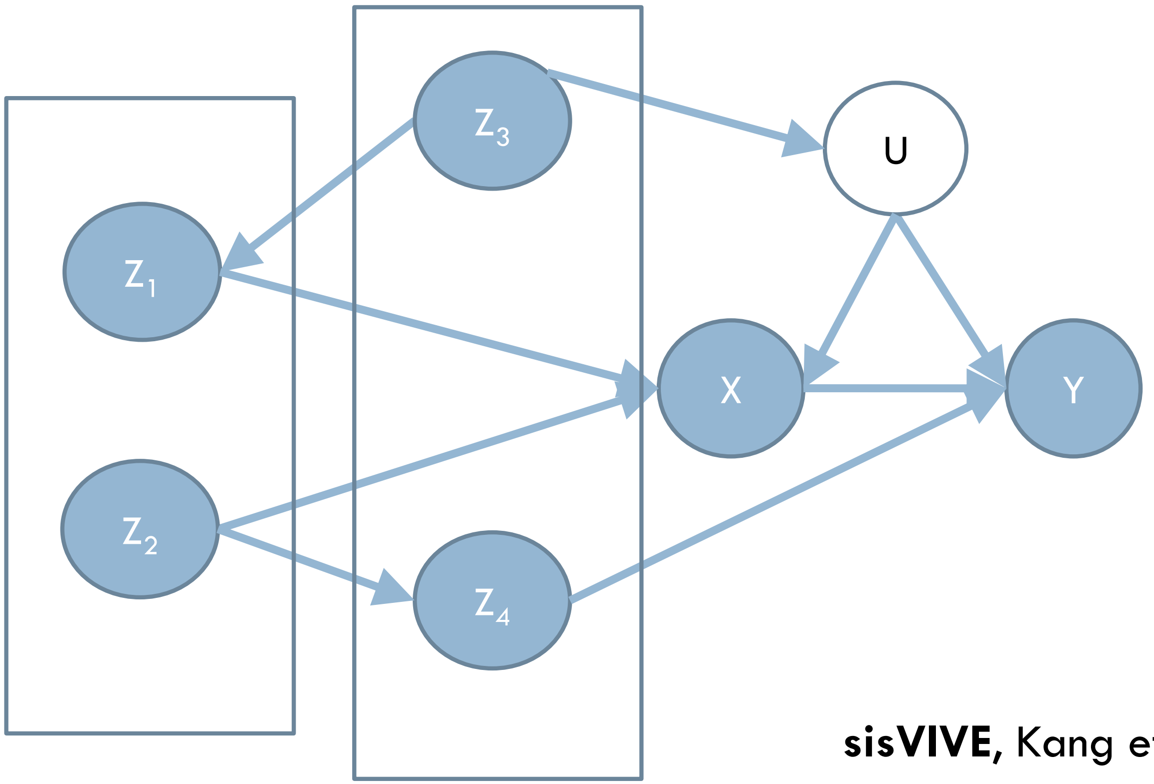
# The Converse Does NOT Hold!




$$\lambda_{yx} \neq \sigma_{w1y} / \sigma_{w1x} = \sigma_{w2y} / \sigma_{w2x}$$

# Strengthening the Assumptions

- Say you split your set Z into two: $Z_V$ and $Z_I$, where $Z_V$ are "valid IVs" given $Z_I$, the possible "invalid" ones.




**sisVIVE**, Kang et al. (JASA, 2015)

# Strengthening the Assumptions

- If in the **true and unknown model** we have **more than half** of Z is valid, we are guaranteed we can use $Z_V$ as instrumental variables (given $Z_I$).

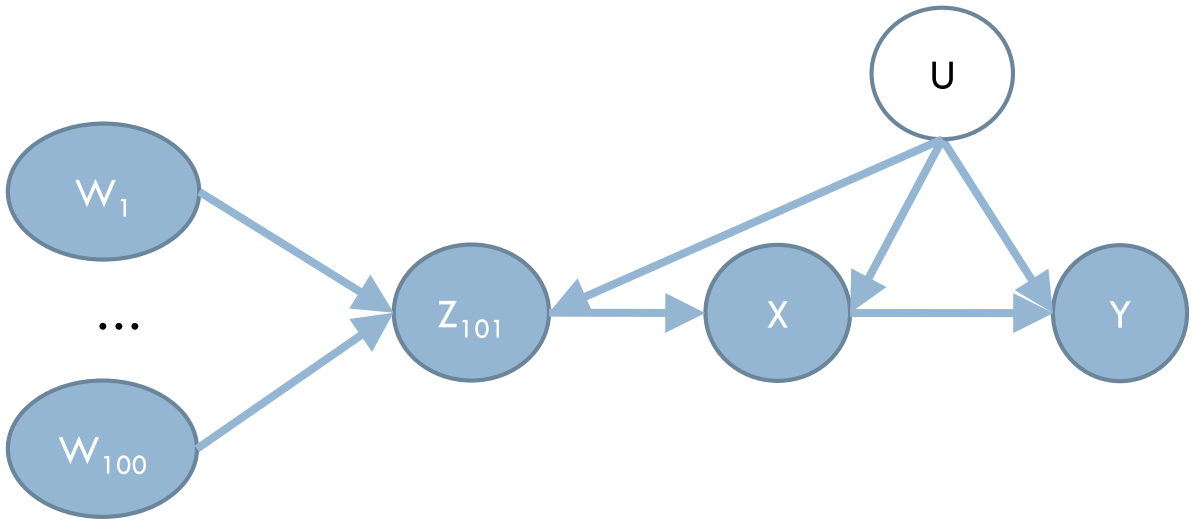
Kang et al. (JASA, 2015)

# An "Equivalent" Algorithm to sisVIVE

**Algorithm 1** IV-BY-MAJORITY$_\infty$

**Input:** set of random variables $\mathbf{V} \cup \{X, Y\}$
**Output:** the causal effect of $X$ on $Y$, or a value (NA) indicating lack of knowledge
**for** each $W_i \in \mathbf{V}$ **do**
    $\mathbf{Z}_i \leftarrow \mathbf{V} \backslash \{W_i\}$
    $\beta_i \leftarrow \sigma_{w_i y . \mathbf{z}_i} / \sigma_{w_i x . \mathbf{z}_i}$
**end for**
**if** more than half of set $\{\beta_i\}$ is equal to the same value $\beta$ **then**
    **return** $\beta$
**end if**
**return** NA

# Still Strong, Sometimes Too Strong




- All of $W_1$, …, $W_{100}$ are valid IVs, *if* we don't condition on $Z_{101}$
- But sisVIVE requires a variable is either an IV or a conditioning variable…

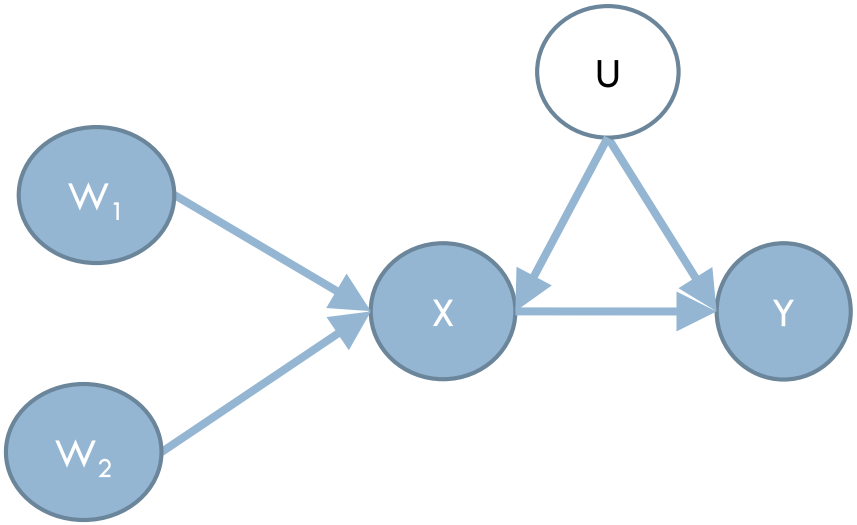
# Alternative: TETRAD-IV

**Algorithm 2** TETRAD-IV$_\infty$

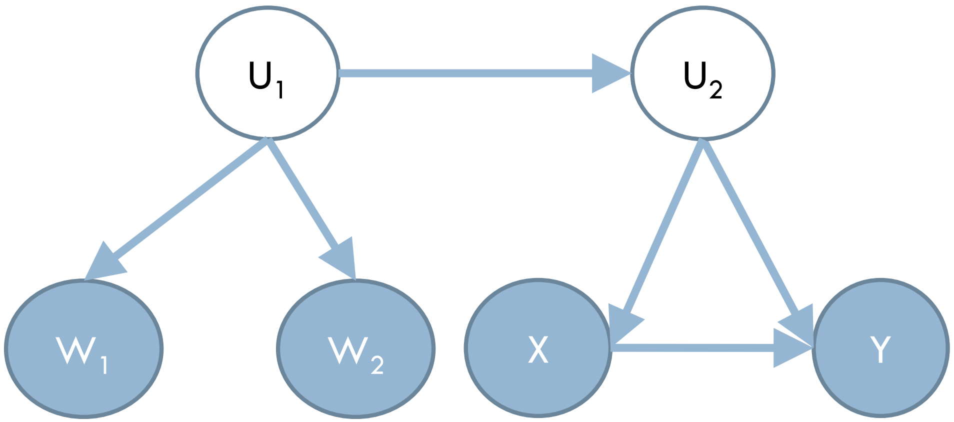
**Input:** set of random variables $\mathbf{V} \cup \{X, Y\}$
**Output:** $\mathcal{C}$, a set of candidate differential causal effects of $X$ on $Y$
Initialize $\mathcal{C} \leftarrow \emptyset$
**for** each pair $\{W_i, W_j\} \subseteq \mathbf{V}$ **do**
    **for** every set $\mathbf{Z} \subseteq \mathbf{V} \backslash \{W_i, W_j\}$ **do**
        **if** $\sigma_{w_i x.\mathbf{z}} = 0$ **or** $\sigma_{w_j x.\mathbf{z}} = 0$ **then**
            **next**
        **end if**
        **if** $\sigma_{w_i x.\mathbf{z}} \sigma_{w_j y.\mathbf{z}} \neq \sigma_{w_i y.\mathbf{z}} \sigma_{w_j x.\mathbf{z}}$ **then**
            **next**
        **end if**
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{\sigma_{w_i y.\mathbf{z}} / \sigma_{w_i x.\mathbf{z}}\}$
    **end for**
**end for**
**return** $\mathcal{C}$

# Interpretation

- What is the graphical converse of the tetrad constraint?
  - Known: the Tetrad Representation Theorem, via the notion of "**choke point**".




