General Identifiability with Arbitrary Surrogate Experiments

Sanghack Lee
with Juan Correa and Elias Bareinboim

Columbia University

AAAI 2020
(presented at UAI 2019)
Overview

• **Causality** & Everyday Life, Science, Artificial Intelligence 🤖.
Overview

- **Causality** & Everyday Life, Science, Artificial Intelligence 🤖.

- **Causal Effect Identifiability** concerns about precisely determining the effect of intervention given information (e.g., causal assumptions and an observational data).
Overview

- **Causality** & Everyday Life, Science, Artificial Intelligence 🤖.

- **Causal Effect Identifiability** concerns about precisely determining the effect of intervention given information (e.g., causal assumptions and an observational data).

- **General Identifiability** considers identifying a causal effect given an arbitrary combination of observational and experimental data.
Overview

- **Causality** & Everyday Life, Science, Artificial Intelligence 🤖.

- **Causal Effect Identifiability** concerns about precisely determining the effect of intervention given information (e.g., causal assumptions and an observational data).

- **General Identifiability** considers identifying a causal effect given an arbitrary combination of observational and experimental data.

- We provided a graphical **necessary and sufficient condition** under which a causal effect of interest can be estimable. We devised a **sound and complete algorithm** which outputs a formula for the causal effect made with probabilities obtained from available data.
Understanding Data

unknown
real-world
Understanding Data

unknown

real-world

observation

X, Y, Z

experiment

unperformed

experiment
Understanding Data

unknown
real-world

observation
observation

X,Y,Z

intervention

experiment

experiment
unperformed
unperformed
Understanding Data

- **real-world**
  - **unknown**
  - **observation** $X,Y,Z$
  - **experiment** unperformed
  - **experiment** unperformed
Understanding Data

real-world

X Z Y
Understanding Data

real-world

X
Z
Y
Understanding Data
Understanding Data

intervention

real-world

X → Z → Y
Causal Framework

SCM provides an abstraction of causality in the real-world.

**Definition (Structural Causal Model (Pearl))**

SCM $\mathcal{M}$ is a 4-tuple $\langle U, V, F, P(U) \rangle$

- $U = \{U_1, \ldots, U_m\}$ are exogenous variables;
- $V = \{V_1, \ldots, V_n\}$ are endogenous variables;
- $F = \{f_1, \ldots, f_n\}$ are functions determining $V$,
  
  $$v_i \leftarrow f_i(pa^i, u^i)$$

  where $PA^i \subseteq V \setminus \{V_i\}$, $U^i \subseteq U$; and
- $P(U)$ is a joint distribution over $U$. 
Causal Framework

SCM provides an abstraction of causality in the real-world.

**Definition (Structural Causal Model (Pearl))**

SCM $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$

- $\mathbf{U} = \{U_1, \ldots, U_m\}$ are **exogenous** variables;
- $\mathbf{V} = \{V_1, \ldots, V_n\}$ are **endogenous** variables;
- $\mathbf{F} = \{f_1, \ldots, f_n\}$ are **functions** determining $\mathbf{V}$,

  \[ v_i \leftarrow f_i(pa^i, u^i) \]

  where $PA^i \subseteq \mathbf{V} \setminus \{V_i\}$, $U^i \subseteq \mathbf{U}$; and

- $P(\mathbf{U})$ is a joint **distribution** over $\mathbf{U}$
Causal Framework

SCM provides an abstraction of causality in the real-world.

**Definition (Structural Causal Model (Pearl))**

SCM \( \mathcal{M} \) is a 4-tuple \( \langle U, V, F, P(U) \rangle \)

- \( U = \{U_1, \ldots, U_m\} \) are **exogenous** variables;
- \( V = \{V_1, \ldots, V_n\} \) are **endogenous** variables;
- \( F = \{f_1, \ldots, f_n\} \) are **functions** determining \( V \),
  
  \[ v_i \leftarrow f_i(pa^i, u^i) \]

  where \( PA^i \subseteq V \setminus \{V_i\}, U^i \subseteq U \); and
- \( P(U) \) is a joint **distribution** over \( U \)
Causal Framework

SCM provides an abstraction of causality in the real-world.

