Title: Modeling Latent Selection with Structural Causal Models

Speakers: Leihao Chen

Series: Quantum Foundations, Quantum Information

Date: September 16, 2024 - 1:45 PM

URL: https://pirsa.org/24090107

朗读此页内容 1 / 30

# Modeling Latent Selection with Structural Causal Models

Leihao Chen
University of Amsterdam
(Joint work with Onno Zoeter and Joris M. Mooij)

Causalworlds 2024

September 16, 2024

# Background I: Causal Inference

- Mathematical models for causal inference in current talk: acyclic Structural Causal Models (works for cyclic SCMs).

Correlation does not imply Causation. ($p(y \mid \mathrm{do}(x)) \neq p(y \mid x)$):

1. **Common Cause**
2. **Causal Cycle**
3. **Selection Bias**: conditioning on common effect induces spurious dependency (Berkson's paradox: "All handsome men are jerks?").

Bongers et al. (2021) studied cyclic SCMs with latent variables but no selection bias. The **goal of our work** is to consider selection bias.

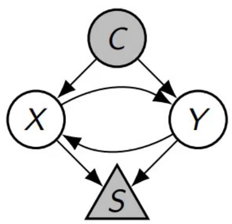



Figure 1: Three ways to induce dependency between $X$ and $Y$

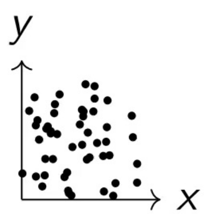



Figure 2: $X, Y \sim \mathrm{Uni}[0,1]$ and $X \perp\!\!\!\perp Y$.

# Background I: Causal Inference

- Mathematical models for causal inference in current talk: acyclic Structural Causal Models (works for cyclic SCMs).

Correlation does not imply Causation. ($p(y \mid \mathrm{do}(x)) \neq p(y \mid x)$):

1. **Common Cause**
2. **Causal Cycle**
3. **Selection Bias**: conditioning on common effect induces spurious dependency (Berkson's paradox: "All handsome men are jerks?").

Bongers et al. (2021) studied cyclic SCMs with latent variables but no selection bias. The **goal of our work** is to consider selection bias.

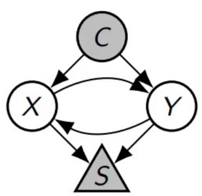



Figure 3: Three ways to induce dependency between $X$ and $Y$

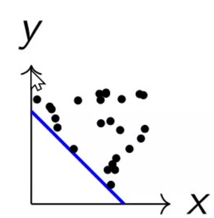


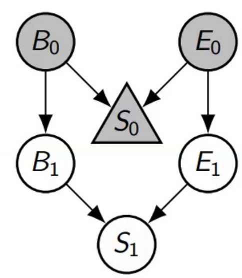

Figure 4: Select $S := X + Y > 0.8$

# Motivating Example: Car Mechanic Example

**Goal**: having an SCM on observed variables $(B_1, E_1, S_1)$ and performing causal reasoning based on it for subpopulation $S_0 = 0$. E.g., computing $\mathrm{P}(S_1 \mid \mathrm{do}(B_1 = 1))$ and $\mathrm{P}(S_1 \mid \mathrm{do}(E_1 = 1))$ to help with repairing cars.

- $B_0 \in \{0, 1\}$: battery works or not; $E_0 \in \{0, 1\}$: start engine works or not; $S_0 \in \{0, 1\}$: car starts or not.
- $B_0, E_0, S_0$ are measured in the morning; $B_1, E_1, S_1$ are measured in the afternoon.
- Only cars failed to start in the morning ($S_0 = 0$) were sent to car mechanic in the afternoon.

$$M : \begin{cases} U_B \sim \mathrm{Ber}(1 - \delta), \\ U_E \sim \mathrm{Ber}(1 - \epsilon), \\ B_0 = U_B, \;\; E_0 = U_E, \\ S_0 = B_0 \wedge E_0, \\ B_1 = B_0, \;\; E_1 = E_0, \\ S_1 = B_1 \wedge E_1, \end{cases}$$




$G(M)$

# Marginalization: Causal Model Abstraction

**Marginalization**: powerful tool for model abstracting (Bongers et al., 2021). Effectively abstract away latent details.

