Link Prediction and Causality

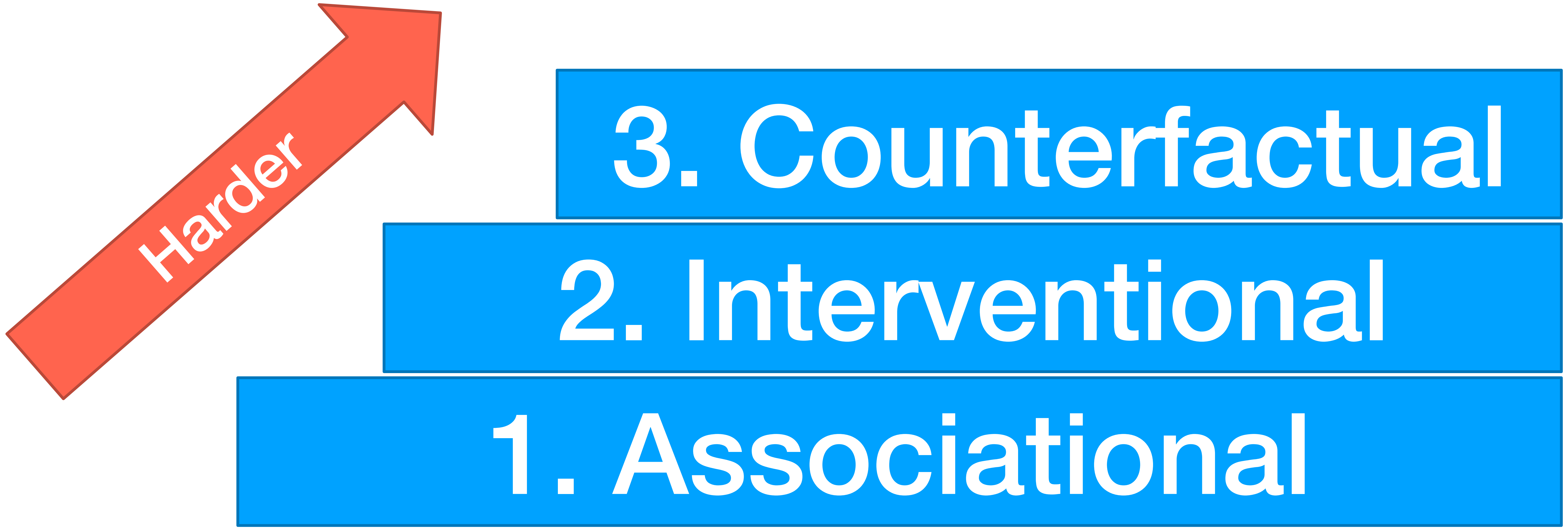
Bruno Ribeiro
Purdue University
(Visiting Associate Professor @ Stanford
on sabbatical until July 2024, **Gates 452**)

Stanford CS224W,
November 30th, 2023

Outline

- A short introduction to causality
- Causality in out-of-distribution graph tasks
- Temporal Link Prediction = Static Link Prediction (Associational)
- Causal link prediction: models & the challenge of cascading dependencies

The 3 rungs of the ladder of causation



Rung 1: Associational

- Traditional graph machine learning tasks

  Assume $X \not\perp\!\!\!\perp Y$

  Task: Predict output Y from input X

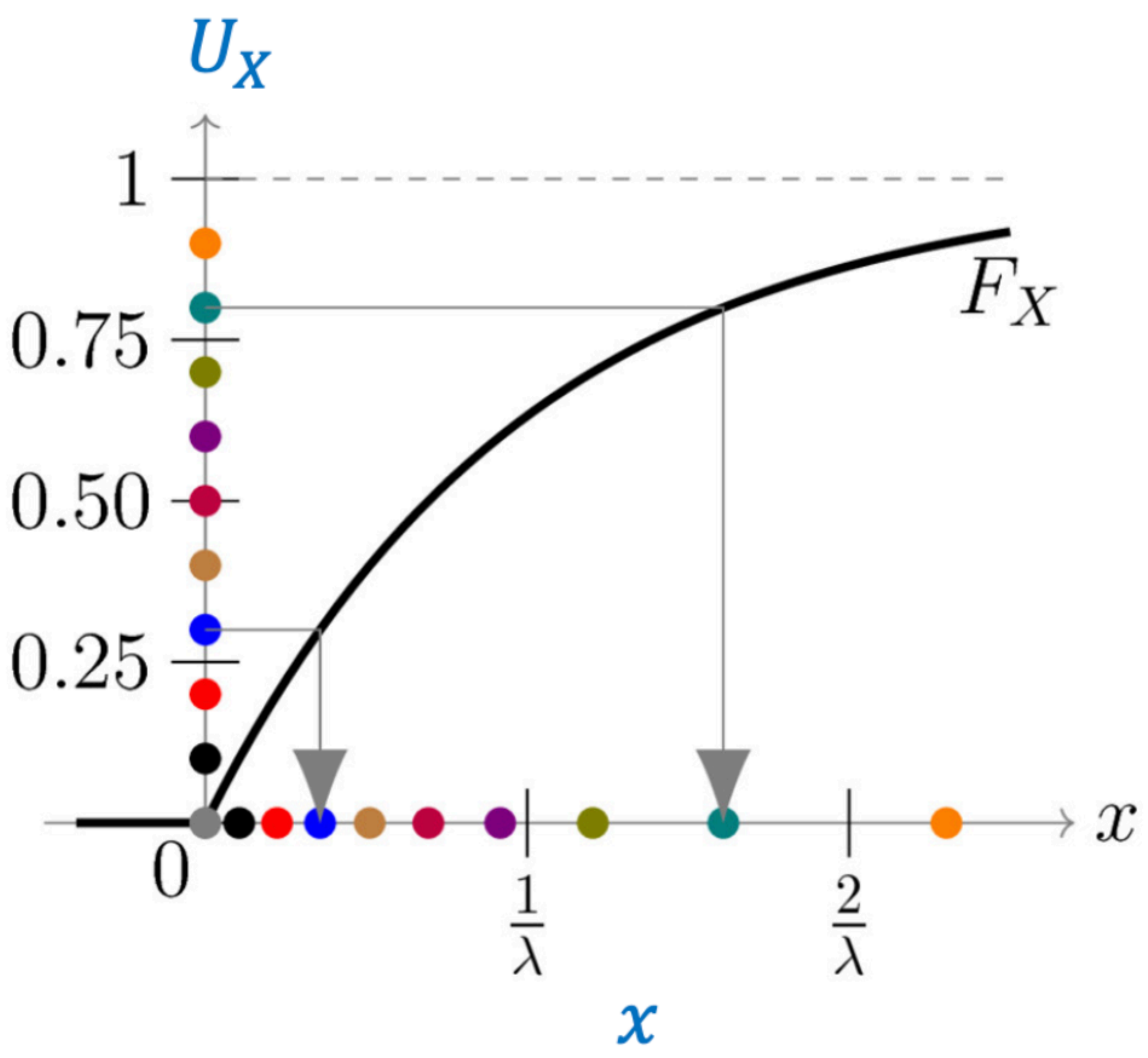
  Data: samples of (X,Y)