X is a choke point for $\{W_1, W_2\} \times \{X, Y\}$



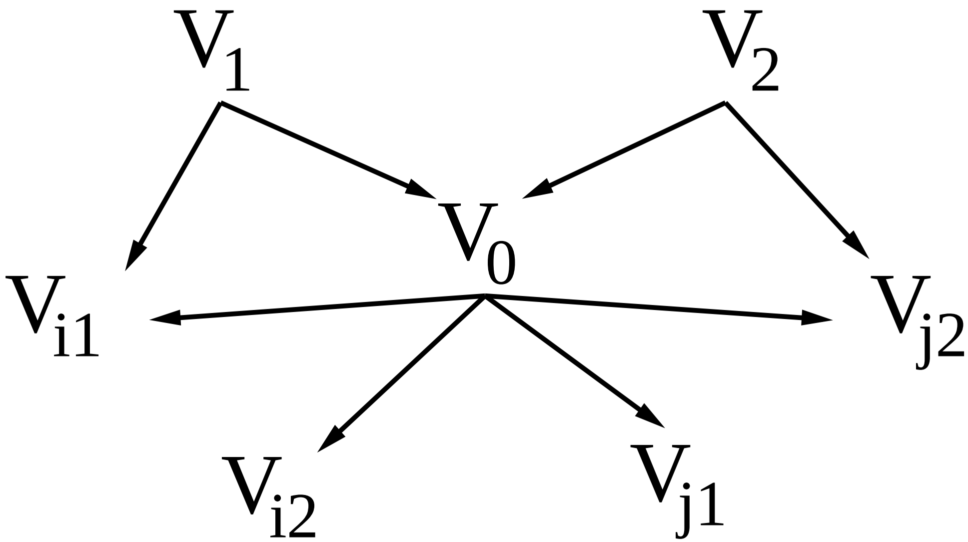

$U_1$ is a choke point for $\{W_1, W_2\} \times \{X, Y\}$

# Interpretation

- What is the graphical converse of the **conditional** tetrad constraint?
  - Cannot appeal to the known result anymore: DAGs are not closed under conditioning.
  - Instead, *re-interpret* a more recent result by Sullivant et al. (Annals of Stats, 2010)

# Sullivant et al.'s Trek Separation

- Cross-covariance of two sets **A** and **B** will drop rank if "small enough" sets "t-separate" **A** from **B.**




Here, $V_0$ "t-separates" $\{V_{i1}, V_{i2}, V_0\}$ from $\{V_{j1}, V_{j2}, V_0\}$

The rank of cross-covariance of these two sets will be (typically) 2.

# Conditional Tetrad Constraint Interpretation

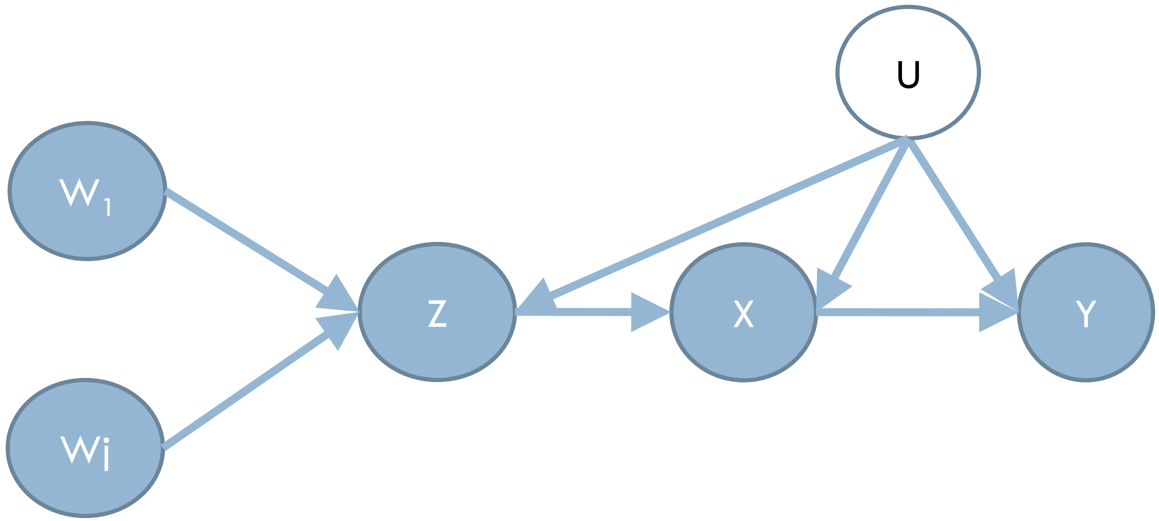
- If $\sigma_{ix.\mathbf{z}}\,\sigma_{jy.\mathbf{z}} = \sigma_{iy.\mathbf{z}}\,\sigma_{jx.\mathbf{z}}$, there will be a set that includes **Z** that t-separates $\{W_i, W_j, \mathbf{Z}\}$ from $\{X, Y, \mathbf{Z}\}$.

$$|\Sigma_{\{Wi,\ Wj,\ \mathbf{Z}\},\ \{X,\ Y,\ \mathbf{Z}\}}| = |\Sigma_{\mathbf{ZZ}}|\,|\sigma_{ix.\mathbf{z}}\,\sigma_{jy.\mathbf{z}} - \sigma_{iy.\mathbf{z}}\,\sigma_{jx.\mathbf{z}}|$$

- This is a necessary but not sufficient condition to guarantee Criterion 2:
  - "Z d-separates W from Y in the graph where we remove X → Y"

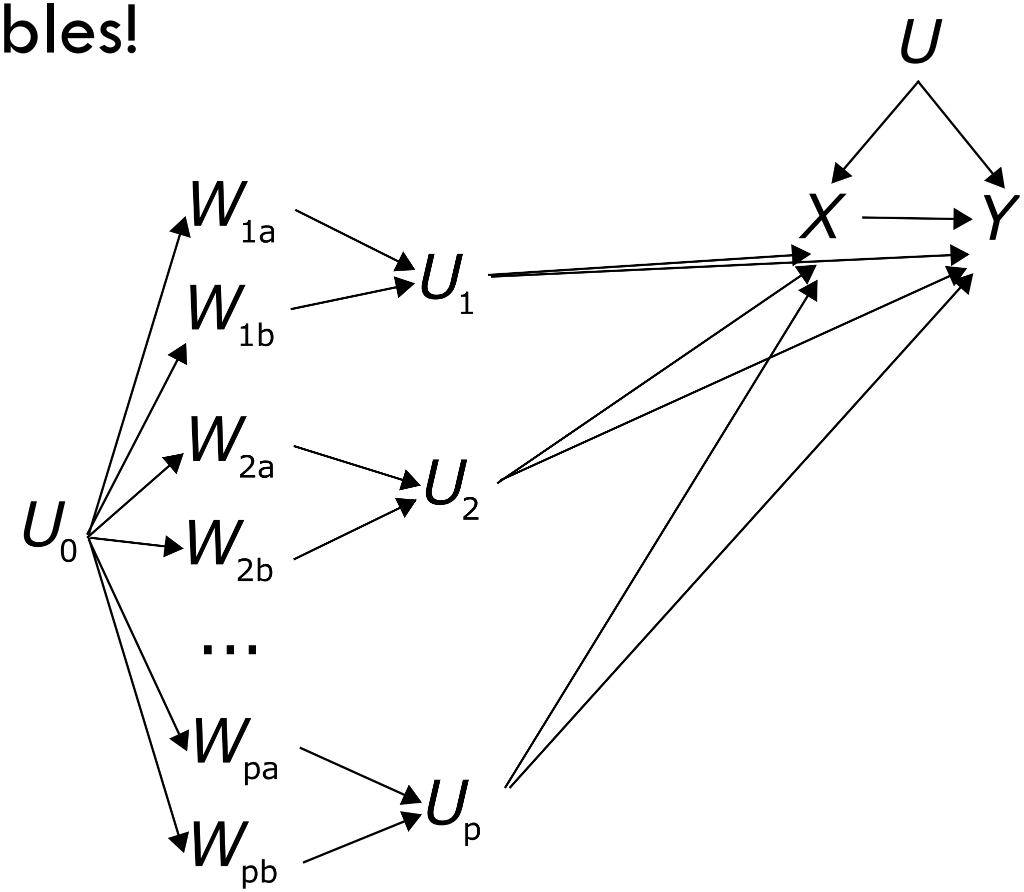
# Tetrad Equivalence Class

- Each TETRAD-IV output can be explained by these "choke sets". If they differ, it is because of
  - a latent element in this choke set (choke set is Z and "$U_z$", instead of Z and X), which links "IVs" to Y
  - a rogue non-directed path activated by conditioning



# Tetrad Equivalence Class

- Size can increase linearly with the number of variables!



