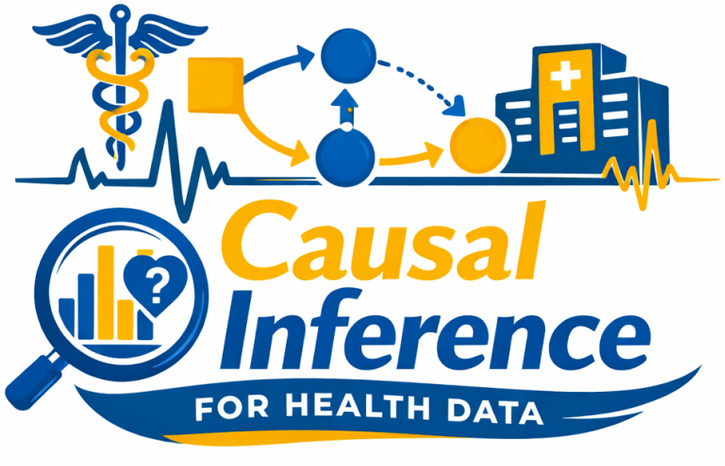
Causal Inference for Health Data

(STATS C160/C260 — Winter 2026)

Lecture 16: Missing Data

Drago Plečko

UCLA College | Physical Sciences Statistics & Data Science



How common is missingness in health data?

Methods

PMID: 25407057

Methods

PMID: 41211173

Methods

PMID: 39543512

We performed a retrospective study of randomised clinical trial reports of psychological

Conclusion: Missingness is very common in health data.

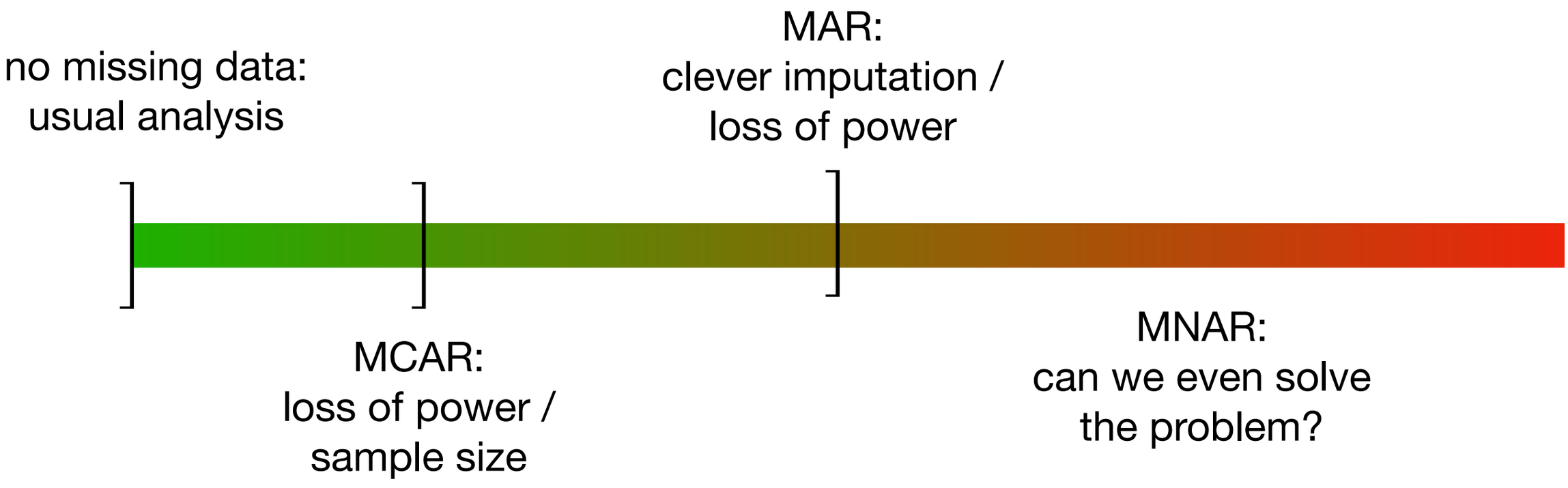
the manuscript.

Results

We identified 182 randomised clinical trials (233 primary outcomes), of which 206 outcomes (88.4%) were assessed at high risk of bias due to missing data.

Challenges of Missing Data

- What kind of challenges does missing data pose? In the first instance, we cannot run our analysis as usual.
- In this lecture, we will study the missing data spectrum.



Note: this is a containment hierarchy

$$NM \subset MCAR \subset MAR \subset MNAR$$

Missingness Graphs (m-graphs)

Definition: A missingness graph (m-graph) is a causal diagram partitioned into:

- V_o, covariates always observed;
- V_m, covariates partially missing;
- $\boldsymbol{U} = \{U_1,\ldots,U_m\}$ exogenous variables;
- R, represeting missingness mechanisms; R_i never a parent of variables $V \cup U$;
- V^*, proxy variables that are actually observed:

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases}$$

Comparison with classical SCM setup

Element in SCM		Appearance in m-graph
Observables V	partitioned →	Always obs. V_o + Partially miss. V_m
Unobservables U	remains the same →	Unbservables U
mechanisms $\mathcal{F}$	extended with R, V^* →	mechanisms $\mathcal{F}^{+R}$
$P(\mathbf{u})$ distribution	remains the same →	$P(\mathbf{u})$ distribution

Example: Obesity Reporting

- Suppose we are trying to analyze obesity in a group of students,
- We collect age, sex, and obesity,
- Age/sex are always available, but students may sometimes not report their weight.

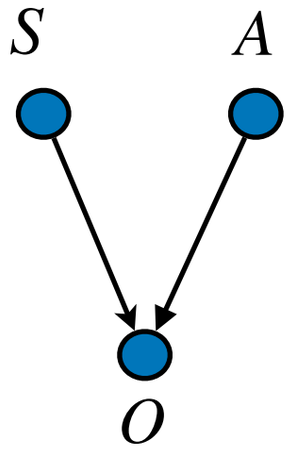
$$V_o = \{\textbf{Age}, \textbf{Sex}\}$$

$$V_m = \{\textbf{Obesity}\}$$

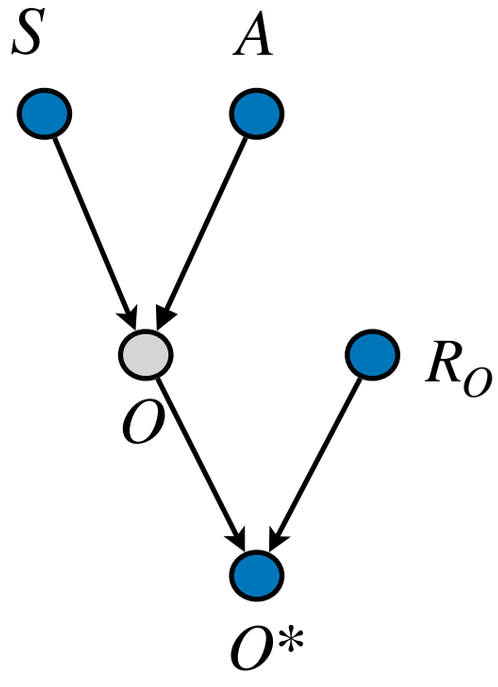
Age	Sex	Obesity
16	*F*	*Obese*
15	*F*	***m***
15	*M*	***m***
14	*F*	*Not Obese*
13	*M*	*Not Obese*
15	*M*	*Obese*
14	*F*	*Obese*