Background: Inverse Transform Sampling

- Data generation algorithm:
  - Let $U_X \sim \text{Uniform}(0,1)$ be a random uniform value in the interval [0,1]
  - Then,

$$X := F_X^{-1}(U_X)$$

    is a random sample with distribution $P(X = x)$.
  - Exponential distribution example: $P(X \leq x) = F_X(x) = 1 - e^{-\lambda x}$ with inverse $x = F_X^{-1}(U_X) = -\frac{1}{\lambda}\ln(1 - U_X)$:



Rung 2: Interventional

- Tasks where we must predict the effect of an intervention

Assume $X \not\perp\!\!\!\perp Y$

Task: Predict output Y from acting on input X

Data: samples of (Y,**do**(X=x))

Rung 2: Interventional (cont)

Imagine two hypothetical data generators for same $P(X, Y)$

$U_y, U_x \sim$ i.i.d. Uniform$(0,1)$

$$X := f_x(U_x) \quad Y := f_y(X, U_Y) \quad = \quad Y := f_y(U_x) \quad X := f_x(Y, U_Y)$$

- $\mathrm{do}(X = x)$ changes f_x to a constant in data generation

$$X := x \quad Y := f_y(X, U_Y) \quad \neq \quad Y := f_y(U_x) \quad X := x$$

Rung 3: Counterfactual

- Tasks where we must imagine the effect of an intervention at an event that has "already happened"

Assume $X \not\perp\!\!\!\perp Y$

Task: Predict output Y from acting on input X

Data: $Y(X = x) \,|\, X = x', Y = y'$ or $Y(X = x) \,|\, X = x'$

Rung 3: Counterfactual

Imagine two hypothetical data generators for same $P(X, Y)$

$U_y, U_x \sim$ i.i.d. Uniform$(0,1)$

$$X := f_x(U_x) \qquad\qquad Y := f_y(U_x)$$
$$Y := f_y(X, U_Y) \quad = \quad X := f_x(Y, U_Y)$$

- Now assume we know $X = x', Y = y'$

  This knowledge changes distribution of U_x and U_y

$$X := x \qquad\qquad Y := f_y\big(U_x \,|\, (X = x', Y = y')\big)$$
$$Y := f_y(X, U_Y \,|\, (X = x', Y = y')) \quad \neq \quad X := x$$

Causal DAG

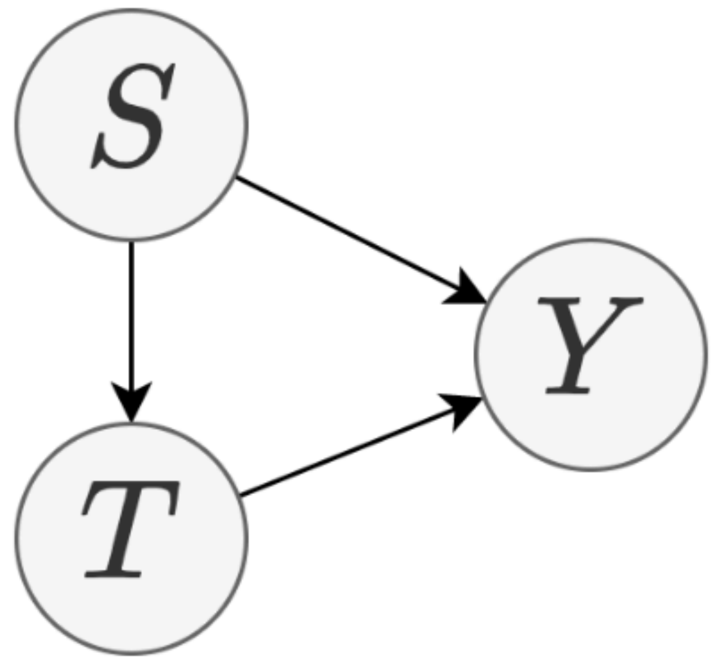
- Representing causal dependencies using graphs (example in the extra notes)

    - S = Kidney stone size
    - T = Treatment type
    - Y = Treatment outcome

$$
\begin{aligned}
S &= F_S^{-1}(U_{\text{stone size}}), \\
T &= F_T^{-1}(S, U_{\text{treatment}}), \\
Y &= F_C^{-1}(T, S, U_{\text{outcome}}),
\end{aligned}
$$