- **Preserving causal semantics**: The marginalized model $M_{\setminus L}$ has the same causal semantics (observations/interventional/counterfactual) as the original model $M$ on remaining variables $X_{V \setminus L}$.
- **Interact well with intervention and marginalization**: $(M_{\setminus L})_{\mathrm{do}(X_T = x_T)} \equiv (M_{\mathrm{do}(X_T = x_T)})_{\setminus L}$ and $(M_{\setminus L_1})_{\setminus L_2} \equiv (M_{\setminus L_2})_{\setminus L_1} \equiv M_{\setminus L_1 \cup L_2}$.
- **Preserving model class**: $M$ is linear/acyclic/simple $\Longrightarrow M_{\setminus L}$ is is linear/acyclic/simple.

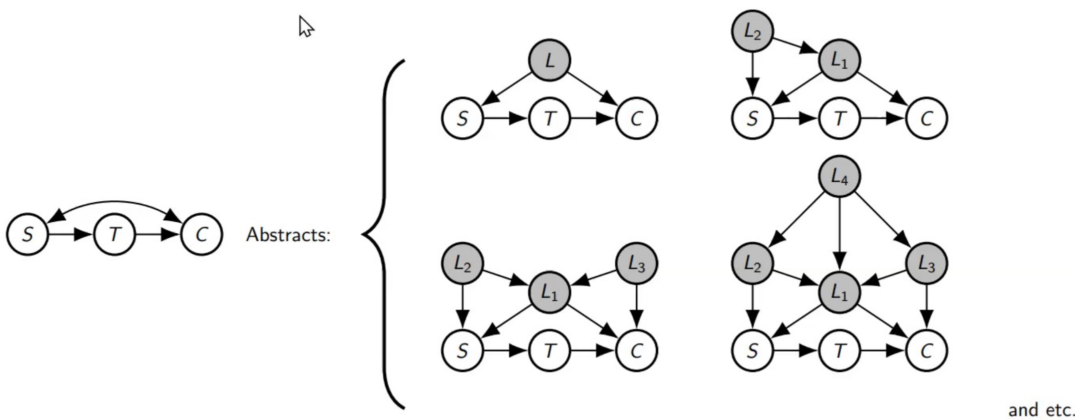
▶ **SCM marginalization and graph marginalization interact well**: $G(M_{\setminus L})$ is a subgraph of $G(M)_{\setminus L}$.

$$
\begin{array}{ccc}
X_A \underset{\mathrm{P}_M(X_V)}{\perp\!\!\!\perp} X_B \mid X_C & \underset{d/\sigma\text{-Faithful}}{\overset{d/\sigma\text{-Markov}}{\rightleftarrows}} & A \overset{d/\sigma}{\underset{G(M)}{\perp}} B \mid C \\
\Updownarrow & & \Updownarrow \\
X_A \underset{\mathrm{P}_{M_{\setminus L}}(X_{V \setminus L})}{\perp\!\!\!\perp} X_B \mid X_C & \underset{d/\sigma\text{-Faithful}}{\overset{d/\sigma\text{-Markov}}{\rightleftarrows}} & A \overset{d/\sigma}{\underset{G(M)_{\setminus L}}{\perp}} B \mid C
\end{array}
$$

# Marginalization: Causal Model Abstraction

- The first model with only observed variables can represent **infinitely many** models with latent variables.
- The **same** $d$-**separation** and the **same identification result** (ID-algorithm):

$$P(C = c \mid do(S = s)) = \sum_t P(C = c \mid T = t) P(T = t \mid S = s)$$



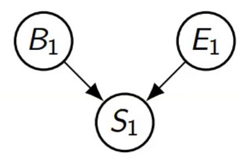

(!) Spoiler alert: Bidirected edges can also represent latent selection bias.