Definition (Structural Causal Model (Pearl))

SCM $\mathcal{M}$ is a 4-tuple $\langle U, V, F, P(U) \rangle$

- $U = \{U_1, \ldots, U_m\}$ are **exogenous** variables;
- $V = \{V_1, \ldots, V_n\}$ are **endogenous** variables;
- $F = \{f_1, \ldots, f_n\}$ are **functions** determining $V$,
  
  $$v_i \leftarrow f_i(pa^i, u^i)$$

  where $PA^i \subseteq V \setminus \{V_i\}$, $U^i \subseteq U$; and
- $P(U)$ is a joint **distribution** over $U$.
Causal Framework

SCM provides an abstraction of causality in the real-world.

Definition (Structural Causal Model (Pearl))

SCM $\mathcal{M}$ is a 4-tuple $\langle U, V, F, P(U) \rangle$

- $U = \{U_1, \ldots, U_m\}$ are **exogenous** variables;
- $V = \{V_1, \ldots, V_n\}$ are **endogenous** variables;
- $F = \{f_1, \ldots, f_n\}$ are **functions** determining $V$,
  \[ v_i \leftarrow f_i(\text{pa}^i, u^i) \]
  where $\text{pa}^i \subseteq V \setminus \{V_i\}$, $u^i \subseteq U$; and
- $P(U)$ is a joint **distribution** over $U$
Causal Framework

SCM $\mathcal{M}$

$\langle U, V, F, P(U) \rangle$

unknown

$P(V)$

$P_{X_1}(V)$

$P_{X_2}(V)$

$P_{X_2, X_4}(V)$

$do(\emptyset)$

$do(x_1)$

$do(x_2)$

$do(x_2, x_4)$
Causal Framework

\[ P(V) \to do(\emptyset) \to do(x_1) \to do(x_2) \to do(x_2, x_4) \to P_{X_2, X_4}(V) \]

causal relationships

unknown

\[ P(X_1)(V) \to P(X_2)(V) \]

[1873x39]6

Causal Framework

Do not hallucinate.
(Classic) Causal Effect Identifiability

1. Query $Q$
   \[ P_x(y) = P(y|do(x)) \]

2. Causal Diagram $\mathcal{G}$

3. Data
   \[ P(V) \]

A non-parametric assumption: no assumption on $P$
(Classic) Causal Effect Identifiability

1. Query $Q$
   
   $P_x(y) = P(y|do(x))$

2. Causal Diagram $\mathcal{G}$

3. Data $P(V)$

Causal Inference Engine

Formula $f_\mathcal{G}$ s.t.

$P_x(y) = f_\mathcal{G}(P(V))$

No evidence

A non-parametric assumption: no assumption on $P$
(Classic) Causal Effect Identifiability

1. Query $Q$
   
   $P_x(y) = P(y|do(x))$

2. Causal Diagram $G$

3. Data
   
   arbitrary combinations?

Causal Inference Engine

formula $f_G$ s.t.

$P_x(y) = f_G(P(V))$

Yes

No

evidence

a non-parametric assumption: no assumption on $P$
g-identifiability
Definition (g-Identifiability)

Let $\mathcal{Z} = \{Z_i\}_{i=1}^m$ be a collection of sets of variables. $P_x(y)$ is said to be g-identifiable from $\mathbb{P}$ in $\mathcal{G}$, if $P_x(y)$ is uniquely computable from distributions

$$\mathbb{P} = \{ P(V \mid do(Z_i)) \}_{z_i \in \mathcal{Z}}$$

in any causal model which induces $\mathcal{G}$. 
Definition (g-Identifiability)

Let $\mathbb{Z} = \{Z_i\}_{i=1}^m$ be a collection of sets of variables. $P_x(y)$ is said to be **g-identifiable** from $\mathbb{P}$ in $\mathcal{G}$, if $P_x(y)$ is uniquely computable from distributions

$$\mathbb{P} = \{ P(V \mid \text{do}(Z_i)) \}_{Z_i \in \mathbb{Z}}$$

in any causal model which induces $\mathcal{G}$. 
Related Work: *-Identifiability

Causal identifiability has been studied in the literature:

**ID** An observational distribution [TP’02, SP’06, HV’06]:

\[ P(V) \]
Related Work: *-Identifiability

Causal identifiability has been studied in the literature:

**ID** An observational distribution [TP’02, SP’06, HV’06]:

\[ P(V) \]

**zID** A set of manipulable variables [BP’12]:

\[ \{ P_{Z'}(V) \}_{Z' \subseteq z} \]
Related Work: *-Identifiability

**Causal identifiability** has been studied in the literature:

- **ID** An observational distribution [TP’02, SP’06, HV’06]:
  \[ P(V) \]