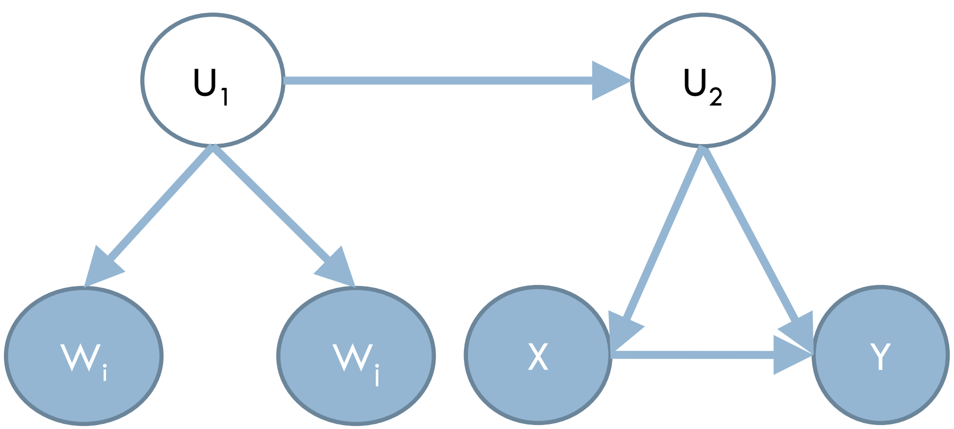
# Tetrad Equivalence Classes

- If there is at least one genuine pair of conditional IVs in the solution, then the output set provides upper and lower bounds on causal effect.
  - This is a much weaker assumption than the one in sisVIVE.

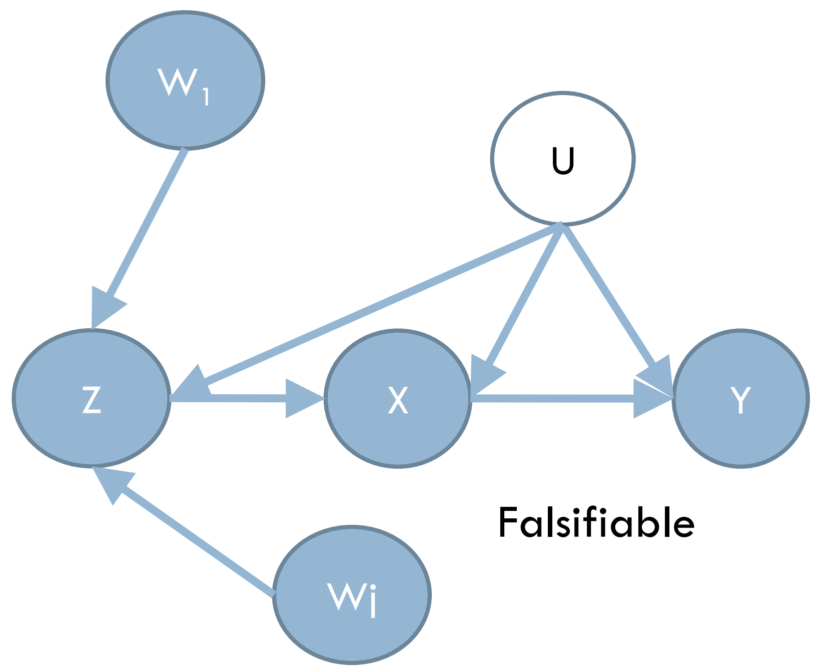
- Also: <INCLUDE FAVOURITE PET IDENTIFIYING ASSUMPTION HERE>
  - “Largest set wins”
  - “Strongest association wins”
  - “Exclude implausibly large effects”
  - “Most common sign wins”
  - Etc.

# Non-Gaussianity

- We can generalize the main result of Entner et al. (2012), and exclude solutions that are due to **non-directed active paths** by a testable condition.



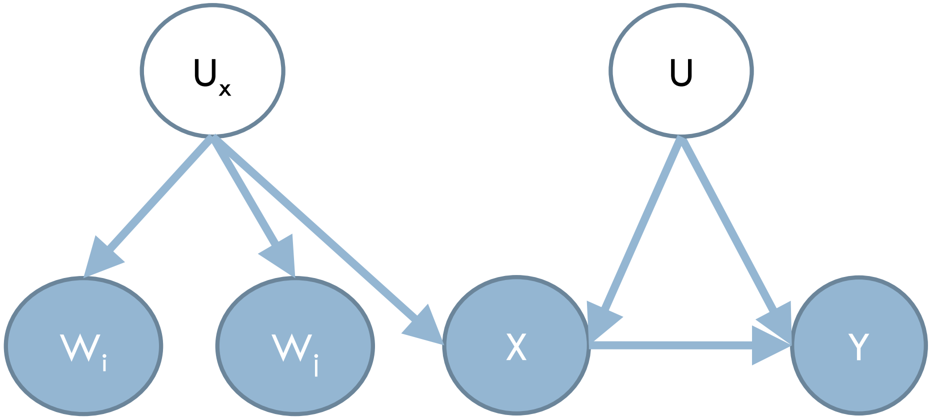

Falsifiable



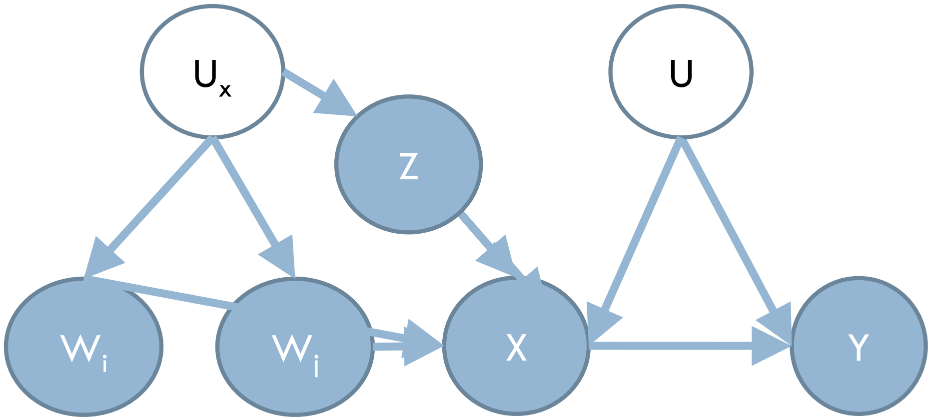

Falsifiable

# Backdoor vs IV trade-off

- Unfortunately, this also excludes some genuine IVs.
- Those will not be excluded if backdoors *with treatment X* are blocked.




Rejected




Preserved

# Empirical Results

- This is work in progress.
- Practical implementation does not use tests of tetrad constraints: much of the signal is weak, tests perform horribly.
  - Without going in details, it clusters empirical estimates of causal effects, assumes a minimal number of IVs.
- Practical implementation does not do combinatorial search on Z: again too much error. Instead, an all-or-nothing is suggested: discard solutions that fail the non-Gaussianity tests.
- It does well in sample sizes relatively large, and seems to be comfortably better than sisVIVE when its assumptions fail. Non-Gaussianity tests require very large sample sizes though.
- Contact me for current manuscript (soon to be re-arXived)

# THE NON-LINEAR DISCRETE (BINARY) CASE

NIPS, 2014; JMLR, 2016

# The Problem

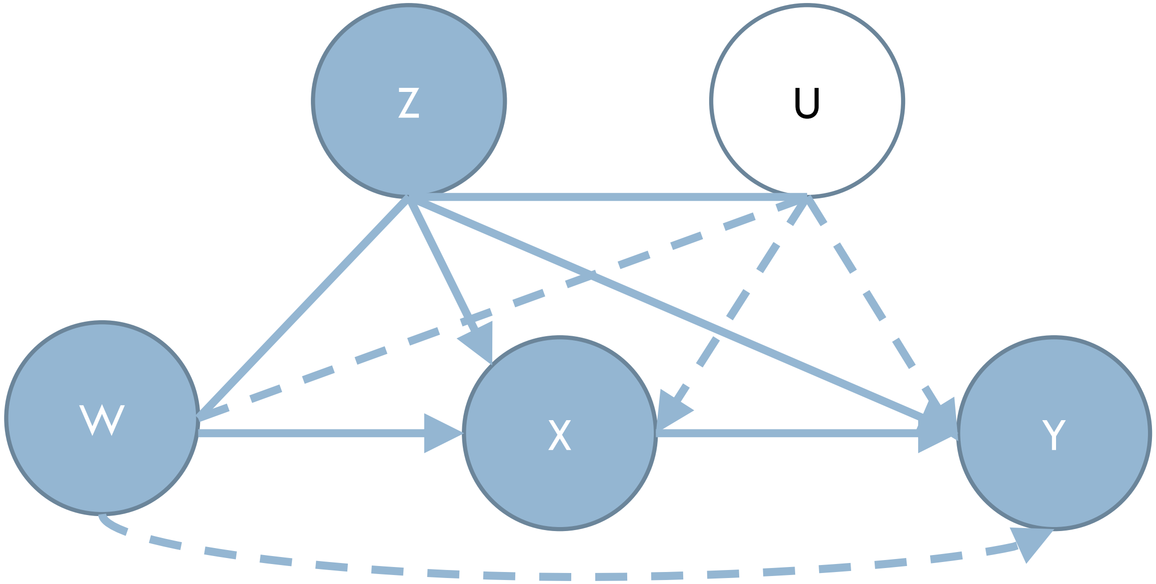
- Given binary X precedes binary Y causally, estimate average causal effect (ACE) **using observational data**


$$ACE \equiv E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)] =$$

$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0))$$

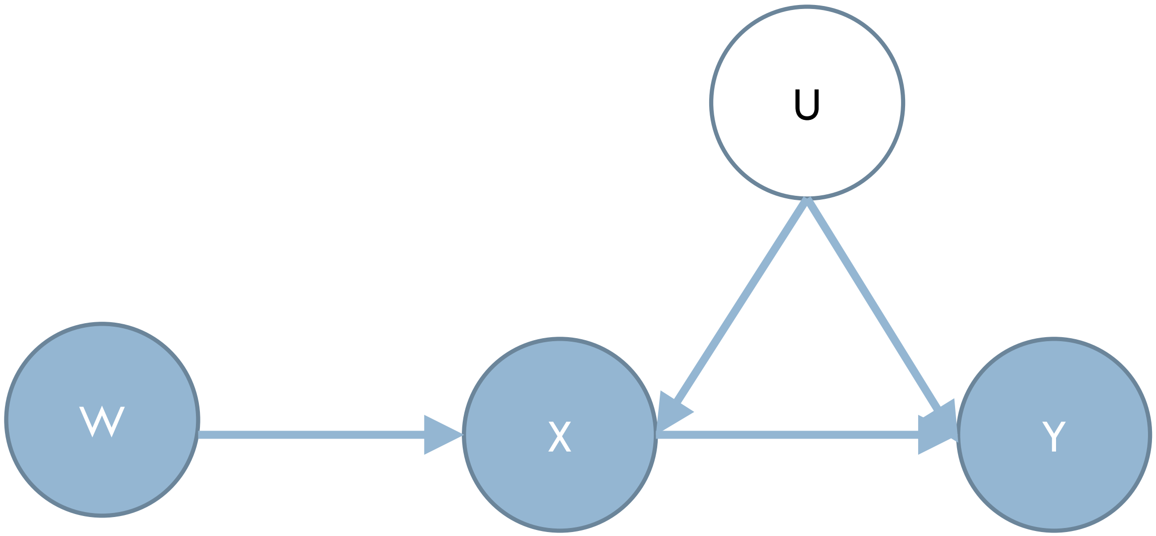
# Goal

- To get an estimate of **bounds** of the ACE
- Rely on the identification of an auxiliary variable W (**witness**), an auxiliary set Z (**background set**), and **assumptions about strength of dependencies** on latent variables



# Instrumental Variables in Discrete Systems

- But where do the missing edges come from?