# Marginalized Model

1. **Graphically**: $B_1$ and $E_1$ are separated even if they are dependent given $S_0 = 0$;
2. **Causal Semantics**: inconsistent with the original model under subpopulation $(S_0 = 0)$

$$\mathrm{P}_{M^*}(B_1, E_1, S_1) \neq \mathrm{P}_M(B_1, E_1, S_1 \mid S_0 = 0)$$
$$\mathrm{P}_{M^*}(S_1 = 1 \mid \mathrm{do}(B_1 = 1)) \neq \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(B_1 = 1), S_0 = 0)$$
$$\mathrm{P}_{M^*}(S_1 = 1 \mid \mathrm{do}(E_1 = 1)) \neq \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(E_1 = 1), S_0 = 0)$$

$$M^* : \begin{cases} U_B \sim \mathrm{Ber}(1 - \delta), \\ U_E \sim \mathrm{Ber}(1 - \epsilon), \\ B_1 = U_B, \ E_1 = U_E, \\ S_1 = B_1 \wedge E_1, \end{cases}$$




$G(M^*)$

# Wait a Minute: What we Shall Achieve in the Talk

(!) Marginalization cannot deal with latent selection bias.

(?) Marginalization effectively abstract away the latent common cause, can we effectively abstract away latent selection bias similarly?

Transformations $(M, X_S \in \mathcal{S}) \mapsto M_{|X_S \in \mathcal{S}}$ and $(G, S) \mapsto G_{|S}$? **Effectively abstract away** latent selection bias...:

- The conditioned SCM $M_{|X_S \in \mathcal{S}}$ encodes the correct **causal semantics** (observational, interventional and counterfactual) under the subpopulation;
- Interact well with other operations on SCMs/DMGs (mar/int/cond);
- Preserve important model classes (lin/acyc/simp);
- One can read off **causal information** from **causal graphs**.

$$X_A \underset{\mathrm{P}_M(X_V)}{\perp\!\!\!\perp} X_B \mid X_C, X_S \in \mathcal{S} \quad \underset{?}{\overset{?}{\rightleftarrows}} \quad A \underset{G(M)}{\overset{d/\sigma}{\perp}} B \mid C \cup S$$

$$? \Updownarrow \qquad\qquad\qquad\qquad ? \Updownarrow$$

$$X_A \underset{\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O)}{\perp\!\!\!\perp} X_B \mid X_C \quad \underset{?}{\overset{?}{\rightleftarrows}} \quad A \underset{G(M)_{|S}}{\overset{d/\sigma}{\perp}} B \mid C$$

# Correct Surrogate Model

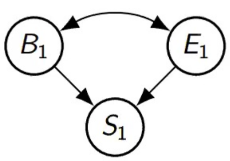
1. **Graphically**: $B_1$ and $E_1$ are connected;
2. **Causal Semantics**: consistent with the original model under subpopulation ($S_0 = 0$)

$$\mathrm{P}_{\tilde{M}}(B_1, E_1, S_1) = \mathrm{P}_M(B_1, E_1, S_1 \mid S_0 = 0)$$
$$\mathrm{P}_{\tilde{M}}(S_1 = 1 \mid \mathrm{do}(B_1 = 1)) = \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(B_1 = 1), S_0 = 0)$$
$$\mathrm{P}_{\tilde{M}}(S_1 = 1 \mid \mathrm{do}(E_1 = 1)) = \mathrm{P}_M(S_1 = 1 \mid \mathrm{do}(E_1 = 1), S_0 = 0).$$

$$\tilde{M} : \begin{cases} (U_B, U_E) \sim \tilde{\mathrm{P}}(U_B, U_E) \\ B_1 = U_B, E_1 = U_E, S_1 = B_1 \wedge E_1. \end{cases}$$

$\tilde{\mathrm{P}}(U_B, U_E) = \mathrm{P}_M(U_B, U_E \mid S_0 = 0)$ :

| $\tilde{\mathrm{P}}(U_B, U_E)$ | $U_E = 0$ | $U_E = 1$ |
|---|---|---|
| $U_B = 0$ | $\frac{\delta\epsilon}{\delta+(1-\delta)\epsilon}$ | $\frac{\delta(1-\epsilon)}{\delta+(1-\delta)\epsilon}$ |
| $U_B = 1$ | $\frac{(1-\delta)\epsilon}{\delta+(1-\delta)\epsilon}$ | 0 |