- **zID** A set of manipulable variables [BP’12]:
  \[ \{ P_{z_i}(V) \}_{z_i \subseteq z} \]

- **mzID** A collection of manipulable variables [BP’14]:
  \[ \{ \{ P_{z_i}(V) \}_{z_i \subseteq z_i} \}_{i=1}^{m} \]
Related Work: *-Identifiability

Causal identifiability has been studied in the literature:

**ID**  An observational distribution [TP’02, SP’06, HV’06]:

\[ P(V) \]

**zID**  A set of manipulable variables [BP’12]:

\[ \{ P_{Z'_i}(V) \}_{z'_i \subseteq z} \]

**mzID**  A collection of manipulable variables [BP’14]:

\[ \{ \{ P_{Z'_i}(V) \}_{z'_i \subseteq z_i} \}_{i=1}^m \]

**gID**  A collection of arbitrary experiments [LCB’19]:

\[ \{ P_{Z_i}(V) \}_{z_i \in \mathcal{Z}} \]
Example: Drug-Drug Interactions

(a)

$Y$ cardiovascular disease; $B$ blood pressure; $X_1$ taking an antihypertensive drug; and $X_2$ the use of an anti-diabetic drug.
Example: Drug-Drug Interactions

\[ P_{x_1,x_2}(y) \iff \{ P_{x_1}(v), P_{x_2}(v) \} \]

\( Y \) cardiovascular disease; \( B \) blood pressure; \( X_1 \) taking an antihypertensive drug; and \( X_2 \) the use of an anti-diabetic drug.

**Goal:** assess the effect of prescribing both treatments (\( \bullet \bullet \)) on the risk of cardiovascular diseases from individual drug experiments, either (\( \bullet \)) or (\( \bullet \)).
Example: Drug-Drug Interactions

\[
P_{X_1, X_2}(y) = \sum_b P_{X_2}(y|b) P_{X_1}(b)
\]

\(Y\) cardiovascular disease; \(B\) blood pressure; \(X_1\) taking an antihypertensive drug; and \(X_2\) the use of an anti-diabetic drug.

Goal: assess the effect of prescribing both treatments (\(\bullet \bullet\)) on the risk of cardiovascular diseases from individual drug experiments, either \(\bullet\) or \(\bigcirc\).
Example: Drug-Drug Interactions

Y cardiovascular disease; B blood pressure; $X_1$ taking an antihypertensive drug; and $X_2$ the use of an anti-diabetic drug.

Goal: assess the effect of prescribing both treatments (●●) on the risk of cardiovascular diseases from individual drug experiments, either ● or ●.
Example: Drug-Drug Interactions

\[ (a) \quad \checkmark \quad \quad (b) \quad \checkmark \quad \quad (c) \quad \times \quad \quad (d) \quad \times \]

\[ \begin{align*} \text{Goal:} \quad & \text{assess the effect of prescribing both treatments (σ⊙) on the risk of} \\
& \text{cardiovascular diseases from individual drug experiments, either σ or ⊙.} \end{align*} \]
g-identifiability
– a sound algorithm
Algorithm for gID

We developed a two-phase algorithm w/ probability axioms & do-calculus:

1. A given query is modified, and factorized into subqueries;
2. Each subquery is identified by one of the available distributions.
   * It FAILs only if there exists a subquery that cannot be identified by any of the available distributions.

$$Q \ P_x(y)$$
Algorithm for gID

We developed a two-phase algorithm w/ probability axioms & do-calculus:

1. A given query is modified, and factorized into subqueries;
2. Each subquery is identified by one of the available distributions.
* It FAILs only if there exists a subquery that cannot be identified by any of the available distributions.
Algorithm for gID

We developed a two-phase algorithm w/ probability axioms & do-calculus:

1. A given query is modified, and factorized into subqueries;

2. Each subquery is identified by one of the available distributions.

* It FAILs only if there exists a subquery that cannot be identified by any of the available distributions.
Algorithm for gID

We developed a two-phase algorithm w/ probability axioms & do-calculus:

1. A given query is modified, and factorized into subqueries;
2. Each subquery is identified by one of the available distributions.