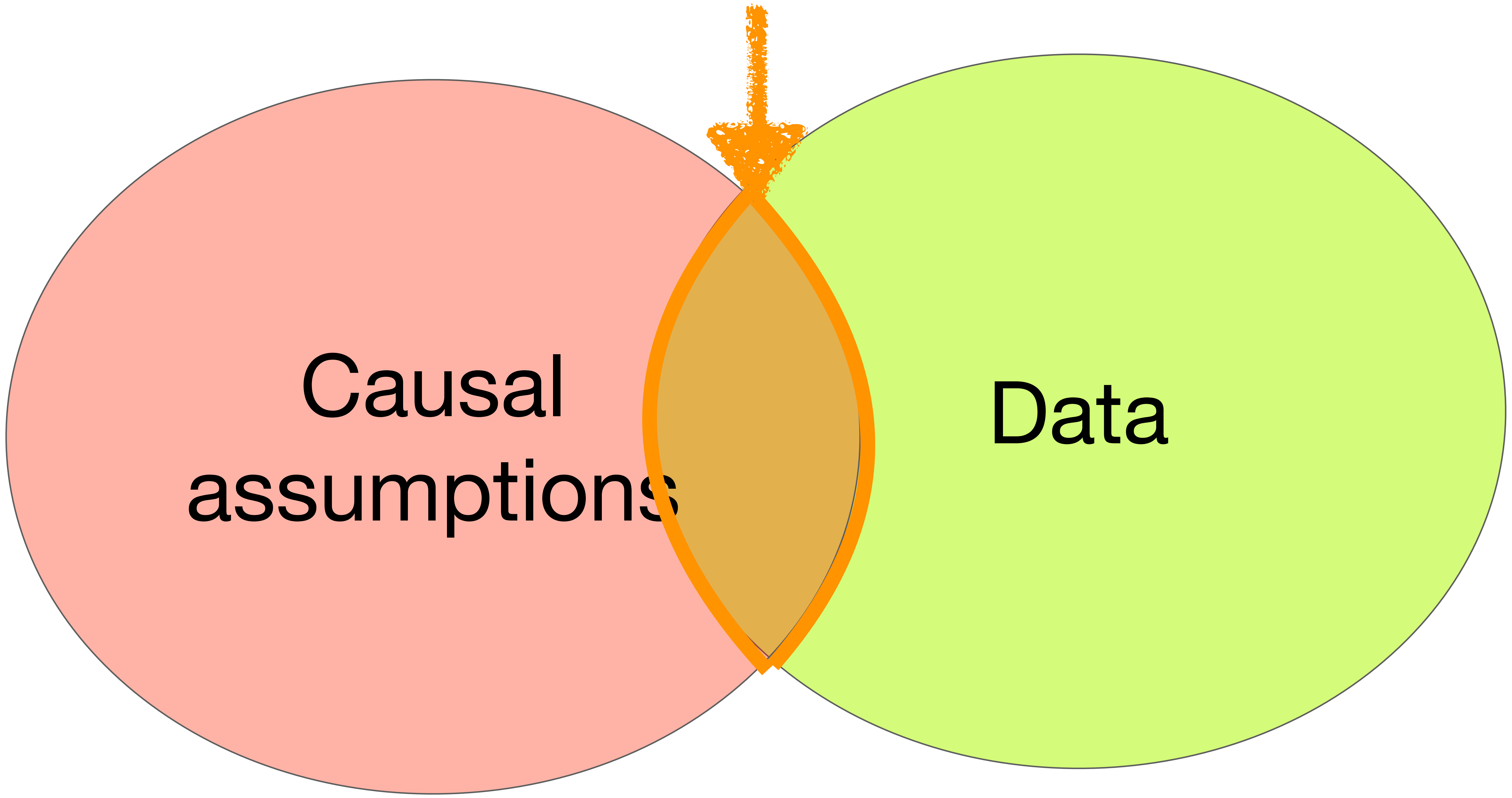
where $U_{\text{stone size}}, U_{\text{treatment}}, U_{\text{outcome}} \in [0, 1]$ are independent variables.

The above data generation can be described by an execution graph, called the **causal Directed Acyclic Graph (DAG)**:



Causality Challenge: *Identifiability*

Identifiable queries:
Causal queries we can **answer** with our data



Some graph tasks are causal

Link prediction for **decision-making interventions**
(e.g., search & recommendations) tends to be causal

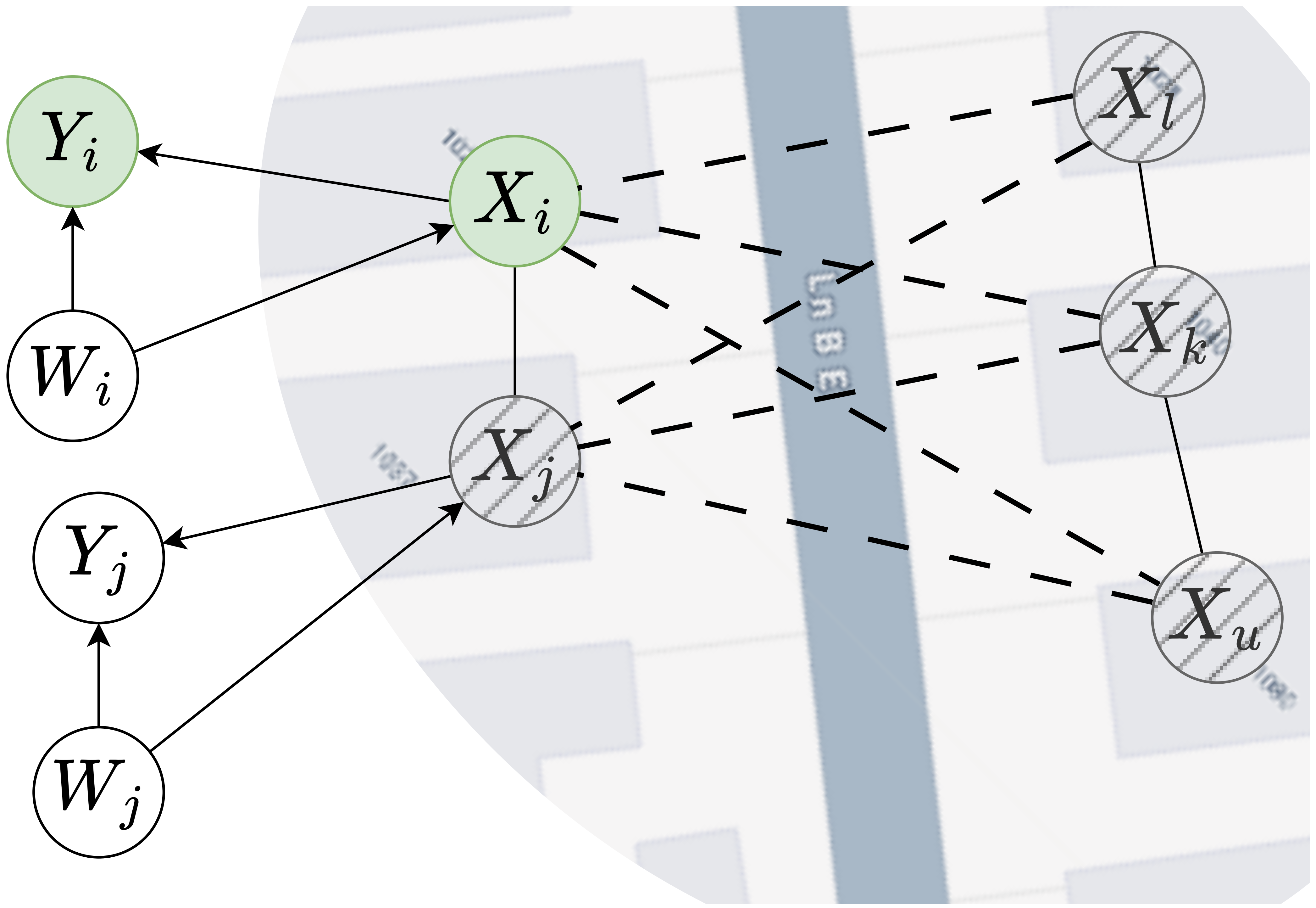
$$P(\mathrm{Accept}(i,j) = \mathrm{yes} \mid do(\mathrm{show\ recommendation} = j \mathrm{\ to\ user} = i))$$

Can we identify (*answer*) these queries?