$G(\tilde{M})$

# Some Thoughts about Car Mechanic Example

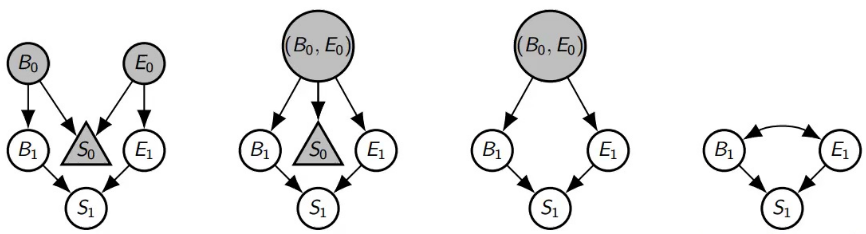
1. (!) Marginalization cannot deal with latent selection bias.
2. (!!!) **Bidirected edges** cannot only represent latent common cause but also latent selection bias.
3. $\exists\ \tilde{M}$ representing $(M, S_0 = 0)$ and leading to correct predictions.

Effectively abstract away irrelevant latent modeling details:

1. the latent variables $B_0$, $E_0$ and $S_0$,
2. their causal mechanisms, and
3. the filtering step on $S_0 = 0$.

Note that we could also have obtained the model $\tilde{M}$ directly from $M$, by

1. replacing $\mathrm{P}_M(U_B, U_E)$ by $\mathrm{P}_M(U_B, U_E \mid S_0 = 0)$,
2. marginalizing out $B_0, E_0$ and $S_0$.



# Structural Causal Model

## Definition (Bongers et al. (2021))

A **Structural Causal Model (SCM)** is a tuple $M = (V, W, \mathcal{X}, \mathrm{P}, f)$ such that

- $V, W$ are disjoint finite sets of labels for the, the **endogenous variables** and the **exogenous random variables**, respectively;
- the **state space** $\mathcal{X} = \prod_{\dot\cup W \dot\cup V} \mathcal{X}_i$ is a product of standard measurable spaces $\mathcal{X}_i$;
- the **exogenous distribution** $\mathrm{P}$ is a probability distribution on $\mathcal{X}_W$ that factorizes as a product $\mathrm{P} = \bigotimes_{w \in W} \mathrm{P}(X_w)$ of probability distributions $\mathrm{P}(X_w)$ on $\mathcal{X}_w$;
- the **causal mechanism** is specified by the measurable function $f : \mathcal{X} \to \mathcal{X}_V$.

**Notation**:

$$\mathrm{P}_M(X_{V \setminus S} \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}) := \frac{\mathrm{P}_M(X_{V \setminus S}, X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T))}{\mathrm{P}_M(X_S \in \mathcal{S} \mid \mathrm{do}(X_T = x_T))}$$
$$\neq \mathrm{P}_M(X_{V \setminus S}(x_T) \mid X_S \in \mathcal{S}).$$

# Main Definition: Conditioning Operation on SCMs

Main Definition: $\mathrm{P}_M(X_S \in \mathcal{S}) > 0$. Let $g : \mathcal{X}_W \to \mathcal{X}_V$ and $g^S : \mathcal{X}_{V \setminus S} \times \mathcal{X}_W \to \mathcal{X}_S$ be the (essentially unique) solution function of $M$ w.r.t. $V$ and $S$ respectively. We define the **conditioned SCM** $M_{|X_S \in \mathcal{S}} := \left(\hat{V}, \hat{W}, \hat{\mathcal{X}}, \hat{\mathrm{P}}, \hat{f}\right)$ by:

- $\hat{V} := V \setminus S$;
- $\hat{W} := \{\hat{w}_1, \ldots, \hat{w}_n\}$ where $\hat{w}_i := \{H_i\}$ for $i = 1, \ldots, n$ and $\mathcal{H} = \{H_i\}_{i=1}^n$ is the largest element in $(\mathfrak{P}, \vee)$;
- $\hat{\mathcal{X}} := \mathcal{X}_{\hat{V}} \times \hat{\mathcal{X}}_{\hat{W}} := \mathcal{X}_{\hat{V}} \times \bigtimes_{i=1}^n \mathcal{X}_{\hat{w}_i}$, where $\mathcal{X}_{\hat{w}_i} := \mathcal{X}_{H_i}$;
- $\hat{\mathrm{P}} := \bigotimes_{i=1}^n \hat{\mathrm{P}}(X_{\hat{w}_i})$, where $\hat{\mathrm{P}}(X_{\hat{w}_i}) := \mathrm{P}_M(X_{H_i} \mid X_S \in \mathcal{S})$;
- $\hat{f}(x_{\hat{V}}, x_{\hat{W}}) := f_{\hat{V}}(x_{\hat{V}}, g^S(x_{\hat{V}}, x_{H_1}, \ldots, x_{H_n}), x_{H_1}, \ldots, x_{H_n})$.

$\mathfrak{P} = \{\mathcal{J} = \{J_1, \ldots, J_n\} : \mathcal{J}$ is a partition of $W$ s.t. $g_S^{-1}(\mathcal{S}) \stackrel{p}{=} \bigtimes_{i=1}^n \mathrm{pr}_{\mathcal{X}_{J_i}}(g_S^{-1}(\mathcal{S}))\}$. Then $(\mathfrak{P}, \vee)$ is a finite join semi-lattice where $\mathcal{I} \vee \mathcal{J} := \{I \cap J : I \in \mathcal{I}$ and $J \in \mathcal{J}\}$.

# Main Result: Causal Semantics of Conditioned SCMs

**Main Result**: Write $O := V \setminus S$. Then we have:

1. **Observational**: $\mathrm{P}_{M|X_S \in \mathcal{S}}(X_O) = \mathrm{P}_M(X_O \mid X_S \in \mathcal{S})$.
2. **Interventional**: $T = T_1 \dot{\cup} T_2 \subseteq O$, $T_1 \subseteq O \setminus \mathrm{Anc}_{G(M)}(S)$ and $T_2 \subseteq \mathrm{Anc}_{G(M)}(S)$. For $x_T \in \mathcal{X}_T$,

$$\mathrm{P}_{M|X_S \in \mathcal{S}}\left(X_{O \setminus T} \mid \mathrm{do}(X_T = x_T)\right) = \mathrm{P}_M\left(X_{O \setminus T}(x_{T_2}) \mid \mathrm{do}(X_T = x_{T_1}), X_S \in \mathcal{S}\right).$$

3. **Counterfactual via twinning**: $T = T_1 \dot{\cup} T_2 \subseteq O$, $T_1 \subseteq V \setminus \mathrm{Anc}_{G(M)}(S)$ and $T_2 \subseteq \mathrm{Anc}_{G(M)}(S) \setminus S$. $\tilde{T} = T_3 \dot{\cup} T_4 \subseteq V'$, $T_3 \subseteq (V \setminus \mathrm{Anc}_{G(M)}(S))'$ and $T_4 \subseteq (\mathrm{Anc}_{G(M)}(S) \setminus S)'$. For any $x_T \in \mathcal{X}_T$ and $x_{\tilde{T}} \in \mathcal{X}_{\tilde{T}}$

$$\mathrm{P}_{\left(M_{|X_S \in \mathcal{S}}\right)^{\mathrm{twin}}}(X_{(O \cup O') \setminus (T \cup \tilde{T})} \mid \mathrm{do}(X_T = x_T, X_{\tilde{T}} = x_{\tilde{T}}))$$
$$= \mathrm{P}_{M^{\mathrm{twin}}}(X_{O \setminus T}(x_{T_2}), X_{O' \setminus \tilde{T}}(x_{T_4}) \mid \mathrm{do}(X_{T_1} = x_{T_1}, X_{T_3} = x_{T_3}), X_S \in \mathcal{S}).$$