Example: Obesity Reporting

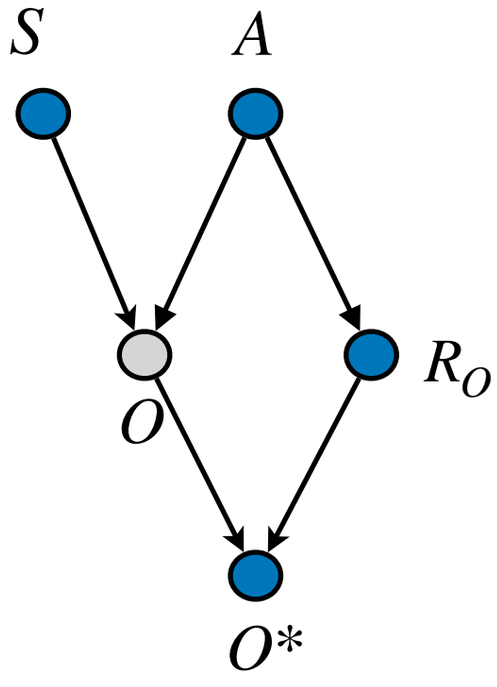
- Mechanism R_O determines how missingness of variable O is achieved, and it can be influenced by different variables:



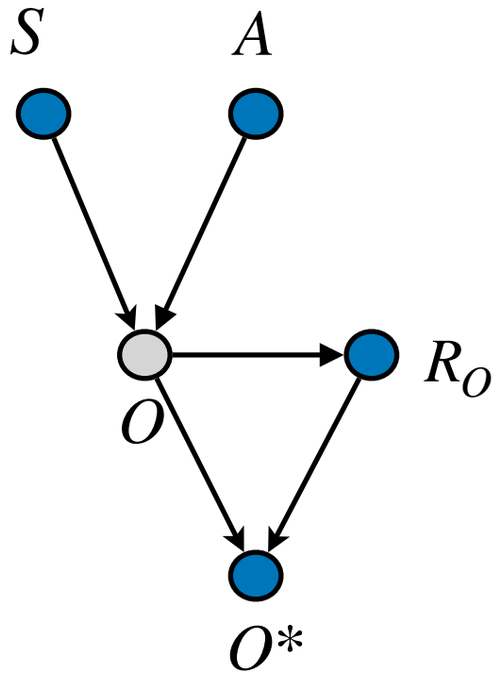
(a) Standard Case (no missingess)



(b) R_O not influenced by any V



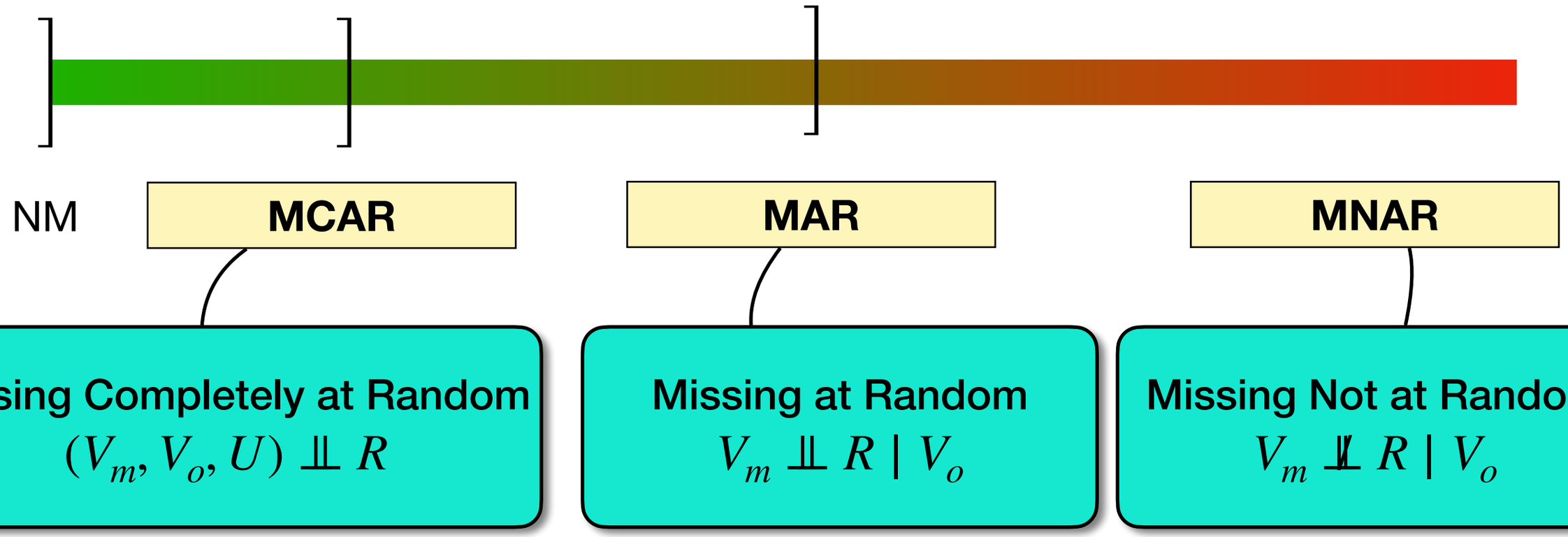
(c) R_O influenced by A



(d) R_O influenced by O

Taxonomy of Missingness

Definition: We define the following missingness settings:

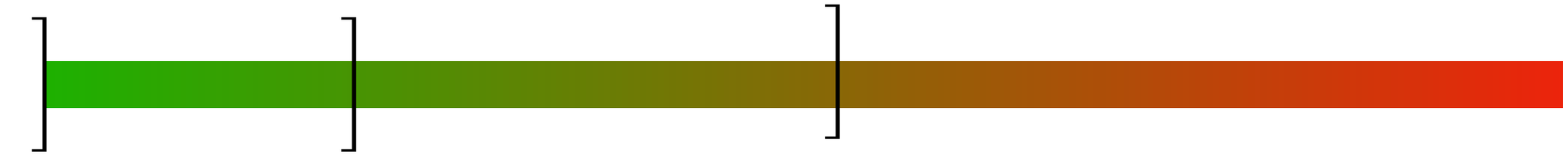


- missingness essentially unrelated to variables in the system

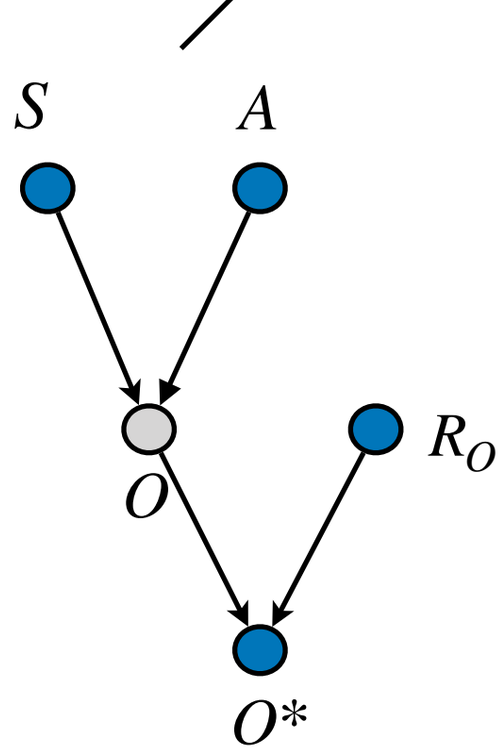
- conditional on always observed variables V_o, missingness does not depend on V_m