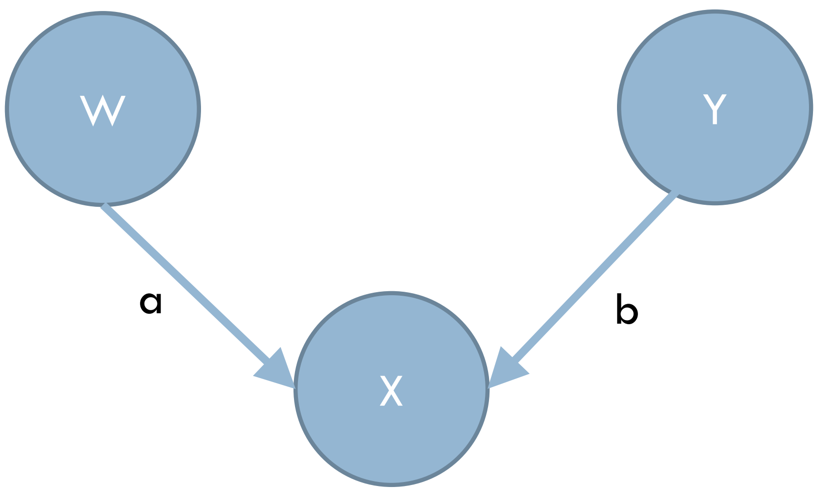

$$L_{P(Y,\, X \mid W)} \leq ACE \leq U_{P(Y,\, X \mid W)}$$

# Exploiting Independence Constraints

- **Faithfulness** provides a way of sometimes finding a point estimator
  - Faithfulness means independence in probability iif "structural" independence (Spirtes et al., 1993)

# Faithfulness

- **W independent of Y, but not when given X:** conclude the following (absentia hidden common causes)



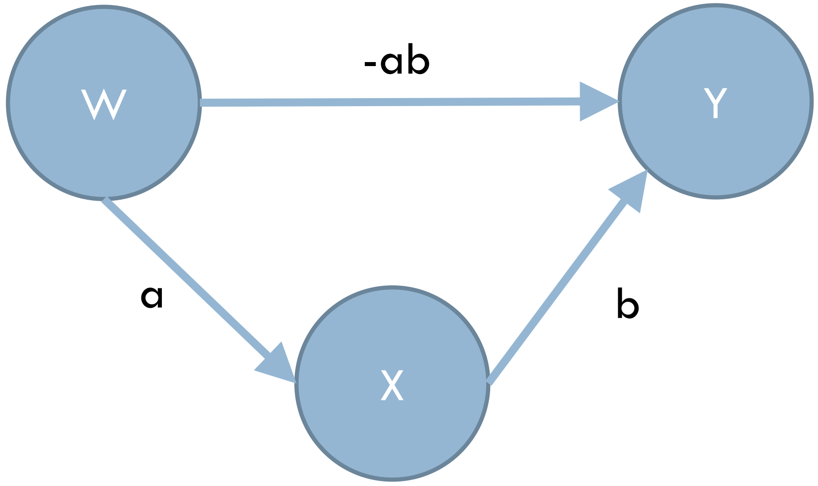

$$X = aW + bY + e_x$$

$$P(W, X, Y) = P(W)P(Y)P(X \mid W, Y)$$

$$P(W, Y \mid X) \propto P(W)P(Y)P(X \mid W, Y)$$

# (Lack of) Faithfulness

- W independent of Y, but not when given X: different structure



# The Problem with Naïve Back-Door Adjustment

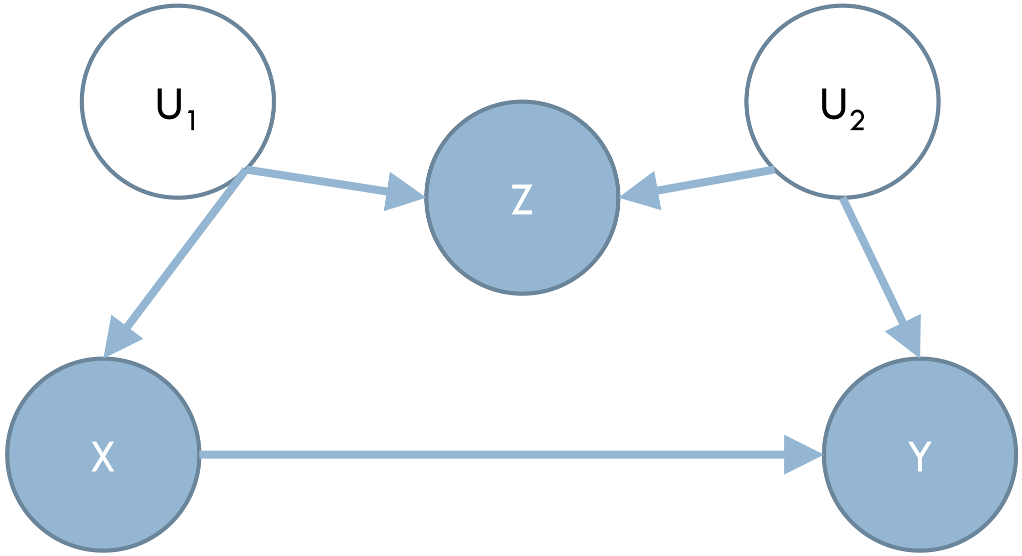
- It is not uncommon in applied sciences to posit that, given a large number of covariates Z that are plausible common causes of X and Y, we should adjust for all

$$P_{est}(Y = 1 \mid do(X = x)) = \sum_{z} P(Y = 1 \mid x, z)P(z)$$

- Even if there are remaining unmeasured confounders, a common assumption is that adding elements of Z will in general decrease bias $|ACE_{true} - ACE_{hat}|$

# The Problem with Naïve Back-Door Adjustment

- Example of failure:

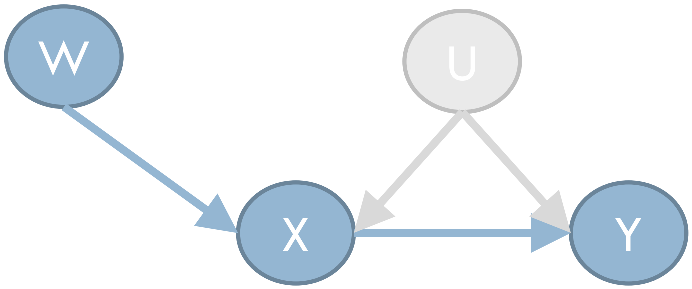



$$P(Y = 1 \mid do(X = x)) = P(Y = 1 \mid X = x) \neq \sum_{z} P(Y = 1 \mid x, z)P(z)$$

Pearl (2009). Technical Report R-348

# Exploiting Faithfulness: A Very Simple Example

- $W \not\perp\!\!\!\perp Y$, $W \perp\!\!\!\perp Y \mid X$ + Faithfulness. Conclusion?




No unmeasured confounding

- Naïve estimator vindicated:
  $ACE = P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$
- This super-simple nugget of causal information has found some practical uses on large-scale problems

# Entner et al.'s Background Finder

- Entner, Hoyer and Spirtes (2013) AISTATS: two simple rules based on finding a **witness** W for a correct **admissible background set** Z
  - Generalizes "chain models" W $\rightarrow$ X $\rightarrow$ Y

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and

(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

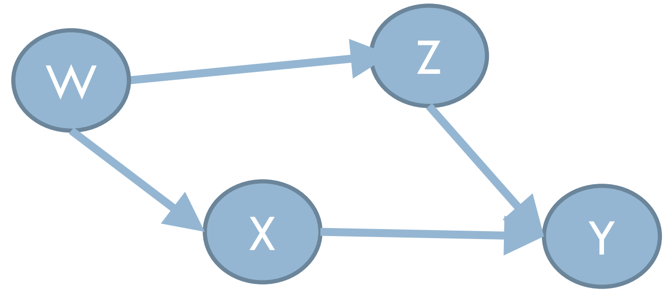
then infer '±' and give $\mathcal{Z}$ as an admissible set.

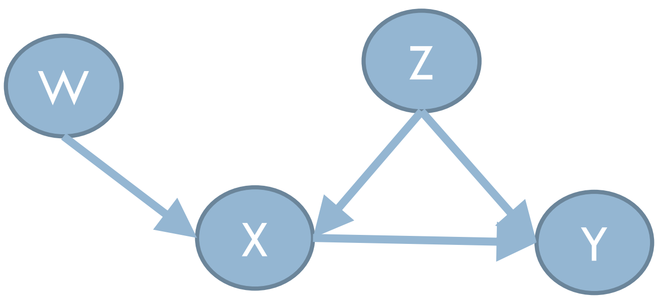
# Rule 1: Illustration

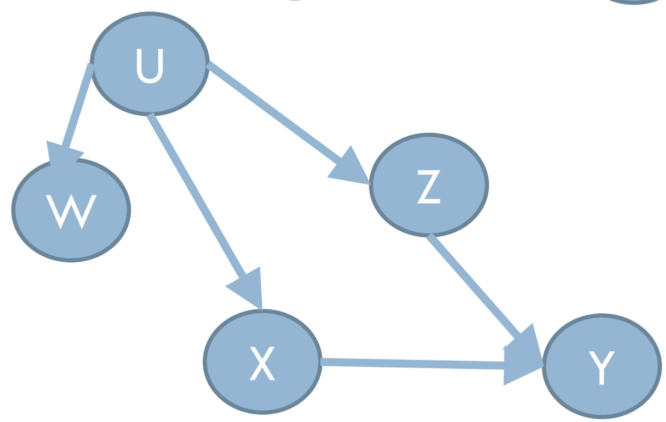
R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and
(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

then infer '±' and give $\mathcal{Z}$ as an admissible set.