Importance of Causality in Decision-making

Zillow House Offer Example (my best-guess)

- Consider a graph where
  - X_i : characteristics of house i
  - Y_i : price of house i
  - W_i: whether homeowner is ready to put house i on the market



Observed variable on market

Observed variable but not on market

Unobserved variable

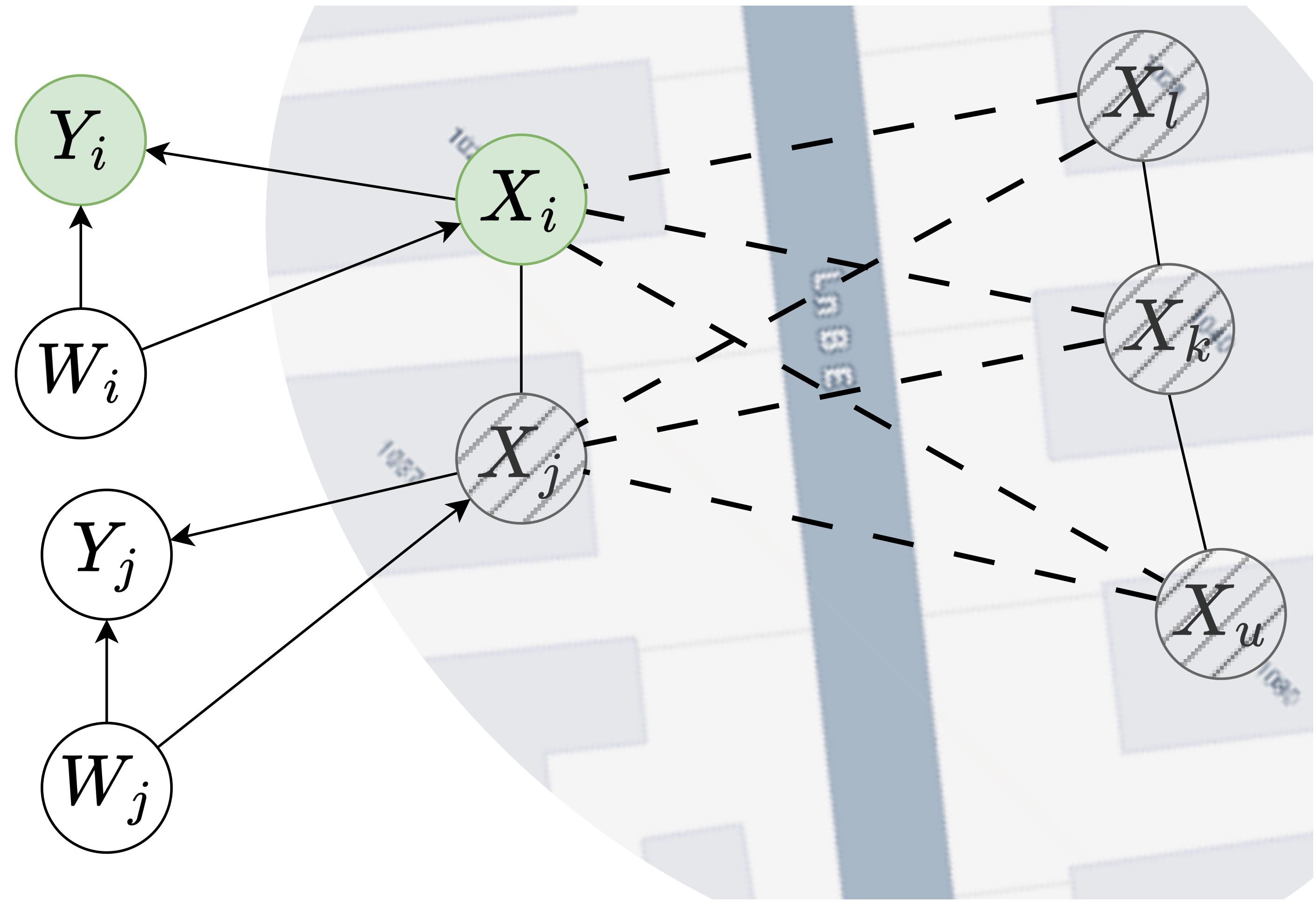
Causal relation

Association by common ancestor (location)

Zillow's Offer Intervention

- Zillow wants to make an **unsolicited** offer $(Y_j - \Delta y)$
  - Since Y_j is unobserved, Zillow can use the predictor $\hat{p}(y \mid X_j, \{Y_m, X_m\}_{m \in N_j})$ learned from houses sold on the market (green observations)

- **But an unsolicited offer is an intervention: do**$(W_j = 1)$
  - Zillow should be predicting instead: $p(y \mid \text{do}(W_j = 1), X_j, \{Y_m, X_m\}_{m \in N_j})$
  - W_j is a confounder between X_j and Y_j
    - $W = 1$ is associated with high prices Y, since owner may *improve* home *livability* to fetch a higher price (not fully reflected on X)