- even after conditioning on V_o , missingness may depend on V_m

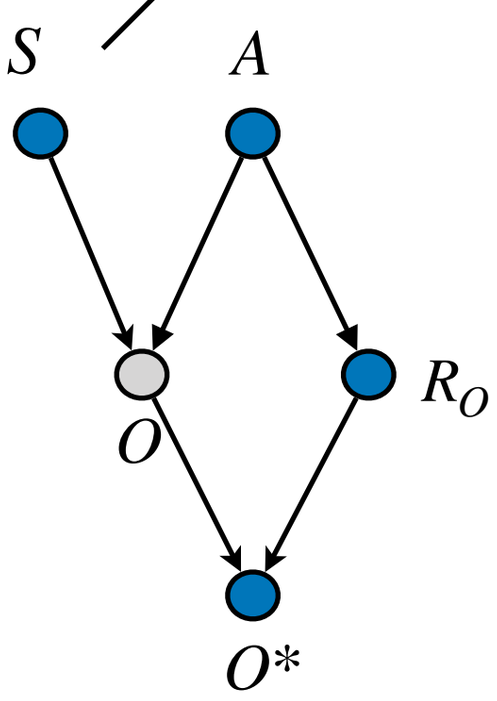
Taxonomy of Missingness: Examples



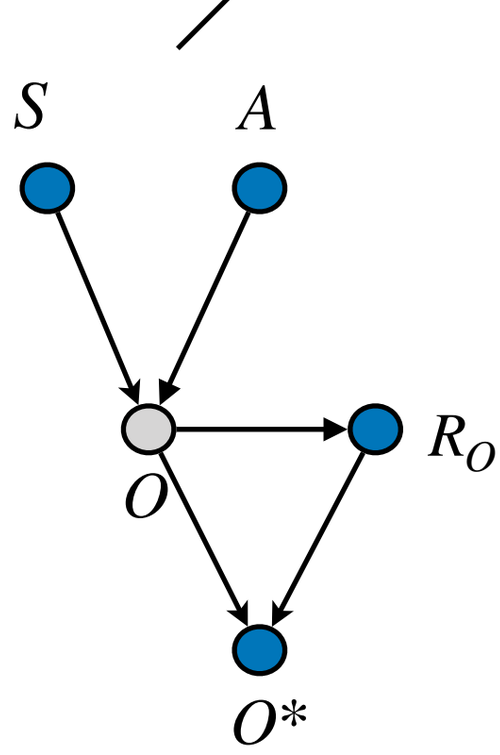
NM | **MCAR** | **MAR** | **MNAR**



(b) R_O not influenced by any V



(c) R_O influenced by A



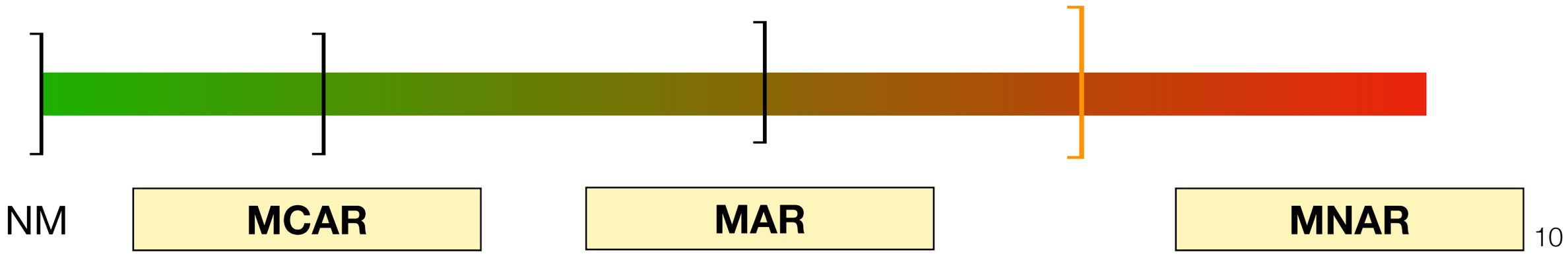
(d) R_O influenced by O

Recoverability and m-Identification

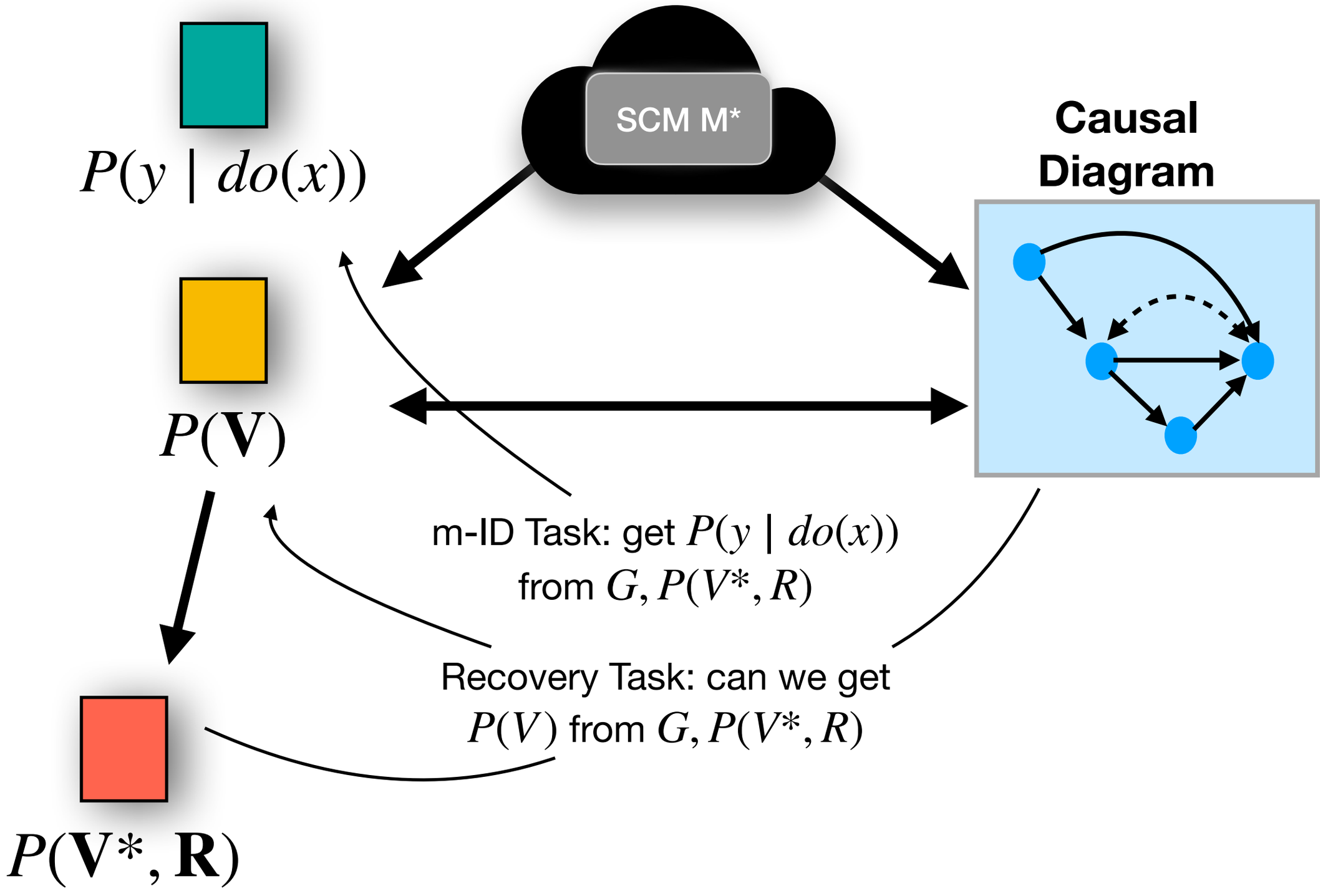
Definition: A distribution $P(V)$ recoverable if it can be uniquely determined from the m-graph G and the observed distribution $P(V^*, R)$.