* It FAILs only if there exists a subquery that cannot be identified by any of the available distributions.
visualization of the 1st phase

\[ X_1 \quad X_2 \quad X_3 \]
\[ Y_1 \quad Y_2 \quad Y_3 \]
visualization of the 1\textsuperscript{st} phase
visualization of the 1st phase
visualization of the 1\textsuperscript{st} phase
visualization of the 1st phase
visualization of the 1\textsuperscript{st} phase

\[ P_{x}(y) = P_{x}(\bullet) = \sum P_{x}(\bullet \bullet \bullet) \]
visualization of the 1st phase
visualization of the 1\textsuperscript{st} phase
visualization of the 1\textsuperscript{st} phase
visualization of the 1st phase

\[ \sum P_X(x) = \sum \prod_c P_{pa(c)}(c) \]

C: confounded variables
visualization of the 1st phase
visualization of the 1\textsuperscript{st} phase
visualization of the 1\textsuperscript{st} phase
visualization of the 1st phase
visualization of the 1st phase
visualization of the 1st phase

\[
\sum P(x) = \sum \prod P(y)
\]
Algorithm for gID (1\textsuperscript{st} phase)

\begin{verbatim}
function \text{gID}(\bullet, \bullet, \mathcal{G}, \mathcal{Z})
    if \exists z \in \mathcal{Z} \bullet = z \cap V then
        return \mathcal{P}_{z \setminus V, \bullet}(\bullet)
    \text{\textgreater \hspace{1cm}} \triangleright \text{check whether a matching experiment exists}

    if V \neq \text{An}(\bullet)_{\mathcal{G}} then
        return \text{gID}(\bullet, \bullet \cap \text{An}(\bullet)_{\mathcal{G}}, \mathcal{G}[\text{An}(\bullet)_{\mathcal{G}}], \mathcal{Z})
    \text{\textgreater \hspace{1cm}} \triangleright \text{retain only the ancestors of } \bullet

    if (W \leftarrow (V \setminus \bullet) \setminus \text{An}(\bullet)_{\mathcal{G}}) \neq \emptyset then
        return \text{gID}(\bullet, \bullet \cup \bullet, \mathcal{G}, \mathcal{Z})
    \text{\textgreater \hspace{1cm}} \triangleright \text{modify to a maximal intervention}

    S \leftarrow \mathcal{C}(\mathcal{G} \setminus \bullet \bullet)
    if |S| > 1 then
        return \sum_{\bullet} \prod_{\bullet \in S} \text{gID}(\bullet, \bullet, \mathcal{G}, \mathcal{Z})
    \text{\textgreater \hspace{1cm}} \triangleright \text{factorize into subqueries}

    for Z \in \mathcal{Z} \text{ such that } Z \cap V \subseteq \bullet \bullet \text{ do}
        return \text{SUB-ID}(\bullet, \bullet \setminus Z, \mathcal{P}_{z \setminus V, \bullet \cap \mathcal{Z}}, \mathcal{G} \setminus (Z \cap \bullet \bullet)) \text{ if not } \text{NONE}

    throw \text{FAIL}
\end{verbatim}

\textsc{sub-ID} is a simplified \textsc{ID} algorithm [SP'06], which returns \text{NONE} if failed.
Algorithm for gID (2nd phase)

function sub-ID(•, •, Q, G)

\{S\} ← C(G \setminus •)

if • = ∅ then
    return \sum_{v \setminus •} Q(v)

if V ≠ An(•)_G then
    return sub-ID(•, \• \cap An(•)_G, \sum_{v \setminus An(•)_G} Q, G[An(•)_G])

if C(G) = V then
    return NONE

if S ∈ C(G) then
    return \sum_{s \setminus •} \prod_{V_i \in •} Q(v_i | V^{(i-1)}_{π}).

if S ⊈ S' ∈ C(G) then
    return sub-ID(•, \• \cap S', \prod_{V_i \in S'} Q(V_i | V^{(i-1)}_{π} \cap S', \prod_{V_i \in S'} V^{(i-1)}_{π} \setminus S'), S')