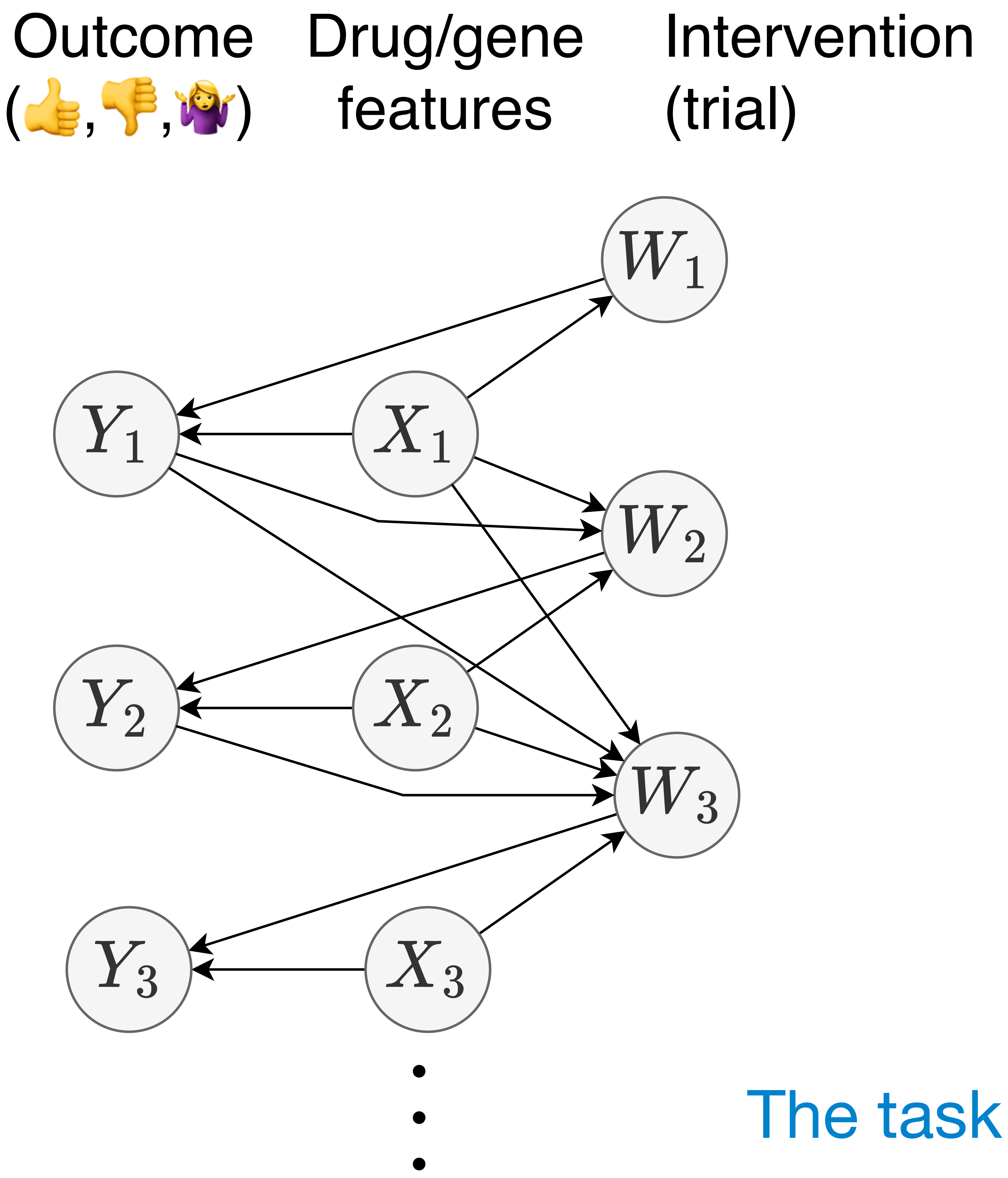


CNN BUSINESS Markets Tech Media Calculators Videos

Zillow’s home-buying debacle shows how hard it is to use AI to value real estate

- “In 2021 for certain homes, Zillow’s "Zestimate" would also represent an initial cash offer from the company to purchase the property.”
- “[Zillow] took a $304 million inventory write-down in the third quarter, which it blamed on having recently purchased homes for prices that are higher than it thinks it can sell them.”

Biomedical Experiment Causal Graph



- At step j, intervening $W_j = 1$ may consider features X_j and the likelihood of success (i.e., account for past success cases)

- Query: $P(Y_4 = y \,|\, X_4, \mathrm{do}(W_4 = 1))$
  - May not be answerable with data due to cascading dependencies
  - $Y_j \,|\, X_j, W_j$ depends on $Y_1, X_1, \ldots, Y_{j-1}, X_{j-1}$

The task is a little easier if we split Y into two variables (outcome) $Y' \in \{$👍 , 👎$\}$ and (observation)$O \in \{$ 🧪, 🤷‍♀️$\}$ but the overall cascading challenge persists

Detour

Other Applications of Causality in Graph Learning

(Out-of-distribution Graph Tasks)

Consider an out-of-distribution graph classification task

Running example:

Training data: (G^{tr}, Y^{tr})

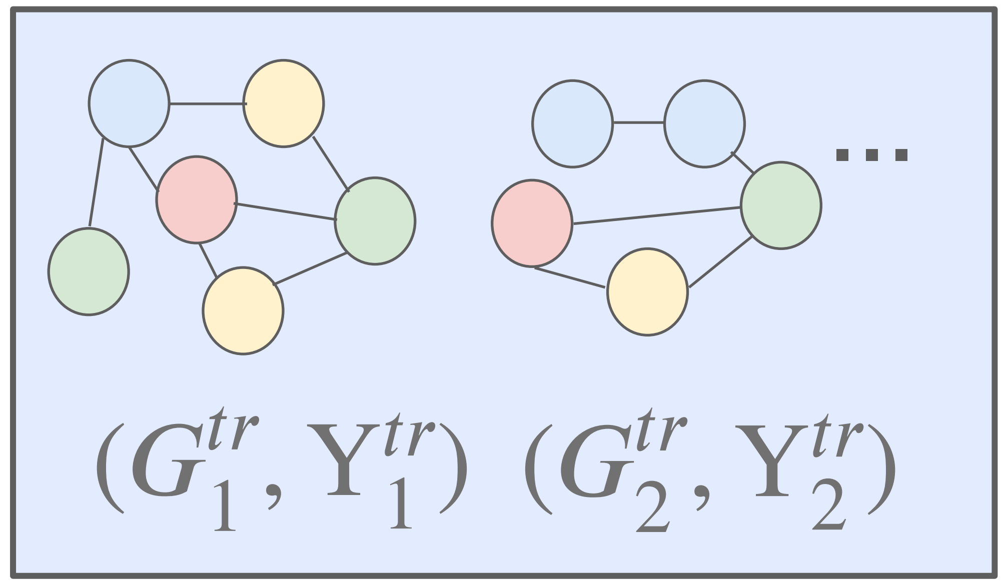
Test data: Predict Y^{te} given G^{te} ,