A causal effect $P(y \mid do(x))$ is m-identifiable if it can be uniquely determined from the m-graph G and the observed distribution $P(V^*, R)$.

boundary of recoverability?



Inferential Tasks



MCAR: Easy to Handle

Proposition. In the setting of missing completely at random (MCAR), we have the following result:

$$P(V) = P(V \mid R = 0) = P(V^* \mid R = 0)$$

Distribution $P(V)$ same as $P(V \mid R = 0)$

Focusing on complete cases is enough

Loss of sample size is the only difficulty

Complete-Case Analysis: Not a Universal Panacea

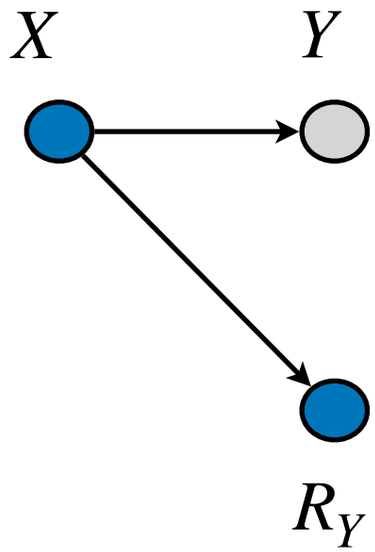
- The complete-case analysis is not a universal solution, and it can give biased results.

$$X \leftarrow \epsilon_X$$

$$Y \leftarrow X + \epsilon_Y$$

$$R_Y \leftarrow \textbf{Bern}(\textbf{expit}(X))$$

$$\epsilon_X, \epsilon_Y \sim N(0,1)$$



$$Y \perp\!\!\!\perp R \mid X \implies \textbf{MAR}$$

What happens to $E[Y \mid R = 0]$ compared to $E[Y]$?

Beyond MCAR:
MAR ≠ Rubin-MAR

- The MAR definition we used is not the common definition of MAR used in statistics textbooks,
- The most common definition is Rubin-MAR, given by:

Definition. The missingness mechanism satisfies Rubin-MAR if

$$P(R \mid V_{obs}, V_{miss}) = P(R \mid V_{obs})$$

where V_{obs}, V_{miss} are the observed and missing values in the data sample.

MAR ≠ Rubin-MAR

MAR

- set of variables V_o, which are considered observed, must be observed for each sample.
- satisfied if $V_m \perp\!\!\!\perp R \mid V_o$ is implied by the m-graph
- structural property

Rubin MAR

- set of variables V_{obs}, which are considered observed, could be unobserved for other samples.
- variables V_{obs} can vary from one simple to another, but must satisfy $P(R \mid V_{obs}, V_{miss}) = P(R \mid V_{obs})$ in each sample
- sample-based property

MAR ≠ Rubin-MAR: Example

$$X \leftarrow \epsilon_X$$

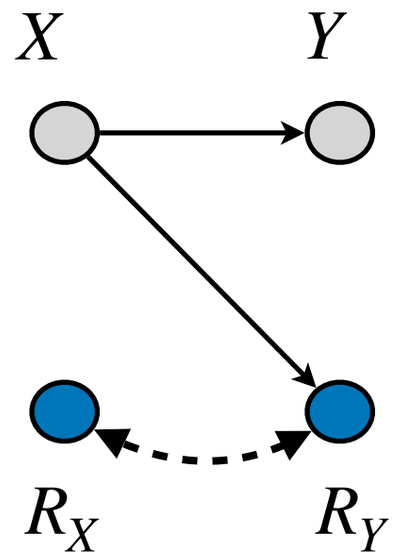
$$Y \leftarrow X + \epsilon_Y$$

$$R_X \leftarrow U_{R_{xy}}$$

$$R_Y \leftarrow \min(1 - U_{R_{xy}}, 1(X > 1))$$

$$\epsilon_X, \epsilon_Y \sim N(0,1)$$

$$U_{R_{xy}} \sim \textbf{Bern}(0.5)$$



$$V_o = \emptyset$$

$$V_m = \{X, Y\}$$

$$R_X, R_Y \not\perp\!\!\!\perp X, Y$$

This is not MAR!

Question: does Rubin-MAR hold?

MAR ≠ Rubin-MAR: Example

Note: $R_X = 1 \implies R_Y = 0,$
$R_Y = 1 \implies R_X = 0$

Evaluating $P(R \mid V_{obs}, V_{miss}) = P(R \mid V_{obs})$:

$$X \leftarrow \epsilon_X$$
$$Y \leftarrow X + \epsilon_Y$$
$$R_X \leftarrow U_{R_{xy}}$$
$$R_Y \leftarrow \min(1 - U_{R_{xy}}, 1(X > 1))$$
$$\epsilon_X, \epsilon_Y \sim N(0,1)$$
$$U_{R_{xy}} \sim \textbf{Bern}(0.5)$$

- $(R_X, R_Y) = (0,0)$

want to show: $P(R_X = 0, R_Y = 0 \mid X, Y) = P(R_X = 0, R_Y = 0 \mid X, Y)$ ✓

follows directly

- $(R_X, R_Y) = (0,1)$

$$P(R_X = 0, R_Y = 1 \mid X, Y) = P(R_X = 0 \mid X, Y) P(R_Y = 1 \mid R_X = 0, X, Y)$$
$R_X \perp\!\!\!\perp Y \mid X$ $R_Y \perp\!\!\!\perp Y \mid X, R_X$
$$= P(R_X = 0 \mid X) P(R_Y = 1 \mid R_X = 0, X)$$
$$= P(R_X = 0, R_Y = 1 \mid X)$$ ✓

MAR ≠ Rubin-MAR: Example

- $(R_X, R_Y) = (1,0)$

$$P(R_X = 1, R_Y = 0 \mid X, Y) = P(R_X = 1 \mid X, Y) P(R_Y = 0 \mid R_X = 1, X, Y)$$