▷ check identified

▷ retain only the ancestors of •

▷ check the existence of a hedge

▷ check identifiable

▷ modify input (query, distribution, and graph)
Example: Durg-Drug Interactions

$Y$: heart disease, $B$: blood pressure, $X_1, X_2$: drugs

$$P_{X_1, X_2}(y)$$
Example: Durg-Drug Interactions

$Y$: heart disease, $B$: blood pressure, $X_1, X_2$: drugs

$$
P_{x_1,x_2}(y) = \sum_b P_{x_1,x_2}(y, b)$$
Example: Durg-Drug Interactions

$Y$: heart disease, $B$: blood pressure, $X_1, X_2$: drugs

$$P_{x_1,x_2}(y) = \sum_b P_{x_1,x_2}(y, b)$$

$$= \sum_b P_{x_1}(b) P_{x_2,b}(y)$$
Example: Durg-Drug Interactions

$Y$: heart disease, $B$: blood pressure, $X_1, X_2$: drugs

$$P_{x_1,x_2}(y) = \sum_b P_{x_1,x_2}(y, b)$$

$$= \sum_b P_{x_1}(b)P_{x_2,b}(y)$$
Example: Durg-Drug Interactions

$Y$: heart disease, $B$: blood pressure, $X_1, X_2$: drugs

$$P_{X_1,X_2}(y) = \sum_{b} P_{X_1,X_2}(y,b)$$

$$= \sum_{b} P_{X_1}(b)P_{X_2,b}(y)$$

$$= \sum_{b} P_{X_1}(b)P_{X_2}(y|b)$$
non-g-identifiability
– the failure condition & a prohibiting structure
To prove $P_x(y)$ is not g-identifiable from $\mathcal{Z}$ in $\mathcal{G}$,
Proving Non-g-Identifiability

To prove $P_x(y)$ is not g-identifiable from $Z$ in $G$, we construct two causal models $M_1$ and $M_2$ compatible with $G$. 

SCM $M^1$ $\xleftarrow{G} \xrightarrow{G} \xrightarrow{G}$ SCM $M^2$
To prove $P_x(y)$ is **not g-identifiable** from $Z$ in $G$, we construct two causal models $M_1$ and $M_2$ compatible with $G$ such that $P_{z_1}^1(v) = P_{z_1}^2(v)$ for all $Z \in \mathcal{Z}$, $z \in \mathcal{X}_Z$. 

---

**Diagram: Proving Non-g-Identifiability**

**SCM $M_1$** → $G$ → **SCM $M_2$**

$P_{Z_1}(V)$ → ... → $P_{Z_n}(V)$

available data $\mathcal{P}$
To prove \( P_x(y) \) is **not g-identifiable** from \( Z \) in \( G \), we construct two causal models \( M_1 \) and \( M_2 \) compatible with \( G \) such that 
\[ P_z^1(v) = P_z^2(v) \]
for all \( Z \in \mathbb{Z}, z \in X_Z \), but 
\[ P_x^1(y) \neq P_x^2(y). \]
Recall the failed factor ...

\[ Q \quad P_x(y) \quad \ldots \quad P_0(\bullet) \quad \ldots \quad P_{Z_1} \quad \ldots \quad P_{Z_{m-1}} \quad \ldots \quad P_{Z_m} \]

\[ \exists P_0(\bullet) \text{ such that } \forall P_{Z_i} \in \mathbb{P} \text{ fails} \]
\( P_\bullet (\bullet) \) versus \( P_{Z_i} \in \mathbb{P} \) (phase-2)

There are 3 situations in identifying a factor \( P_\bullet (\bullet) \) with a distribution \( P_{Z_i} \in \mathbb{P} \).

- **3 (the good)** identified, e.g., \( P_D, P_{c, d} (e, f) = P_d (e, f | c) \)
- **7 (the ugly)** \( Z_i \) on, e.g., \( P_E, D \)
- **9 (the bad)** hedge \([SP'06]\), e.g., \( P_C, E, F \)

The original order should be: the good, the bad, and the ugly.
$P_{\greenbullet}(\bullet)$ versus $P_{Z_i} \in \mathbb{P}$ (phase-2)

There are 3 situations in identifying a factor $P_{\greenbullet}(\bullet)$ with a distribution $P_{Z_i} \in \mathbb{P}$.

- ✓ (the good) identified, e.g., $P_D$, $P_{c,d}(e, f) = P_d(e, f|c)$

the original order should be: the good, the bad, and the ugly.
\( P_{\circ} (\bullet) \) versus \( P_{Z_i} \in \mathbb{P} \) (phase-2)

There are 3 situations in identifying a factor \( P_{\circ} (\bullet) \) with a distribution \( P_{Z_i} \in \mathbb{P} \).