under $P(Y^{tr} \mid G^{tr}) = P(Y^{te} \mid G^{te})$

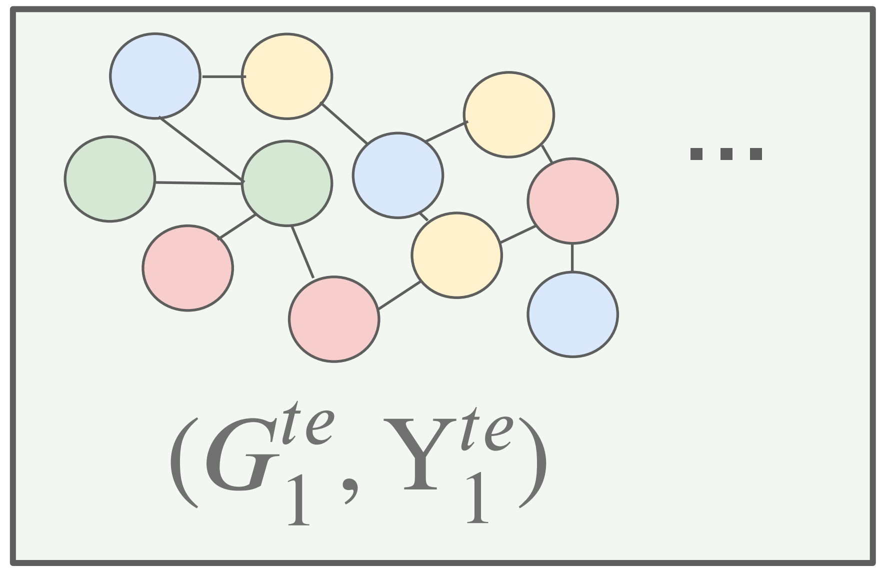
and $\text{supp}(G^{tr}) \neq \text{supp}(G^{te})$

where $\text{supp}(G) := \{ \forall G : P(G) > 0 \}$

Train (small graphs)



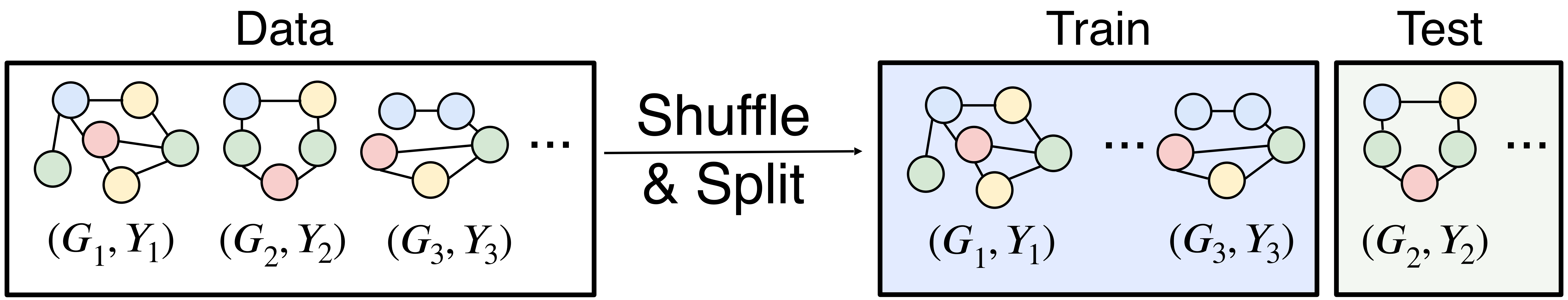
Test (large graphs)



Differences between In-distribution and Out-of-distribution tasks

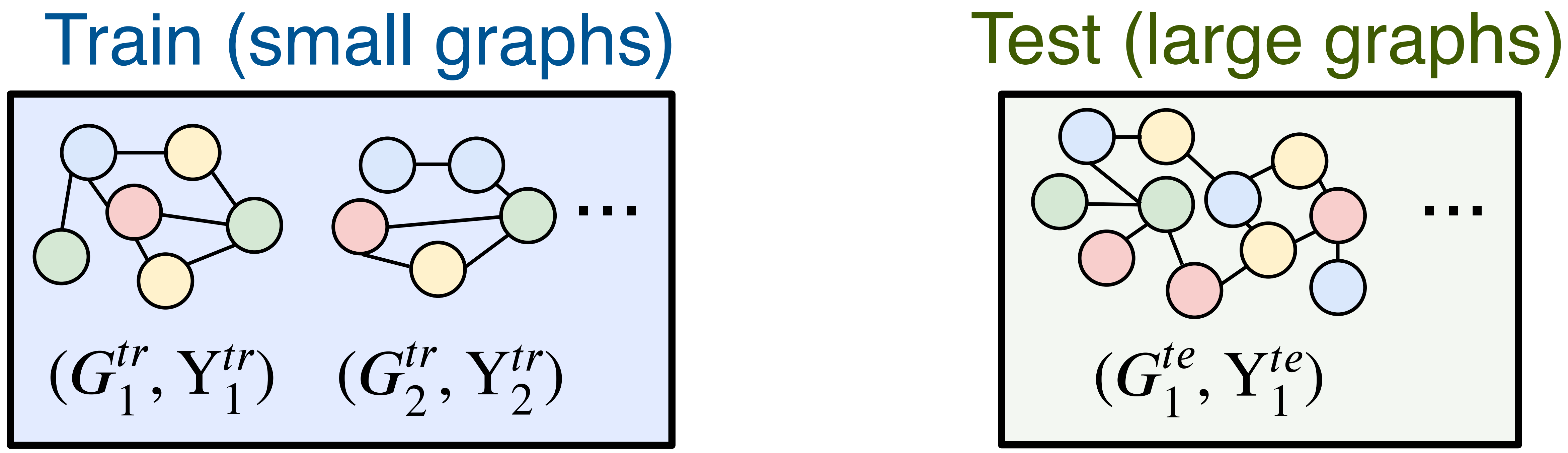
In-distribution graph classification task:

Predicting unseen examples of training distribution



Out-of-distribution graph classification task: (since we have no access to test data):

What would be the label of a graph if it were larger?



or vice-versa

Bevilacqua et al., ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Out-of-distribution tasks are a mix of associational and counterfactual tasks

- Out-of-distribution tasks are associational

Data: (X^{tr}, Y^{tr})