$R_X \perp\!\!\!\perp X \mid Y$ **as** $R_X = 1 \implies R_Y = 0$, **both** $= 1$

$$= P(R_X = 1 \mid Y) P(R_Y = 0 \mid R_X = 1, Y)$$

$$= P(R_X = 1, R_Y = 0 \mid Y)$$ ✓

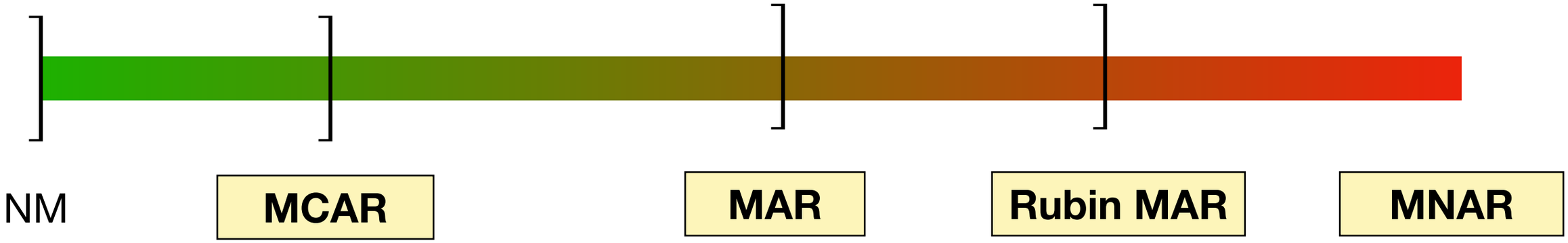
- $(R_X, R_Y) = (1,1)$ does not occur, so nothing to check.

The example is Rubin MAR, but not MAR!

MAR $\Longrightarrow$ Rubin-MAR

Proposition: Any setting that is MAR is also Rubin-MAR, i.e.,

$$\text{MAR} \implies \text{Rubin-MAR}$$



Rubin-MAR is a weaker condition than MAR!

Benefits of the Graphical Definitions

- What kind of benefits does the graphical approach provide compared to typical statistical definitions?

(A) Transparency

- all assumptions are explicitly encoded in the m-graph, at the mechanism level
- in the classical approach, arguing whether $P(R \mid V_{obs}, V_{miss}) = P(R \mid V_{obs})$ holds may be very difficult

(B) Testability

- MAR settings often have testable implications: we can test if our assumptions hold, based on the data
- Rubin-MAR is known to be untestable

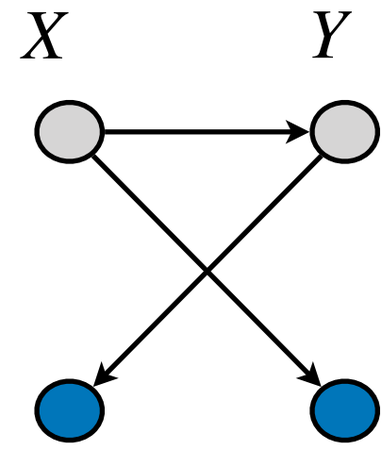
Recovery Beyond Rubin MAR: Deadlock Example

$$X \leftarrow \textbf{Bern}(0.5)$$

$$Y \leftarrow \begin{cases} \textbf{Bern}(2a) & \textbf{if } X = 1 \\ \textbf{Bern}(1-2a) & \textbf{if } X = 0 \end{cases}$$

$$R_X \leftarrow \textbf{Bern}(0.1 + \beta Y)$$

$$R_Y \leftarrow \textbf{Bern}(0.1 + \alpha X)$$



The example is Rubin MNAR — is it recoverable?

We first solve a simpler problem — without the $Y \rightarrow R_X$ edge

Recovery Beyond Rubin MAR: Deadlock Example

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

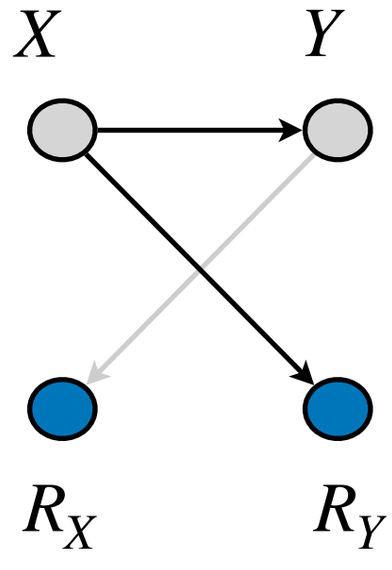
$(Y \perp\!\!\!\perp R_Y, R_X \mid X$ from the graph)

$$= P(Y \mid X, R_Y = 0, R_X = 0)P(X)$$

$(X \perp\!\!\!\perp R_X$ from the graph)

$$= P(Y \mid X, R_Y = 0, R_X = 0)P(X \mid R_X = 0)$$

$$= P(Y^* \mid X^*, R_Y = 0, R_X = 0)P(X^* \mid R_X = 0)$$



both computable from the available data! **so $P(V)$ recoverable**

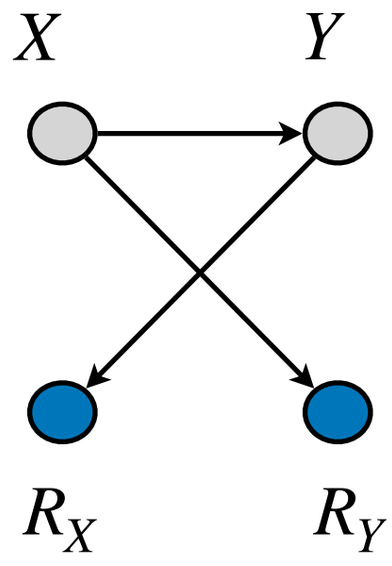
$$P(X, Y) = P(Y \mid X)P(X)$$

$(Y \perp\!\!\!\perp R_Y, R_X \mid X$ from the graph)

$$\neq P(Y \mid X, R_Y = 0, R_X = 0)P(X)$$

$(X \perp\!\!\!\perp R_X$ from the graph)

$$\neq P(Y \mid X, R_Y = 0, R_X = 0)P(X \mid R_X = 0)$$



both steps fail.

is the deadlock non-recoverable?

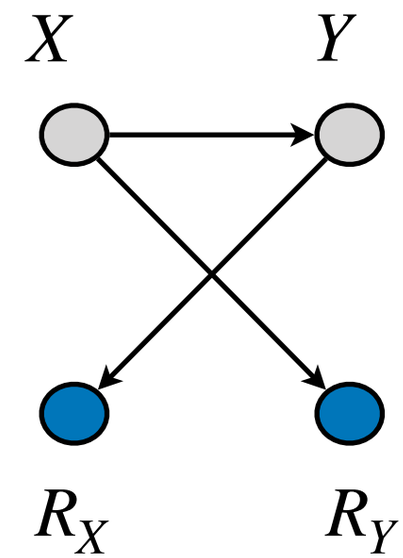
Recovery Beyond Rubin MAR: Deadlock Example

- Which independences do hold?