- Note again the necessity of the dependence of W and Y

# Reverting the Question

- What if **instead of** using W to find Z to make an adjustment by the back-door criterion, **we find** a Z to allow W to be an instrumental variable that gives bounds on the ACE?

# Why do We Care?

- A way to weaken the faithfulness assumption
  - Suppose also by "independence", we might mean "weak dependence" (and by "dependence", we might mean "strong dependence")
- How would interpret the properties of W in this case, given Rule 1?

R1: If there exists a variable $w \in \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W} \setminus \{w\}$ such that

(i) $w \not\perp\!\!\!\perp y \mid \mathcal{Z}$, and

(ii) $w \perp\!\!\!\perp y \mid \mathcal{Z} \cup \{x\}$

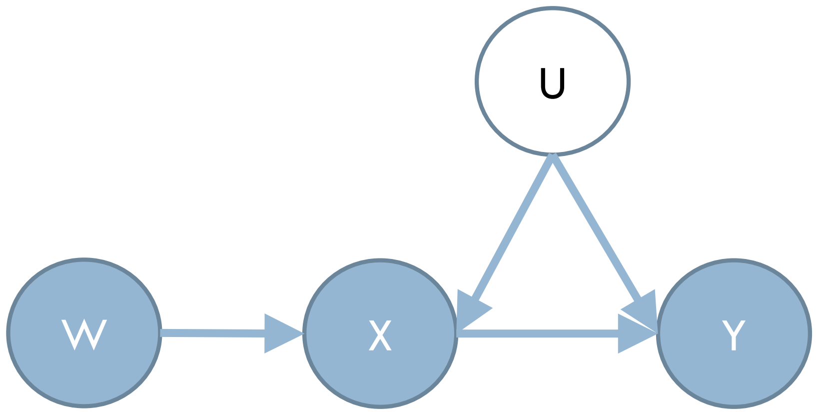
then infer '$\pm$' and give $\mathcal{Z}$ as an admissible set.

# Modified Setup: Main Assumption Statement

- Given Rule 1, assume W is a "conditional IV for X → Y" in the sense that given Z
  - All active paths between W and X are into X
  - There is no "strong direct effect" of W on Y
  - There are no "strong active paths" between W and X, nor W and Y, through common ancestors of X and Y
- The definition of "strong effect/path" creates free parameters we will have to deal with, and a *continuum* of faithfulness-like assumptions.

# Motivation

- Bounds on the ACE in the "standard IV model" can be *quite wide even when* $W \perp\!\!\!\perp Y \mid X$

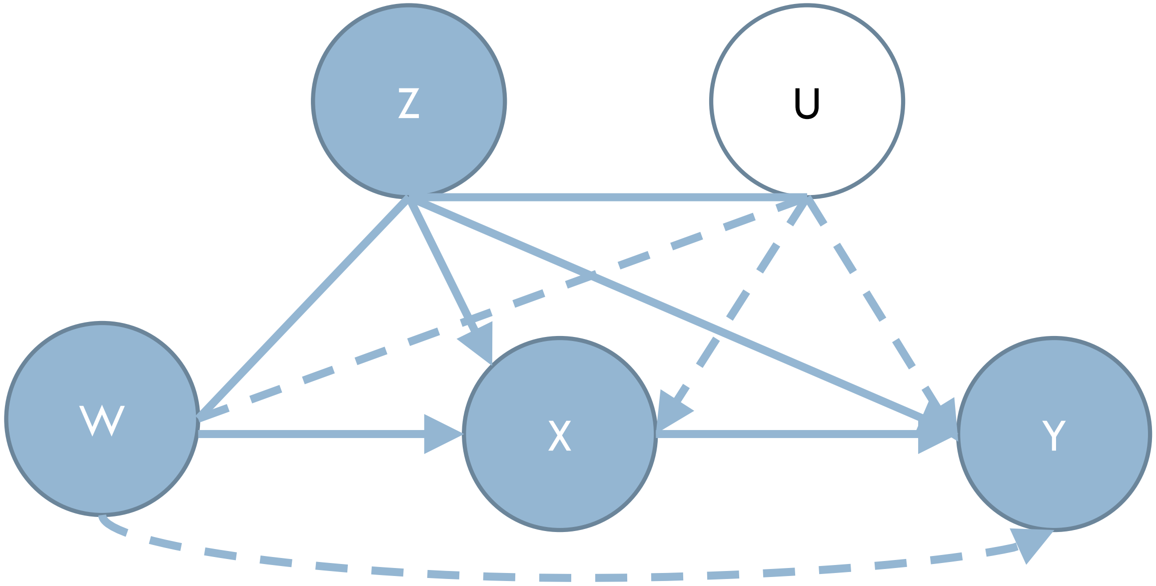



Upper minus lower bound = $1 - |P(X = 1 \mid W = 1) - P(X = 1 \mid W = 0)|$

- This means faithfulness can be quite a strong assumption, and/or "worst-case" analysis can be quite conservative.

# Motivation

- Our analysis can be seen as a way of bridging the two extremes of point estimators of faithfulness analysis and IV bounds without effect constraints.



# The High-Level Idea

- The following might be complicated, but here's a summary:
  - Introduce a **redundant parameterization**, parameters for the two regimes (observational regime, and regime with intervention on X).
  - These parameters cannot be fully unconstrained if we assume "some edges are weak".
    - Machinery behind is linear programming.
  - So statistical inference on the observational regime implies statistical inference on bounds of the ACE.
    - Machinery behind is Bayesian learning with MCMC.

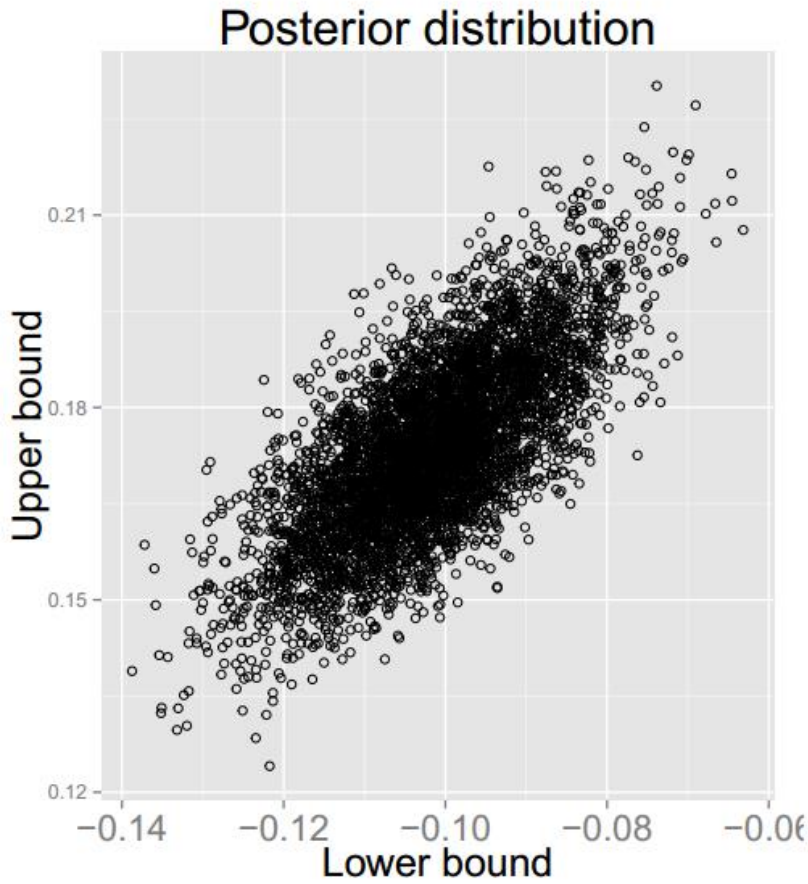
# Illustration of Result: Influenza Data

- Effect of influenza vaccination (X) on hospitalization (Y = 1 means hospitalized)
- Covariate GRP: randomized, doctor of that patient received letter to encourage vaccination
  - Bounds on **average causal effect** using standard methods: [-0.23, 0.64]
- The method we will discuss instead picked DM (diabetes history), AGE (dichotomized at 60 years) and SEX as variables that allowed for adjustment.