- ✓ (the good) identified,
  e.g., \( P_D, P_{c,d}(e, f) = P_{d}(e, f | c) \)

- ✗ (the ugly) \( Z_i \) on \( \bullet \),
  e.g., \( P_{E,D} \)

the original order should be: the good, the bad, and the ugly.
\( P_\bullet (\bullet) \) versus \( P_{Z_i} \in \mathbb{P} \) (phase-2)

There are 3 situations in identifying a factor \( P_\bullet (\bullet) \) with a distribution \( P_{Z_i} \in \mathbb{P} \).

- **✓ (the good)** identified,
  e.g., \( P_D, P_{c,d}(e, f) = P_d(e, f | c) \)

- **✗ (the ugly)** \( Z_i \) on \( \bullet \),
  e.g., \( P_{E,D} \)

- **✗ (the bad)** \( \exists \)hedge [SP’06],
  e.g., \( P_C \)

the original order should be: the good, the bad, and the ugly.
Failure $\Rightarrow \exists$ Thicket

- A **thicket** is the superimposition of **hedges** (the bad structure).
- (if every experiment intersects with ●, confounded ●s form a ‘degenerate thicket.’)

**Hedge**: a fence or boundary formed by closely growing bushes or shrubs. **Thicket**: a dense group of bushes or trees.
• A thicket is the superimposition of hedges (the bad structure).
• (if every experiment intersects with $\bullet$, confounded $\bullet$s form a ‘degenerate thicket.’)
Failure $\Rightarrow \exists$ Thicket

- A **thicket** is the superimposition of **hedges** (the bad structure).
- (If every experiment intersects with $\circ$, confounded $\circ$s form a ‘degenerate thicket.’)

**Hedge**: a fence or boundary formed by closely growing bushes or shrubs. **Thicket**: a dense group of bushes or trees.
A **thicket** is the superimposition of **hedges** (the bad structure).

(if every experiment intersects with ●, confounded ●s form a ‘degenerate thicket.’)

**Hedge**: a fence or boundary formed by closely growing bushes or shrubs. **Thicket**: a dense group of bushes or trees.
# Definition

Let $\mathbf{R}$ be a non-empty set of variables and $\mathcal{Z}$ be a collection of sets of variables in $\mathcal{G}$. A thicket $\mathcal{J} \subseteq \mathcal{G}$ is an $\mathbf{R}$-rooted $c$-component consisting of a minimal $c$-component over $\mathbf{R}$ and hedges

\[ \mathcal{F}_J = \{ \langle \mathcal{F}_Z, \mathcal{J}[\mathbf{R}] \rangle \mid \mathcal{F}_Z \subseteq \mathcal{G} \setminus \mathcal{Z}, \mathcal{Z} \cap \mathbf{R} = \emptyset \}_{\mathcal{Z} \in \mathcal{Z}}. \]

Let $\mathbf{X}$, $\mathbf{Y}$ be disjoint sets of variables in $\mathcal{G}$. A thicket $\mathcal{J}$ is said to be formed for $P_x(y)$ in $\mathcal{G}$ with respect to $\mathcal{Z}$ if $\mathbf{R} \subseteq \text{An}(\mathbf{Y})_{\mathcal{G}_{\mathbf{X}}}$ and every hedgelet of each hedge $\langle \mathcal{F}_Z, \mathcal{J}[\mathbf{R}] \rangle$ intersects with $\mathbf{X}$. 
Given $\mathbb{Z} = \{\{X_1\}, \{X_2\}\}$ with $Q = P(y|\text{do}(x_1, x_2))$ (* $P(\bullet) = P(\bullet)$)

- The query is not g-identifiable, see $B \leftarrow - \rightarrow Y$

$$P_{x_1, x_2}(y) = \sum_b P_{x_1, x_2}(y, b) = \sum_b P_{x_1, x_2}(y, b)$$
Thicket: Drug-Drug Interactions

Given \( \mathcal{Z} = \{\{X_1\}, \{X_2\}\} \) with \( Q = P(y|do(x_1, x_2)) \) (* \( P \odot (\bullet) = P \odot (\bullet) \))

- There is a hedge, which is disjoint with \( \{X_1\} \in \mathcal{Z} \), and also intersects with \( X \).

\[
P_{x_1,x_2}(y) = \sum_b P_{x_1,x_2}(y, b) = \sum_b P_{x_1,x_2}(y, b)
\]
Thicket: Drug-Drug Interactions

Given $\mathcal{Z} = \{\{X_1\}, \{X_2\}\}$ with $Q = P(y|do(x_1, x_2))$ (* $P\circ \bullet = P\bullet \circ \bullet$)