$$(1): \quad R_X \perp\!\!\!\perp R_Y \mid X, Y$$

$$(2): \quad R_X \perp\!\!\!\perp X, R_Y \mid Y$$

$$(3): \quad R_Y \perp\!\!\!\perp Y, R_X \mid X$$



Recoverable

$$P(X,Y) = P(X,Y)\frac{P(X,Y \mid R_X=0,R_Y=0)}{P(X,Y \mid R_X=0,R_Y=0)}$$

$$= \frac{1}{P(R_X=0,R_Y=0 \mid X,Y)} P(X,Y \mid R_X=0,R_Y=0) P(R_X=0,R_Y=0)$$

$$\overset{(1)}{=} \frac{1}{P(R_X=0 \mid X,Y)P(R_Y=0 \mid X,Y)} \overset{(2,3)}{=} \frac{1}{P(R_X=0 \mid Y,R_Y=0)P(R_Y=0 \mid X,R_X=0)}$$

Recovery Result

Theorem. Suppose that

(i) there are no edges between R variables,

(ii) no bidirected edge into R variables,

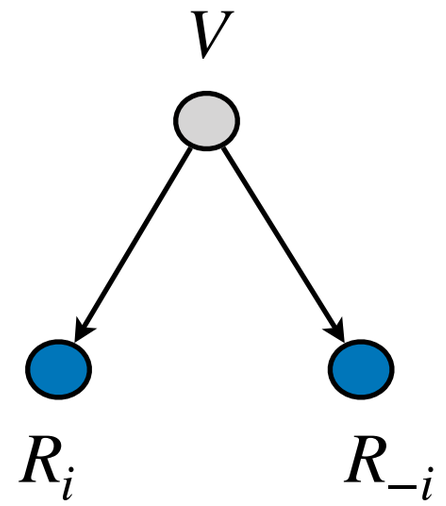
Then, $P(V)$ is recoverable if and only if $X \notin \mathrm{pa}(R_X)$.
Whenever recoverable, we have

$$P(V) = \frac{P(V, R = 0)}{\prod_i P(R_i \mid \mathrm{pa}(R_i), R_{\mathrm{pa}^m(R_i)} = 0)},$$

where $\mathrm{pa}^m(R_i) = \mathrm{pa}(R_i) \cap V_m$.

Proof of Recovery Result

Proof. Note that each R_i is a leaf (terminal) node in the diagram. Since there are also no bidirected edges into R, conditioning on all of V d-separates R_i from any subset of $R_{\_i}$.



Therefore, we have that $P(R = 0 \mid V) = \prod_i P(R_i = 0 \mid V)$.

Further, we also have that $\mathrm{pa}(R_i)$ d-separate R_i from all the other variables in the graph. Therefore, for each term we have

$$P(R_i \mid V) = P(R_i \mid \mathrm{pa}(R_i))$$

By the same d-separate observation, we have that

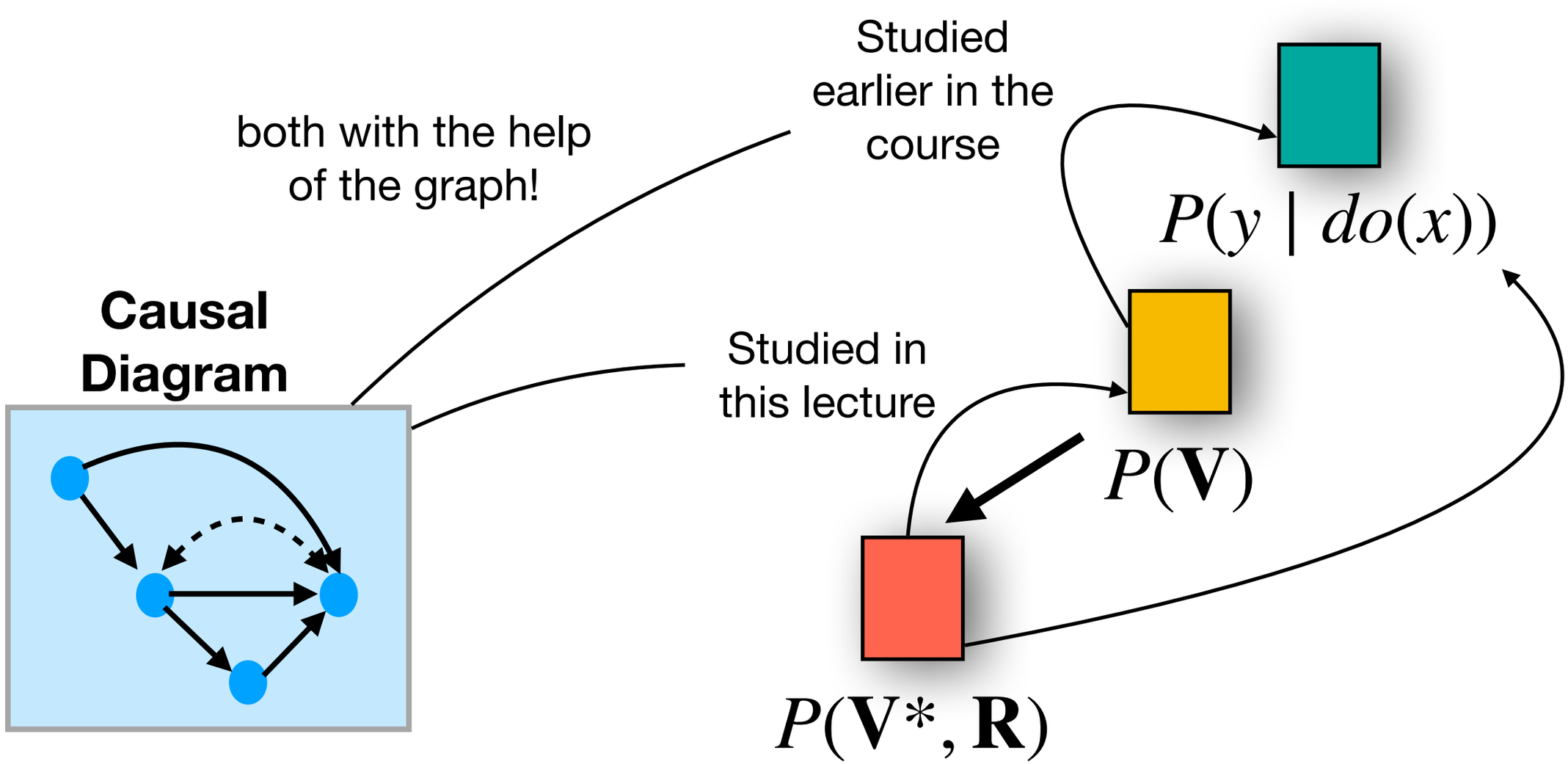
$$P(R_i \mid V) = P(R_i \mid \mathrm{pa}(R_i), R_{\mathrm{pa}^m(R_i)} = 0)$$

from which it follows that

$$P(R = 0 \mid V) = \prod_i P(R_i = 0 \mid \mathrm{pa}(R_i), R_{\mathrm{pa}^m(R_i)} = 0)$$

$P(y \mid do(x))$: m-Identification

- We now have some understanding about recovering $P(V)$ — but how about $P(y \mid do(x))$?



$P(y \mid do(x))$: m-Identification

Proposition. If the effect $P(y \mid do(x))$ is not identifiable from $G(\mathbf{V}), P(\mathbf{V})$ it is not identifiable from $G(\mathbf{V}, \mathbf{V}^*, \mathbf{R}), P(\mathbf{V}^*, \mathbf{R})$ either.

If we cannot identify our query — missingness will not help!