4. **Potential outcome**: Let $T_i \subseteq O$ and $x_{T_i} \in \mathcal{X}_{T_i}$ for $i = 1, \dots, n$. Then we have

$$\mathrm{P}_{M|X_S \in \mathcal{S}}(\{X_{O \setminus T_i}(x_{T_i})\}_{1 \le i \le n}) = \mathrm{P}_M(\{X_{O \setminus T_i}(x_{T_i})\}_{1 \le i \le n} \mid X_S \in \mathcal{S}).$$

# Simple SCMs are not Flexible Enough for Modeling all Conditional Interventional Distributions

**Question**: When $T \cap \mathrm{Anc}_{G(M)}(S) \neq \emptyset$,

$$\mathrm{P}_{M_{|X_S \in \mathcal{S}}}(X_O \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_O(x_T) \mid X_S \in \mathcal{S})$$
$$\neq \mathrm{P}_M(X_O \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S}).$$

Can we always find an SCM $\tilde{M}$ such that

$$\mathrm{P}_{\tilde{M}}(X_O \mid \mathrm{do}(X_T = x_T)) = \mathrm{P}_M(X_O \mid \mathrm{do}(X_T = x_T), X_S \in \mathcal{S})?$$

- The answer is No.
- One can prove it via natural bound of simple SCMs.

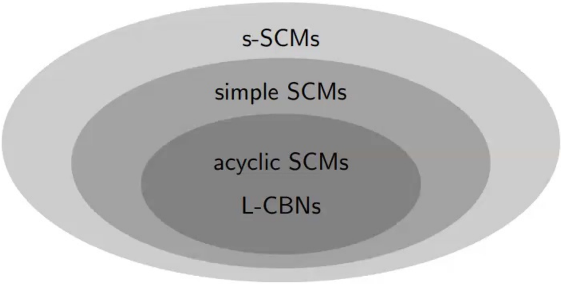



Figure 5: Venn diagram for different causal modeling classes.

# Properties of Conditioning Operation on SCMs

1. **Preserving model class**: $M$ linear/acyclic/simple $\Longrightarrow M_{|X_S \in \mathcal{S}}$ linear/acyclic/simple;
2. **Commuting with intervention**: $T$ non-ancestor of $S \Longrightarrow$ $(M_{|X_S \in \mathcal{S}})_{\mathrm{do}(X_T = x_T)} \equiv (M_{\mathrm{do}(X_T=x_T)})_{|X_S \in \mathcal{S}}$;
3. **Commuting with marginalization**: $(M_{|X_S \in \mathcal{S}})_{\setminus L} \equiv (M_{\setminus L})_{|X_S \in \mathcal{S}}$;
4. **Commuting with conditioning**:$(M_{|X_{S_1} \in \mathcal{S}_1})_{|X_{S_2} \in \mathcal{X}_{S_2}}$, $(M_{|X_{S_2} \in \mathcal{S}_2})_{|X_{S_1} \in \mathcal{X}_{S_1}}$ and $M_{|X_{S_1 \cup S_2} \in \mathcal{S}_1 \times \mathcal{S}_2}$ are counterfactually equivalent.

Remark
1. In item 2, the assumption $T \cap \mathrm{Anc}_{G(M)}(S)$ cannot be relaxed in general.
2. The inelegance of item 4 comes from the fundamental definition of SCMs.

# Application: The Reichenbach Principle of Common Cause

- Two variables are dependent, then one must cause the other or the variables must have a common cause or any combination of these three possibilities (assume no latent selection).
- Assume $M$ is an SCM with two dependent observed endogenous variables $X$ and $Y$. Markov property implies $X \to Y$, $X \leftarrow Y$ or $X \leftrightarrow Y$ in $G(M)$.
- There exist infinitely many SCMs $M^i$, $i \in I$, s.t. $(M^i_{\setminus L_i})_{|X_{S_i} \in S_i} = M$ where $L_i$ is a set of latent variables of $M^i$ and $X_{S_i} \in S_i$ is the latent selection in $M^i$.
- If two variables are dependent, then one must cause the other or the variables must have a common cause or subject to latent selection (or any combination of these four possibilities).