Task: Predict Y^{te} given X^{te} ,

under $P(Y^{tr} \mid X^{tr}) = P(Y^{te} \mid X^{te})$

tr = training distribution
te = test distribution

Out-of-distribution tasks are a mix of associational and counterfactual tasks

- But the **learning** is **counterfactual**

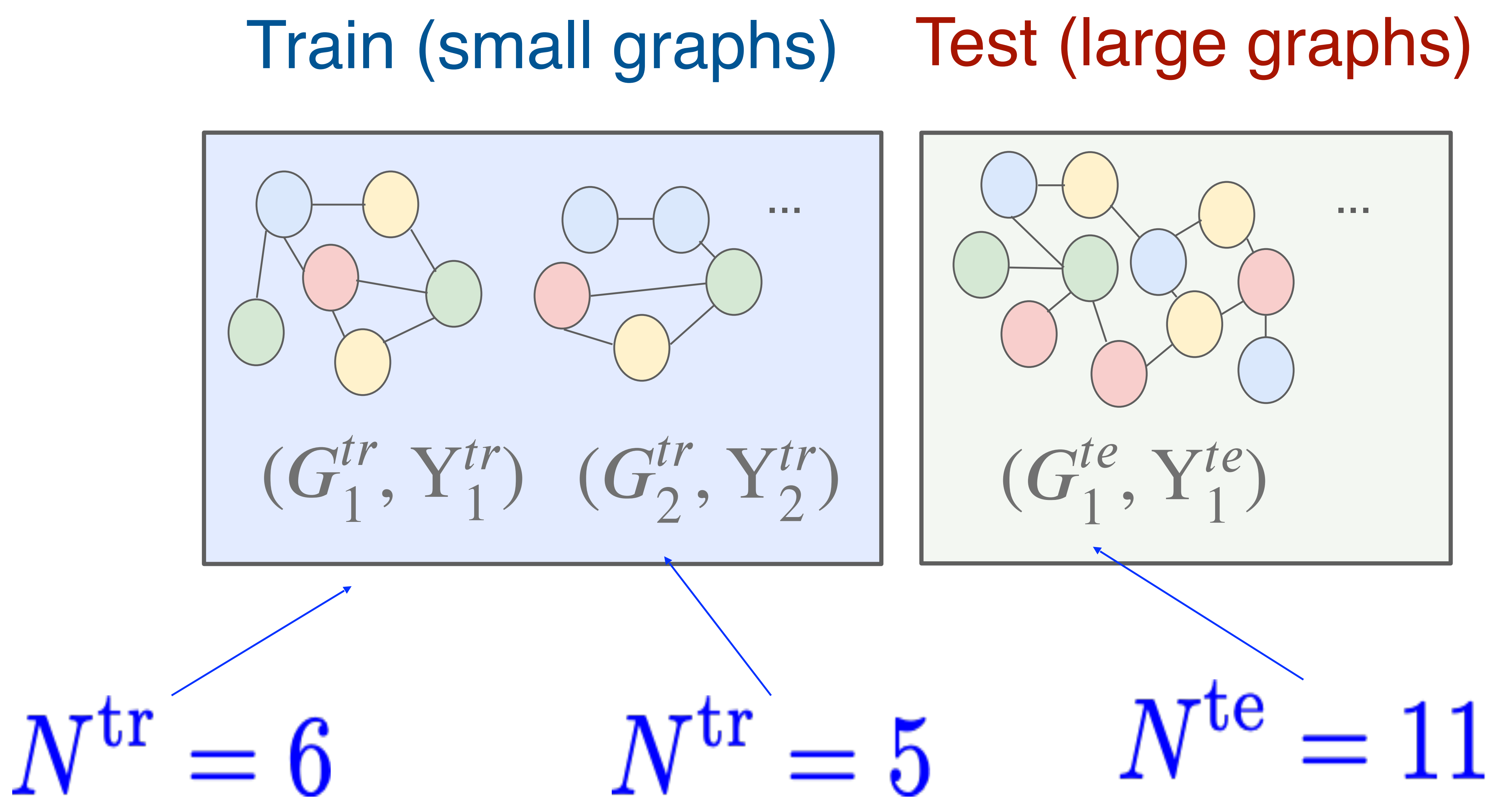
1. Without examples from graphs in test G^{te}
2. The classifier must build a correct predictor for unseen graph sizes

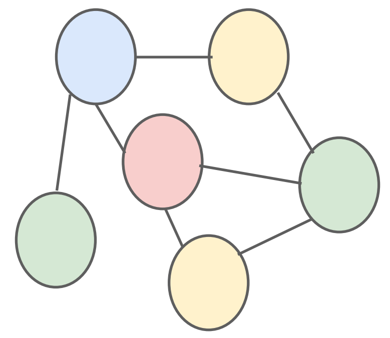
$$Y(N = n) \,|\, N = n^{\text{tr}}, G = g^{\text{tr}}, Y = y^{\text{tr}}$$

Given the size, topology, and label seen in training, what would have been the label if the graph were larger?

Why is Graph OOD Learning a Counterfactual Task?

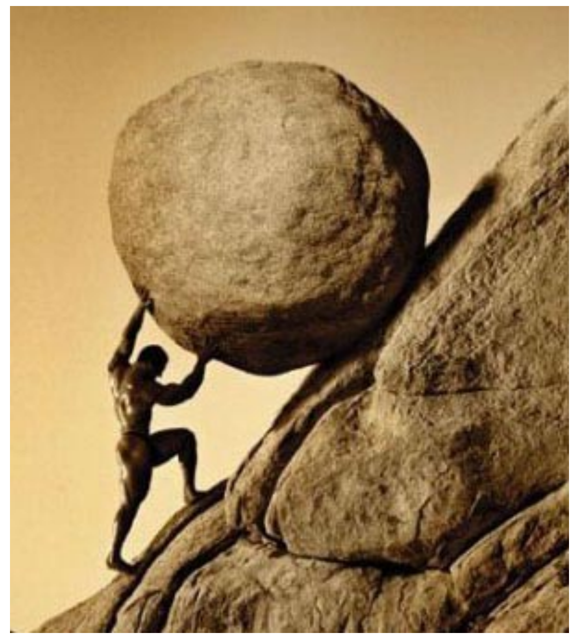
Example:



- Upon seeing graph  in training what would it look like if had N=11 nodes rather than N=6?

Difficult task for data augmentation: How to grow a graph?

Data augmentation question:
What it would look like if graph were larger without changing class label?