Proposition. If the effect $P(y \mid do(x))$ is identifiable from $G(\mathbf{V}), P(\mathbf{V})$ AND $P(\mathbf{V})$ identifiable from $G(\mathbf{V}, \mathbf{V}^*, \mathbf{R}), P(\mathbf{V}^*, \mathbf{R})$, then $P(y \mid do(x))$ is m-identifiable.

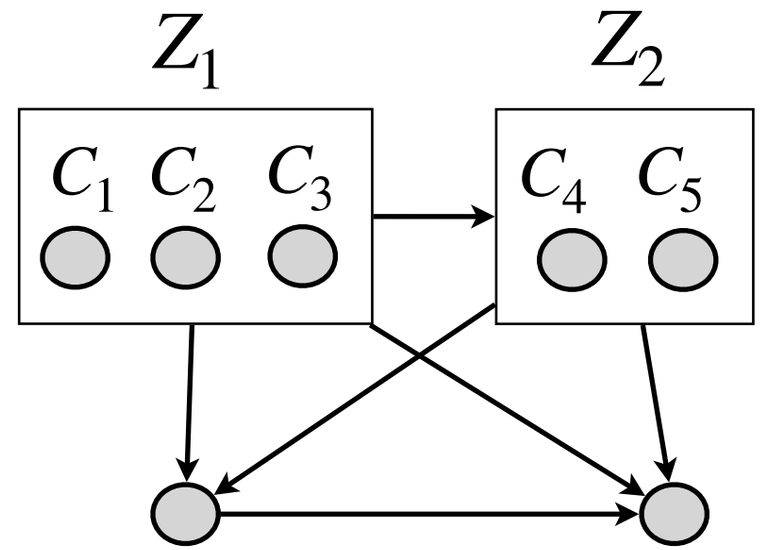
Chain the two steps we learned before.

Recap: VAHCS Study

- The VAHCS study followed a cohort of adolescents in Australia, recruited at age 14-15 and tracked through adolescence into young adulthood (up to age 20–21).
- The point of investigation was a causal question: how much preventing frequent cannabis use during adolescence (X = 1 for using > once per week) would change depression and anxiety in young adulthood (Y = standardized CIS-R score at age 20–21), focusing on female participants.
- Confounders were pre- and mid-adolescent social, behavioural, and mental health factors measured during adolescence: parental education, parental divorce/separation, antisocial behaviour, depression and anxiety, and frequent alcohol use, with incomplete data motivating the use of multiple imputation.

VAHCS m-graph: simplest setting

- Variables:
  - parental education C_1,
  - parental separation C_2,
  - antisocial behavior in adolescence C_3,
  - adolescent depression/anxiety C_4,
  - alcohol use in adolescence C_5,
  - cannabis use in adolescence X,
  - adulthood mental health score Y.



Z_1 **always observed,** $V_o = Z_1$

other variables partially missing $V_m = \{Z_2, X, Y\}$

Z_1, Z_2 **is back-door for** $X, Y \Longrightarrow E[y \mid do(x)] = \sum_{z_1, z_2} E[y \mid x, z_1, z_2] P(z_1, z_2)$

VAHCS m-graph: simplest setting

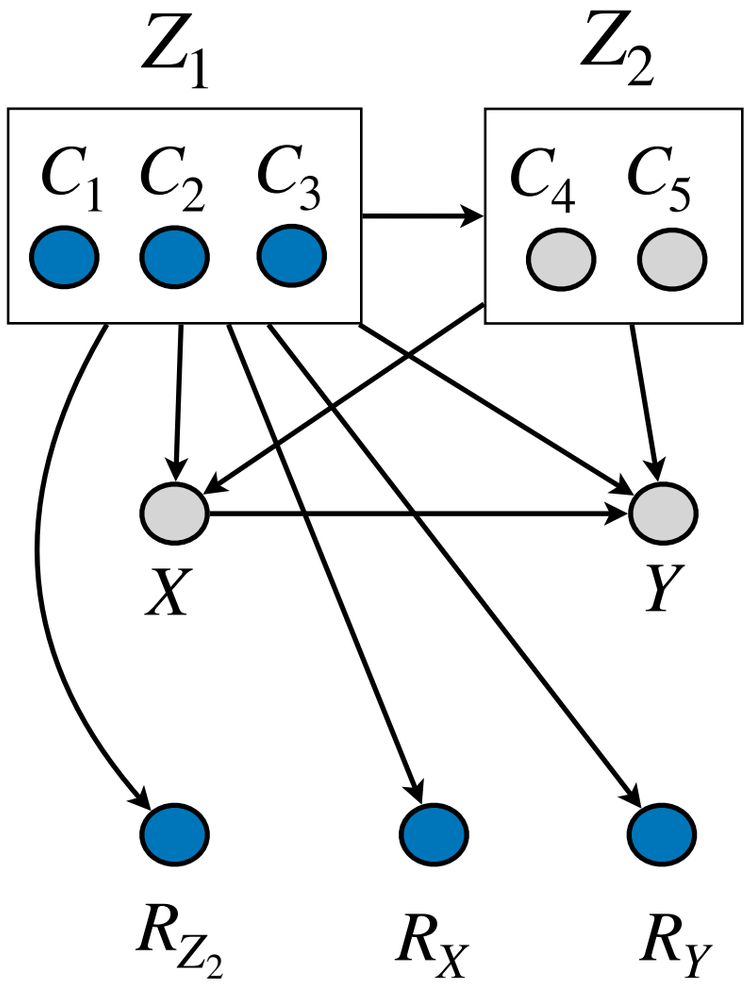
Q: is P(V) recoverable?

$$P(V) = P(V)\frac{P(V, R=0)}{P(V, R=0)}$$

$$= \frac{P(V, R=0)}{P(R=0 \mid V)}$$

$$= \frac{P(V \mid R=0)P(R=0)}{P(R=0 \mid V)}$$

$$= \frac{P(V \mid R=0)P(R=0)}{P(R=0 \mid Z_1)}$$



reweigh fully observed samples!

VAHCS m-graph: less simple setting

Q: is P(V) recoverable?