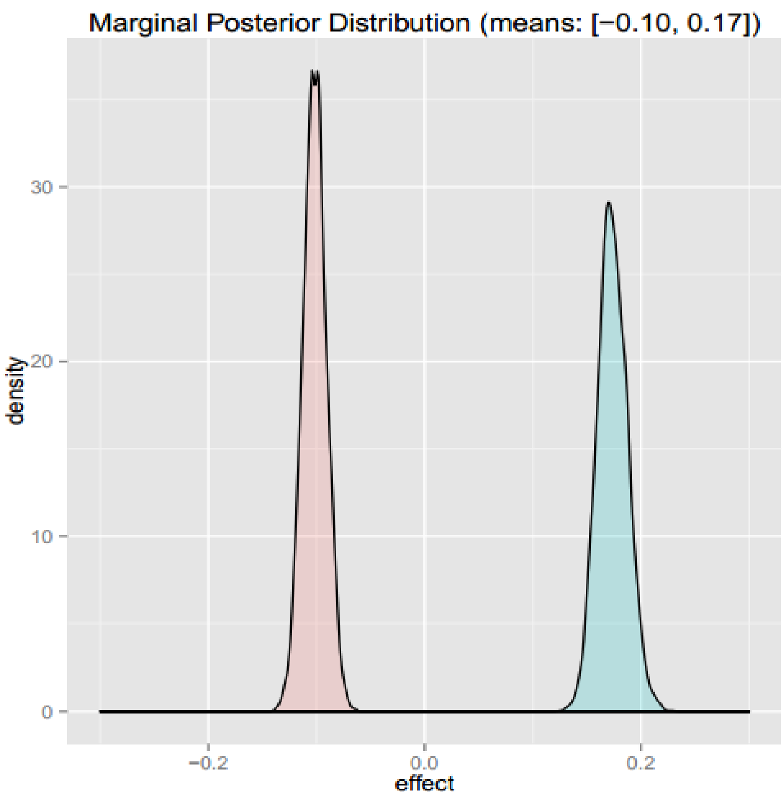
# Influenza Data

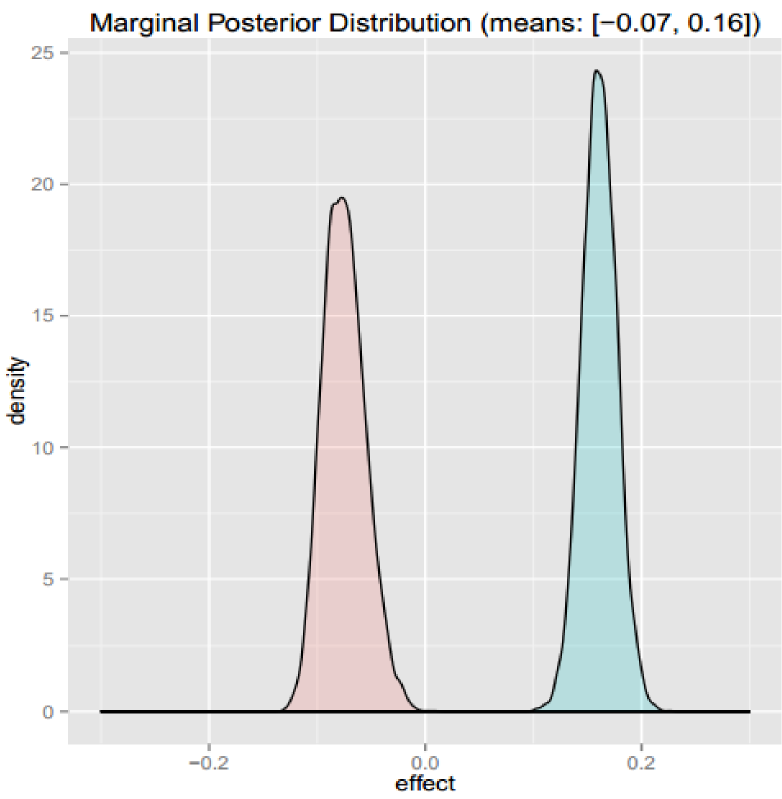
- Our method's estimated interval: [-0.10, 0.17].
- Under some sensitivity analysis postprocessing, the estimate was [-0.02, 0.02].

# Influenza Data: Full Posterior Plots



# Influenza Data: Full Posterior Plots



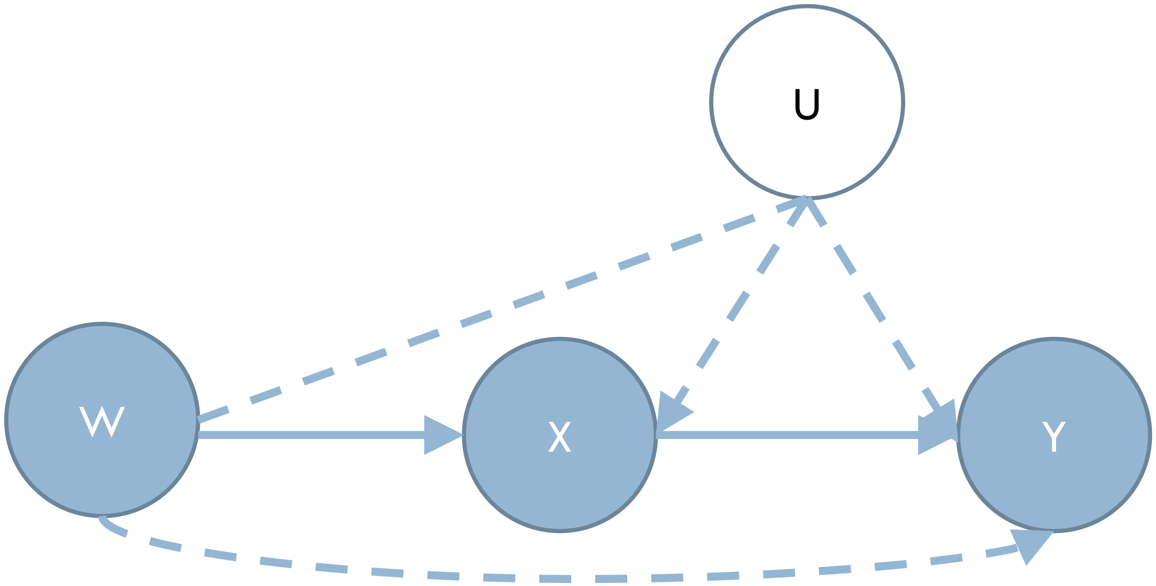



# The High-Level Idea

- The following might be complicated, but here's a summary:
  - Introduce a **redundant parameterization**, parameters for the two regimes (observational, and intervention on X).
  - These parameters cannot be fully unconstrained if we assume "some edges are weak".
    - Machinery behind is linear programming.
  - So statistical inference on the observational regime implies statistical inference on bounds of the ACE.
    - Machinery behind is Bayesian learning with MCMC.

# Expressing Assumptions

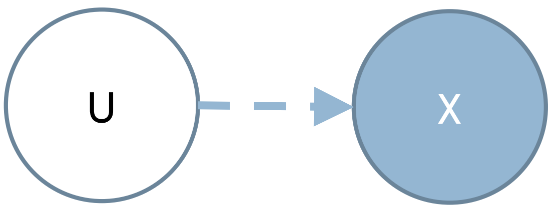
- Some notation first, ignoring Z for now:



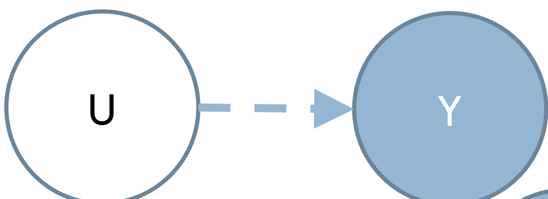

$$
\begin{aligned}
\zeta^{\star}_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^{\star}_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^{\star}_{w} &\equiv P(X = 1 \mid W = w, U)
\end{aligned}
$$

# Stating Assumptions

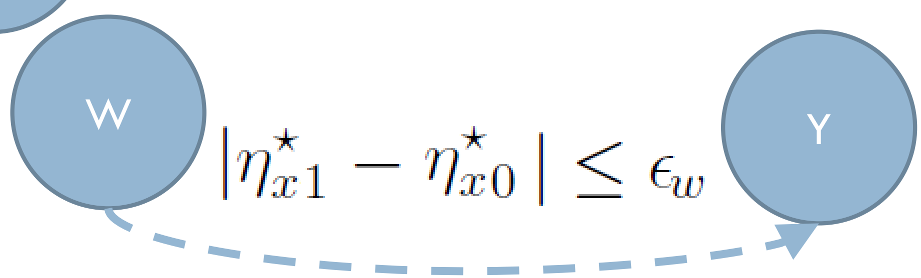
$$
\begin{aligned}
\zeta^{\star}_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^{\star}_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^{\star}_{w} &\equiv P(X = 1 \mid W = w, U)
\end{aligned}
$$




$$|\delta^{\star}_{w} - P(X = 1 \mid W = w)| \leq \epsilon_x$$

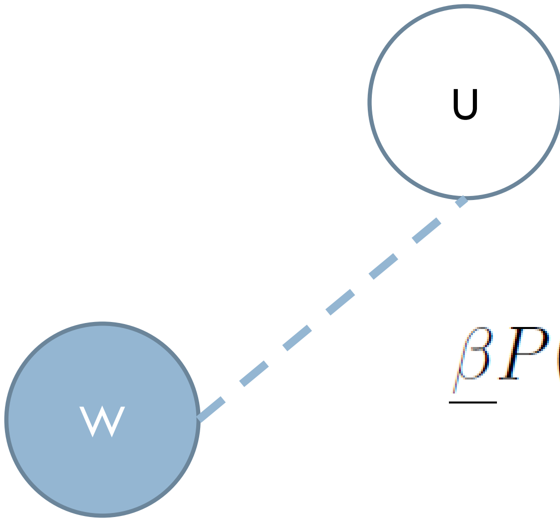



$$|\eta^{\star}_{xw} - P(Y = 1 \mid X = x, W = w)| \leq \epsilon_y$$




$$|\eta^{\star}_{x1} - \eta^{\star}_{x0}| \leq \epsilon_w$$

# Stating Assumptions




$$\underline{\beta} P(U) \leq P(U \mid W = w) \leq \bar{\beta} P(U)$$

# Relation to Observations

$$\begin{aligned}
\zeta^{\star}_{yx.w} &\equiv P(Y = y, X = x \mid W = w, U) \\
\eta^{\star}_{xw} &\equiv P(Y = 1 \mid X = x, W = w, U) \\
\delta^{\star}_{w} &\equiv P(X = 1 \mid W = w, U)
\end{aligned}$$

- Let $\zeta_{yx.w}$ be the expectation of the first entry by $P(U \mid W)$: this is $P(Y = y, X = x \mid W = w)$
- Similarly, let $\eta_{xw}$ be the expectation of the second entry: this is $P(Y = 1 \mid do(X = x), W = w)$

# Context

- The parameterization given was originally exploited by Dawid (2000) and Ramsahai (2012)
- It provides an alternative to the structural equation model parameterization of Balke and Pearl (1997)
- Both approaches work by mapping the problem of testing the model and bounding the ACE by a linear program
- We build on this strategy, with some generalizations