# Application: The Reichenbach Principle of Common Cause

- Two variables are dependent, then one must cause the other or the variables must have a common cause or any combination of these three possibilities (assume no latent selection).
- Assume $M$ is an SCM with two dependent observed endogenous variables $X$ and $Y$. Markov property implies $X \longrightarrow Y$, $X \longleftarrow Y$ or $X \longleftrightarrow Y$ in $G(M)$.
- There exist infinitely many SCMs $M^i$, $i \in I$, s.t. $(M^i_{\setminus L_i})_{|X_{S_i} \in \mathcal{S}_i} = M$ where $L_i$ is a set of latent variables of $M^i$ and $X_{S_i} \in \mathcal{S}_i$ is the latent selection in $M^i$.
- If two variables are dependent, then one must cause the other or the variables must have a common cause or subject to latent selection (or any combination of these four possibilities).

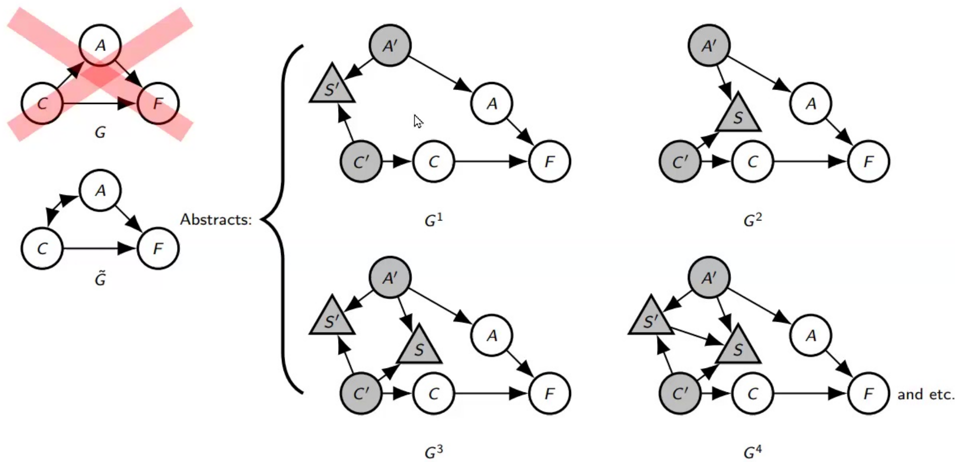
# Application: Causal Modeling of Covid Example

One workflow of Causal Inference:

(1) Ask causal queries;
(2) **Build a causal model**;
(3) The causal model outputs a target estimand;
(4) Use data to estimate the estimand.

- **Causal query**: "What would be the effect on fatality of changing from China to Italy" (von Kügelgen et al., 2021).
- **Estimand**: Total causal effect: $\mathbb{E}[F \mid \mathrm{do}(C = c)] - \mathbb{E}[F \mid \mathrm{do}(C = c')]$. The identification results based on $G$ and $\tilde{G}$ are clearly different.

(!) Bidirected edges can represent latent selection bias.



# Causal Modeling and Other Applications

Causal Modeling:

1. Starting with a complete graph;
2. Using data and prior knowledge to delete edges:
   - No directed causal effect: delete directed edges;
   - No latent common cause or latent selection bias: delete bidirected edges. (In many cases, we know the existence of "non-causal" dependency between two variables but do not know whether it comes from common cause or selection bias.)

Other Applications:

1. Do-calculus;
2. ID-algorithm;
3. Mediation analysis and fairness analysis;
4. Causal discovery and ect.

## Take-home Message

- Structural causal model is a class of causal models that mathematically model causal relationships among variables (common cause, selection bias, causal cycle).
- Selection bias is ubiquitous in many real-world data and dealing with it naively may lead to misleading and counterintuitive results.
- By introducing a conditioning operation on SCM, one can abstract away latent selection, which streamlines causal modeling, causal reasoning and causal model discovery under latent selection bias.

(!) **Bidirected edges** can not only represent latent common cause but also latent selection bias.

Thank you for your attention!
Questions or Comments?