$$P(V) = \frac{P(V \mid R = 0)P(R = 0)}{P(R = 0 \mid V)}$$

$$P(R = 0 \mid V) = P(R_Y = 0 \mid V)P(R_X = 0 \mid V)P(R_{Z_2} = 0 \mid V)$$

$$= P(R_X = 0 \mid Z_1, Z_2)$$

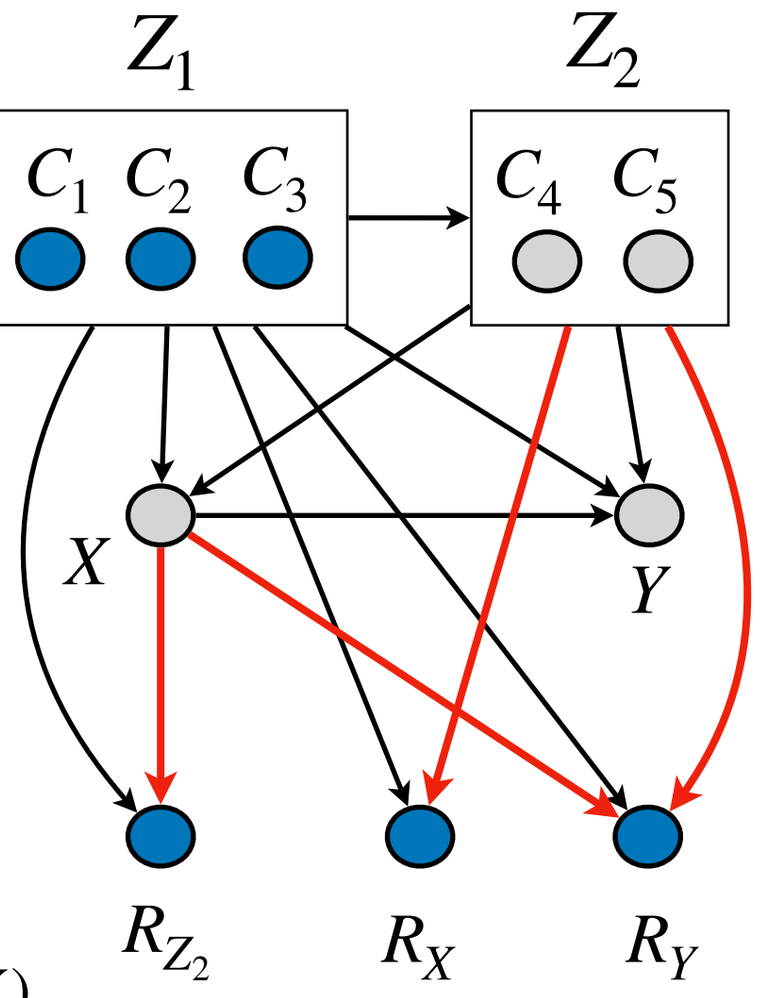
$$= P(R_X = 0 \mid Z_1, Z_2, R_{Z_2} = 0)$$

$$= P(R_Y = 0 \mid Z_1, Z_2, X)$$

$$= P(R_Y = 0 \mid Z_1, Z_2, X, R_X = 0, R_{Z_2} = 0)$$

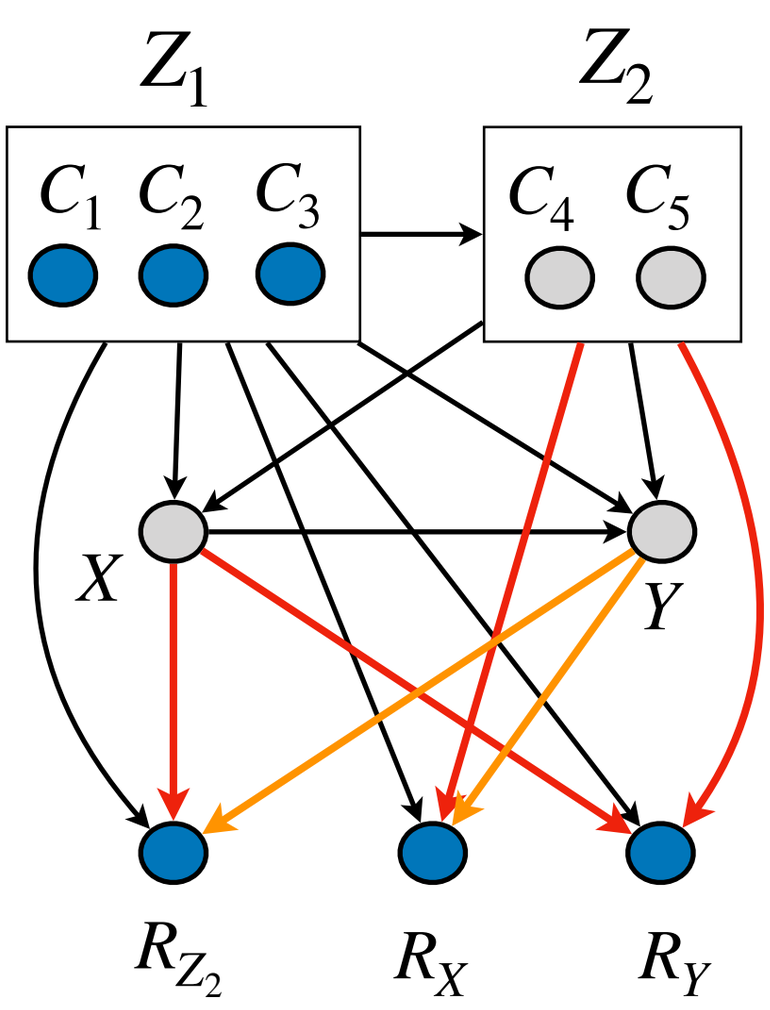
$$= P(R_{Z_2} = 0 \mid Z_1, X)$$

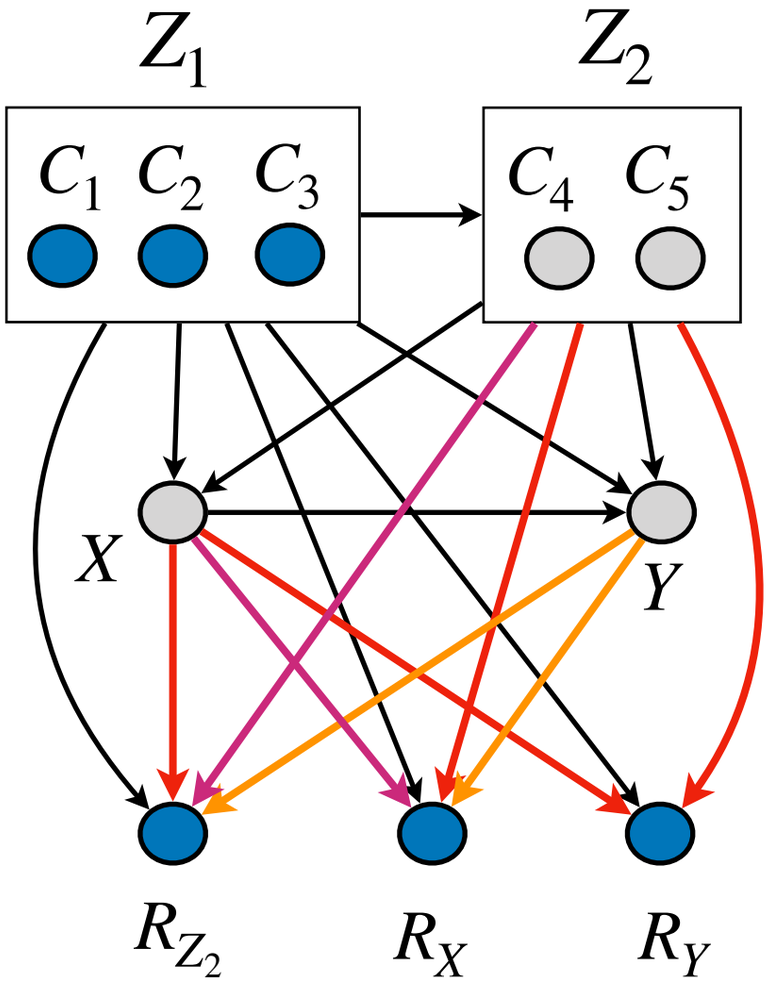
$$= P(R_{Z_2} = 0 \mid Z_1, X, R_X = 0)$$



P(V) again recoverable!

VAHCS m-graphs: FFT





Q: is P(V) recoverable?