# Estimation

- Simpler mapping on $(\delta^*, \eta^*) \rightarrow P(W, X, Y \mid U)$, marginalized, gives constraints on $\zeta \equiv P(W, X, Y)$
- Test whether constraints hold, if not provide no bounds
- Plug-in estimates for $\zeta$ to get $(\zeta, \eta)$ polytope. Find upper bounds and lower bounds on the ACE by solving linear program and maximizing/minimizing objective function

$$f(\eta) = (\eta_{11} - \eta_{01})P(W = 1) + (\eta_{10} - \eta_{00})P(W = 0)$$

# Coping with Non-linearity

- Notice that because of constraints such as

$$|\delta_w^\star - P(X = 1 \mid W = w)| \leq \epsilon_x$$

there will be non-linear constraints in $\zeta \equiv P(W, X, Y)$

- The implied constraints are still linear in $\eta \equiv P(Y \mid do(X), W)$. So linear programming formulation still holds, treating $\zeta$ as a constant.
  - Non-linearity on $\zeta$ can be a problem for estimation of $\zeta$ and derivation of confidence intervals. We will describe later a Bayesian approach that does that simply by rejection sampling

# Algorithm

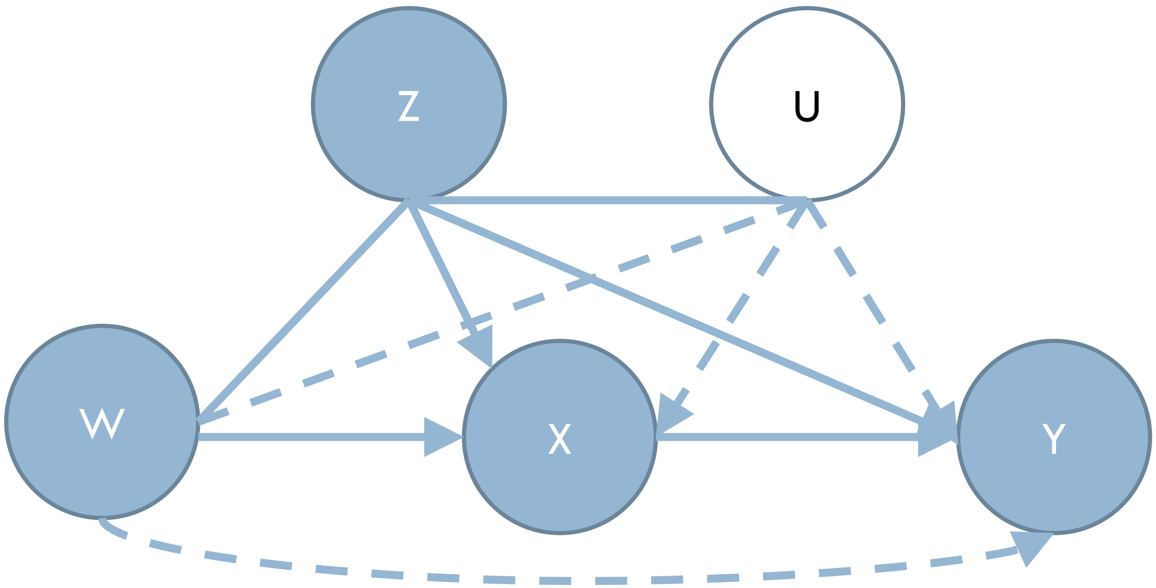
In what follows, we assume dimensionality of Z is small, |Z| < 10

**input** : Binary data matrix $\mathcal{D}$; set of relaxation parameters $\theta$; covariate index set $\mathcal{W}$; cause-effect indices $X$ and $Y$
**output**: A list of pairs (witness, admissible set) contained in $\mathcal{W}$

```
L ← ∅;
for each W ∈ 𝒲 do
    for every admissible set Z ⊆ 𝒲\{W} identified by W and θ given 𝒟 do
        B ← posterior over upper/lowed bounds on the ACE as given by (W, Z, X, Y, 𝒟, θ);
        if there is no evidence in B to falsify the (W, Z, θ) model then
            L ← L ∪ {B};
        end
    end
end
return L
```

# Recap: So far, everything in the population

- "Rely on the identification of an auxiliary variable W (**witness**), an auxiliary set Z (**background set**), and **assumptions about strength of dependencies on latent variables**"

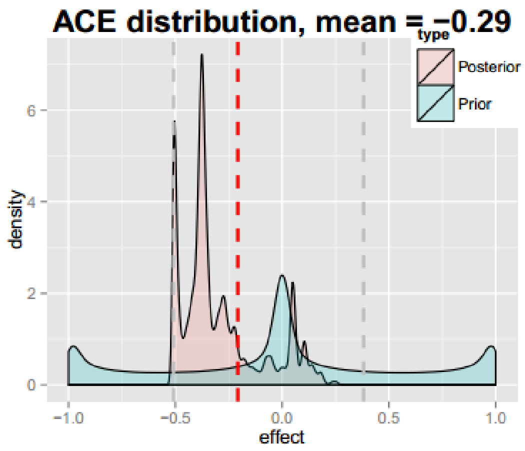


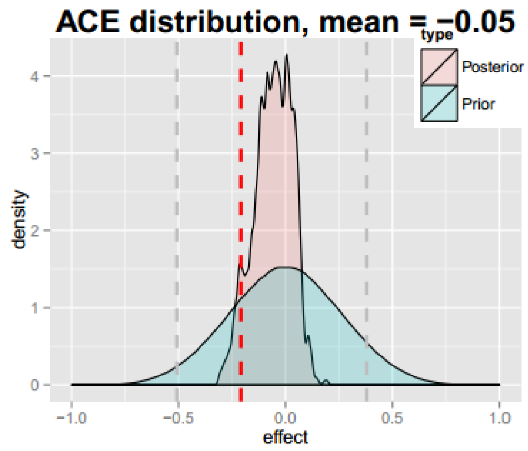
# Bayesian Learning

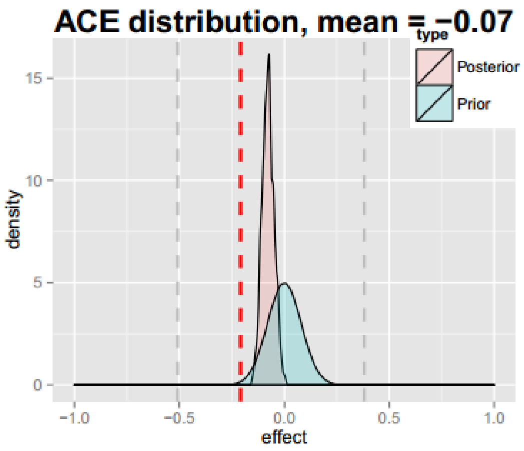
- To decide on independence, we do Bayesian model selection with a contingency table model with Dirichlet priors
- For each pair (W, Z), find posterior bounds for each configuration of Z
  - Use Dirichlet prior for $\zeta$ (for each $Z = z$), conditioned on the constraints of the model, using rejection sampling
    - Propose from unconstrained Dirichlet
  - Reject model if 95% or more of proposed parameters are rejected in the initial round of rejection sampling
  - Feed sample from the posterior of $\zeta$ into linear program to get a sample for the upper bound and lower bound

# Difference wrt ACE Bayesian Learning

- Why not put a prior directly on the latent variable model?
  - Model is unidentifiable → results extremely sensitive to priors
  - Putting priors directly into ζ produces no point estimates, but avoids prior sensibility



# Wrapping Up

- Finally, one is left with different posterior distributions over different bounds on the ACE
- Final step is how to summarize possibly conflicting information. Possibilities are:
  - Report tightest bound
  - Report widest bound
  - Report combined smallest lower bound with largest upper bound
  - Use "posterior of Rule 1" to pick a handful of bounds and discard others

# Recap

- Invert usage of Entner's Rules towards the instrumental variable point of view.
- Obtain bounds, not point estimates.
- Use Bayesian inference, set up a rule to combine possibly conflicting information.

# “Witness Protection Program”

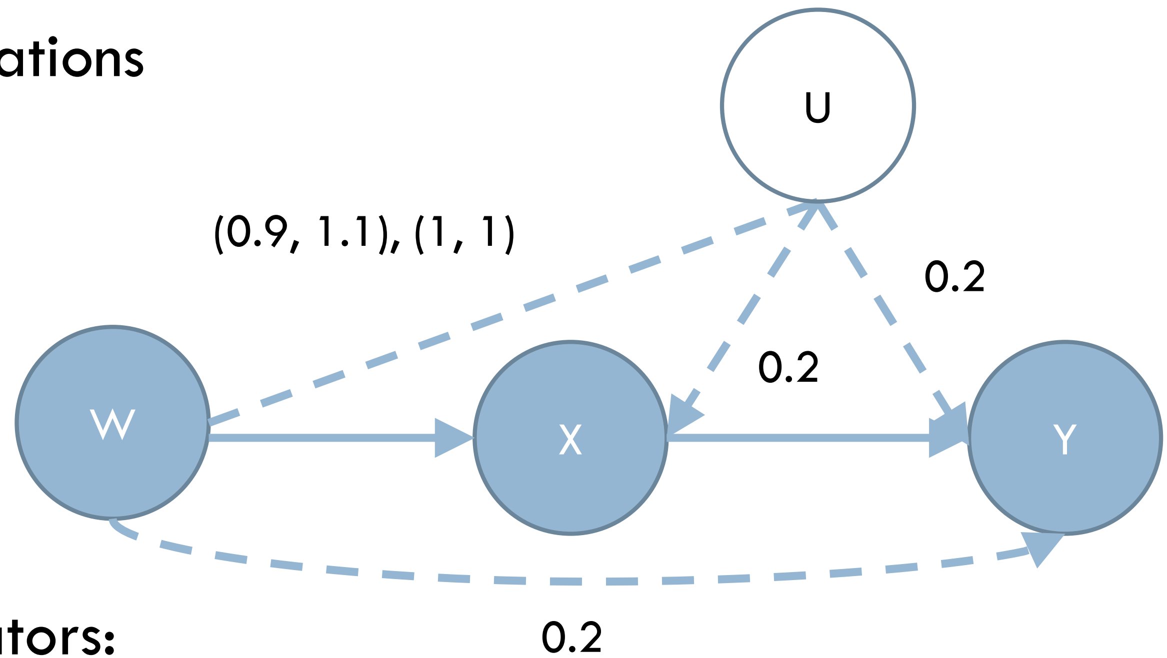
- Because the framework relies on using a linear program to protect a witness variable against violations of faithfulness, we call this the ***Witness Protection Program*** (WPP) algorithm.