- There is another hedge, which is disjoint with $\{X_2\} \in \mathcal{Z}$ and, again, intersects with $X$.

$$P_{x_1, x_2}(y) = \sum_b P_{x_1, x_2}(y, b) = \sum_b P_{x_1, x_2}(y, b)$$
Thicket: Drug-Drug Interactions

Given $\mathcal{Z} = \{\{X_1\}, \{X_2\}\}$ with $Q = P(y|do(x_1, x_2))$ (* $P_\bullet(\bullet) = P_\circ(\circ)$)

- This is a thicket for $P_x(y, b)$ w.r.t. $\mathcal{G}$ and $\mathcal{Z}$.

$$P_{x_1, x_2}(y) = \sum_b P_{x_1, x_2}(y, b) = \sum_b P_{x_1, x_2}(y, b)$$
Thicket: Drug-Drug Interactions

Given $Z = \{\{X_1\}, \{X_2\}\}$ with $Q = P(y | do(x_1, x_2))$ (* $P(\bullet) = P(\circ)$)

- This is a thicket for $P_x(y)$ w.r.t. $G$ and $Z$.

\[ P_{x_1,x_2}(y) = \sum_{b} P_{x_1,x_2}(y, b) = \sum_{b} P_{x_1,x_2}(y, b) \]
Non-g-identifiability

gID algorithm
FAILs

THICKET exists for $P(M_1 \& M_2)$

due to non-gid $P(x(y))$. M1 \& M2 are not presented in this talk
Non-g-identifiability

gID algorithm FAILs

THICKET exists for \( P(\bullet)(\bullet) \)

\[ P(\bullet)(\bullet) \]

\[ M_1 \land M_2 \]

not presented in this talk
Non-g-identifiability

- gID algorithm
  - FAILs

- THICKET exists for $P(x(y))$

- non-gid $P(x(y))$ 
  - $\mathcal{M}^1 & \mathcal{M}^2$

Not presented in this talk
Non-g-identifiability

- gID algorithm FAILs
- THICKET exists for $P_0(y)$ and $M_1 \& M_2$
- non-gid $P_0(y)$ in $M_1 \& M_2$
- non-gid $P_x(y)$ in $M_1' \& M_2'$

not presented in this talk
Conclusions

• We studied general-identifiability — causal effect identifiability given a causal graph and an arbitrary combination of observational and experimental distributions.

• ✓ a necessary and sufficient graphical condition (thicket?).
  ✓ a sound and complete algorithm

• Research Directions: finite-sample efficient formula, studying bounds for the causal effect when not-g-identifiable, incorporating functional assumptions, without a causal graph or partially-specified graphs.

• We further investigated the generalization of this work for transportability (data coming from heterogeneous domains) and for conditional causal effect, e.g., $P_x(y|w)$ (next session@Murray).
Conclusions

- We studied **general-identifiability** — causal effect identifiability given a causal graph and an **arbitrary combination of observational and experimental distributions**.

- A necessary and sufficient graphical **condition**.
  - A sound and complete **algorithm**

- **Research Directions**: finite-sample efficient formula, studying bounds for the causal effect when not-g-identifiable, incorporating functional assumptions, without a causal graph or partially-specified graphs.

- We further investigated the generalization of this work for transportability (data coming from heterogeneous domains) and for conditional causal effect, e.g., $P_x(y|w)$ (next session@Murray).
Conclusions

• We studied general-identifiability — causal effect identifiability given a causal graph and an arbitrary combination of observational and experimental distributions.

• ✓ a necessary and sufficient graphical condition (thicket?).
  ✓ a sound and complete algorithm

• Research Directions: finite-sample efficient formula, studying bounds for the causal effect when not-g-identifiable, incorporating functional assumptions, without a causal graph or partially-specified graphs.

• We further investigated the generalization of this work for transportability (data coming from heterogeneous domains) and for conditional causal effect, e.g., $P_x(y|w)$ (next session@Murray).
References

Pearl’00  J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press


HV’06  Y. Huang, M. Valtorta. Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm. AAAI 2006


BP’14  E. Bareinboim, J. Pearl. Transportability from Multiple Environments with Limited Experiments: Completeness Results NeurIPS 2014

LCB’19  S. Lee, J. D. Correa, E. Bareinboim. General Identifiability with Arbitrary Surrogate Experiments. UAI 2019