# Illustration: Synthetic Studies

- 4 observable nodes, "basic set", form a pool that can generate a possible (witness, background set) pair
- 4 observable nodes form a "decoy set": none of them should be included in the background set
- Graph structures over "basic set" + {X, Y} are chosen randomly
- Observable parents of "decoy set" are sampled from "basic set"
- Each decoy has another four latent parents, $\{L_1, L_2, L_3, L_4\}$
- Latents are mutually independent
- Each latent variable $L_i$ uniformly chooses either X or Y as a child
- Conditional distributions are logistic regression models with pairwise interactions

# Illustration: Synthetic Studies

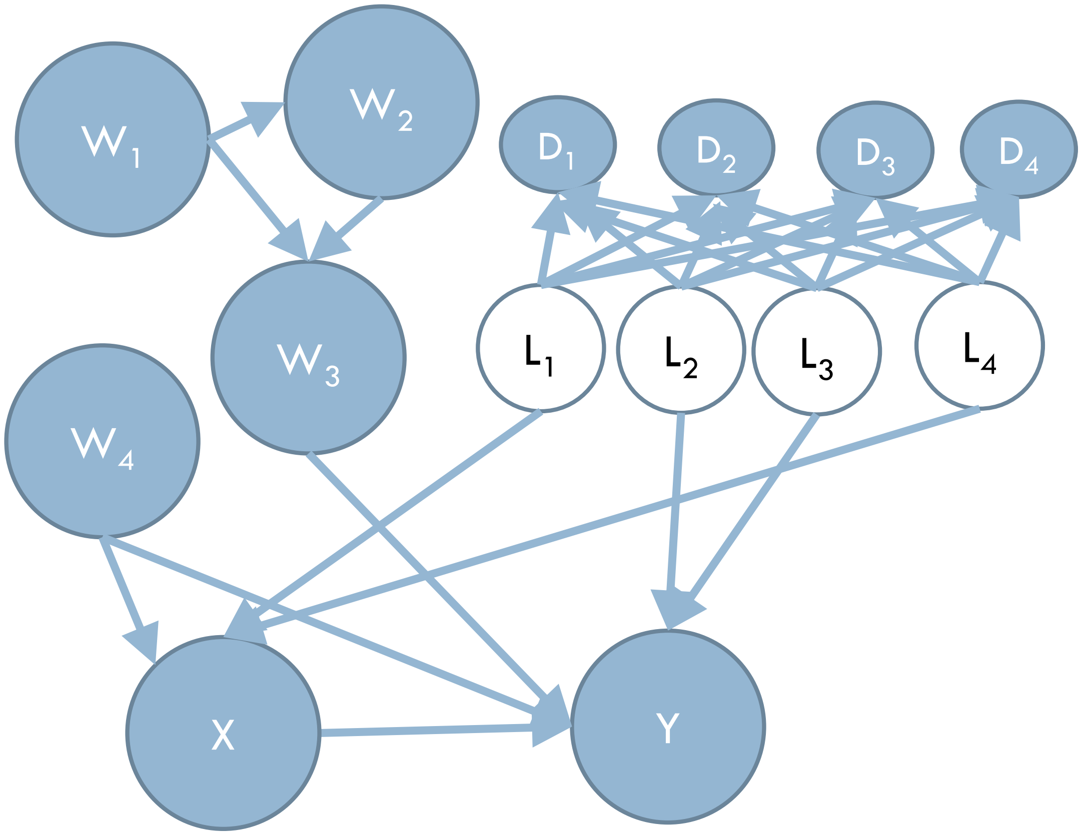
- Relaxations




- Estimators:
  - Posterior expected bounds
  - Naïve 1: back-door adjustment conditioning on everybody
  - Naïve 2: plain $P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0)$
  - Backdoor by faithfulness

# Example

- Note: no theoretical witness solution



# Evaluation

- Bias definition:
  - For point estimators, just absolute value of difference between true ACE and estimate
  - For bounds, Euclidean distance between true ACE and nearest point in the bound
- Summaries (over 100 simulations):
  - Bias average
  - Bias tail mass at 0.1
    - proportion of cases where bias exceeds 0.1
- Notice difficulty of direct comparisons

# Summary

| **Hard, Solvable:** NE1 = (0.18, 1.00), NE2 = (0.19, 0.63) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_\epsilon$ | Found | Faith.1 | | WPP1 | | Width1 | WPP2 | | Width2 |
| 0.05 | 0.75 | 0.15 | 0.57 | 0.12 | 0.43 | 0.06 | 0.05 | 0.19 | 0.35 |
| 0.10 | 0.95 | 0.16 | 0.58 | 0.10 | 0.37 | 0.14 | 0.04 | 0.16 | 0.42 |
| 0.15 | 0.99 | 0.15 | 0.60 | 0.07 | 0.27 | 0.21 | 0.02 | 0.09 | 0.50 |
| 0.20 | 0.99 | 0.15 | 0.61 | 0.05 | 0.17 | 0.28 | 0.02 | 0.08 | 0.59 |
| 0.25 | 1.00 | 0.15 | 0.60 | 0.03 | 0.13 | 0.36 | 0.01 | 0.03 | 0.68 |
| 0.30 | 1.00 | 0.16 | 0.61 | 0.02 | 0.09 | 0.43 | 0.00 | 0.01 | 0.74 |
| **Easy, Solvable:** NE1 = (0.04, 0.13), NE2 = (0.08, 0.29) | | | | | | | | | |
| $k_\epsilon$ | Found | Faith.1 | | WPP1 | | Width1 | WPP2 | | Width2 |
| 0.05 | 0.78 | 0.04 | 0.09 | 0.03 | 0.06 | 0.06 | 0.01 | 0.04 | 0.36 |
| 0.10 | 0.97 | 0.04 | 0.09 | 0.02 | 0.05 | 0.14 | 0.01 | 0.03 | 0.44 |
| 0.15 | 0.99 | 0.05 | 0.10 | 0.01 | 0.04 | 0.21 | 0.00 | 0.01 | 0.52 |
| 0.20 | 0.99 | 0.05 | 0.10 | 0.01 | 0.04 | 0.29 | 0.00 | 0.00 | 0.60 |
| 0.25 | 0.99 | 0.05 | 0.09 | 0.00 | 0.01 | 0.37 | 0.00 | 0.00 | 0.67 |
| 0.30 | 1.00 | 0.05 | 0.09 | 0.00 | 0.01 | 0.44 | 0.00 | 0.00 | 0.75 |

Bias average

Bias tail mass at 0.1

# Summary

| **Hard, Not Solvable:** NE1 = (0.16, 1.00), NE2 = (0.20, 0.88) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $k_\epsilon$ | Found | Faith.1 | | WPP1 | | Width1 | WPP2 | | Width2 |
| 0.05 | 0.67 | 0.20 | 0.90 | 0.17 | 0.76 | 0.06 | 0.04 | 0.14 | 0.32 |
| 0.10 | 0.91 | 0.19 | 0.91 | 0.13 | 0.63 | 0.10 | 0.02 | 0.07 | 0.39 |
| 0.15 | 0.97 | 0.19 | 0.92 | 0.10 | 0.41 | 0.18 | 0.01 | 0.03 | 0.45 |
| 0.20 | 0.99 | 0.19 | 0.95 | 0.07 | 0.25 | 0.24 | 0.01 | 0.01 | 0.51 |
| 0.25 | 1.00 | 0.19 | 0.96 | 0.03 | 0.13 | 0.31 | 0.00 | 0.00 | 0.58 |
| 0.30 | 1.00 | 0.19 | 0.96 | 0.02 | 0.06 | 0.39 | 0.00 | 0.00 | 0.66 |
| **Easy, Not Solvable:** NE1 = (0.09, 0.32), NE2 = (0.14, 0.56) | | | | | | | | | |
| $k_\epsilon$ | Found | Faith.1 | | WPP1 | | Width1 | WPP2 | | Width2 |
| 0.05 | 0.68 | 0.13 | 0.51 | 0.10 | 0.37 | 0.05 | 0.02 | 0.07 | 0.33 |
| 0.10 | 0.97 | 0.12 | 0.53 | 0.08 | 0.28 | 0.10 | 0.01 | 0.05 | 0.39 |
| 0.15 | 1.00 | 0.12 | 0.52 | 0.05 | 0.17 | 0.16 | 0.01 | 0.03 | 0.46 |
| 0.20 | 1.00 | 0.12 | 0.53 | 0.03 | 0.08 | 0.23 | 0.01 | 0.03 | 0.52 |
| 0.25 | 1.00 | 0.12 | 0.48 | 0.02 | 0.05 | 0.31 | 0.00 | 0.02 | 0.59 |
| 0.30 | 1.00 | 0.12 | 0.48 | 0.01 | 0.04 | 0.39 | 0.00 | 0.01 | 0.65 |

# On-going Work

- Finding a more primitive default set of assumptions where assumptions about the relaxations can be derived from
- Doing without a given causal ordering
- Large scale experiments
- Scaling up for a large number of covariates
- Continuous data
- More real data experiments
- R package available at CRAN/GitHub: "**CausalFX**"

# Thank You, and Shameless Ad

**What If? Inference and Learning of Hypothetical and Counterfactual Interventions in Complex Systems**

**A NIPS 2016 Workshop**
**Centre Convencions Internacional Barcelona, Barcelona, Spain**
**December 10th 2016**

**Deadline: October 31st**

**https://sites.google.com/site/whatif2016nips/call-for-papers**