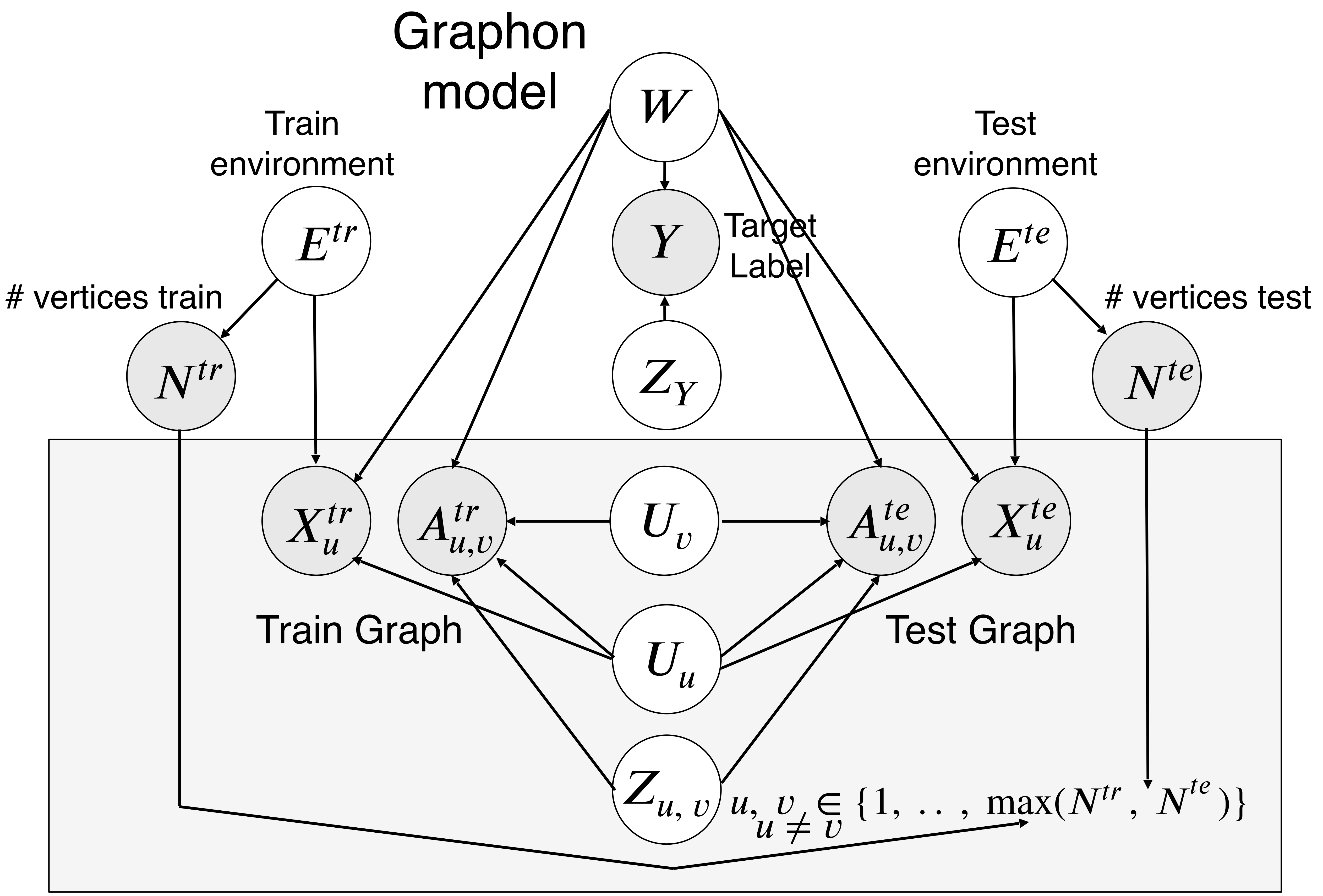
Unnecessarily hard!!

Counterfactual-invariant representation question:
What would be an invariant representation if graph were larger without changing class label?

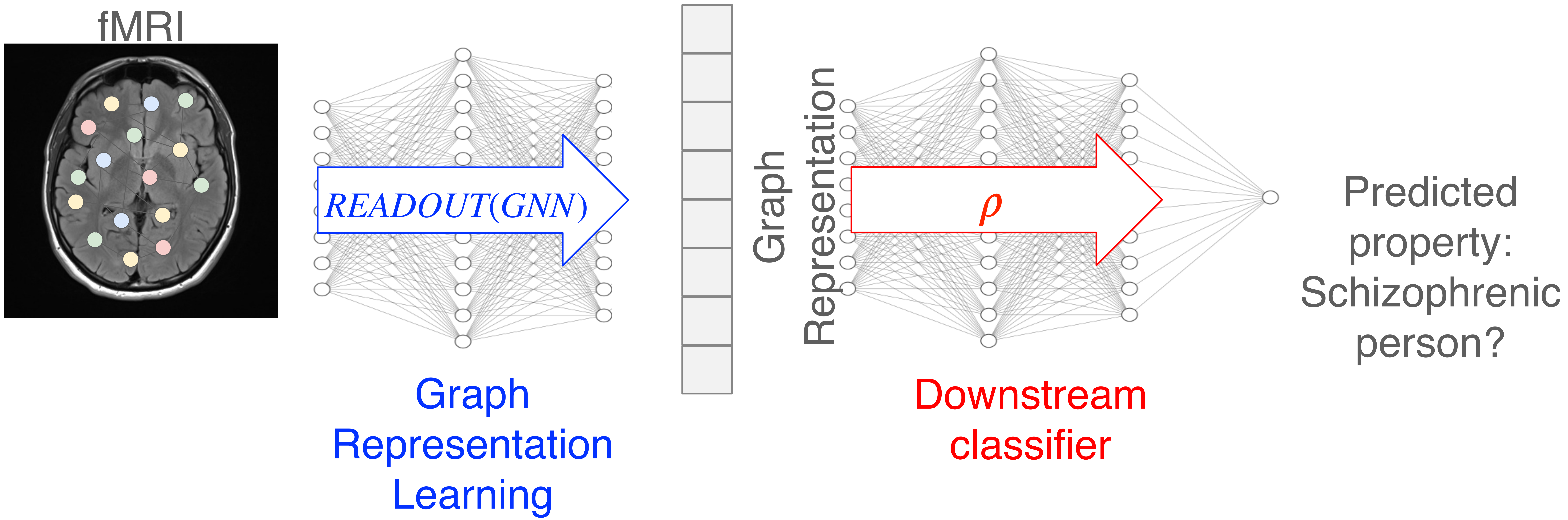
Bevilacqua, Zhou, R., ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

A Causal Mechanism for Graph Sizes

- *Graph formation process (Graphon):*
  - Graph label Y is a function of the graph model W & some random noise
  - Graph size $N^{tr}(N^{te})$ is a function of "environment" $E^{tr}\left(E^{te}\right)$ only
  - Train (test) graphs are generated by W and $E^{tr}\left(E^{te}\right)$ with same random noises

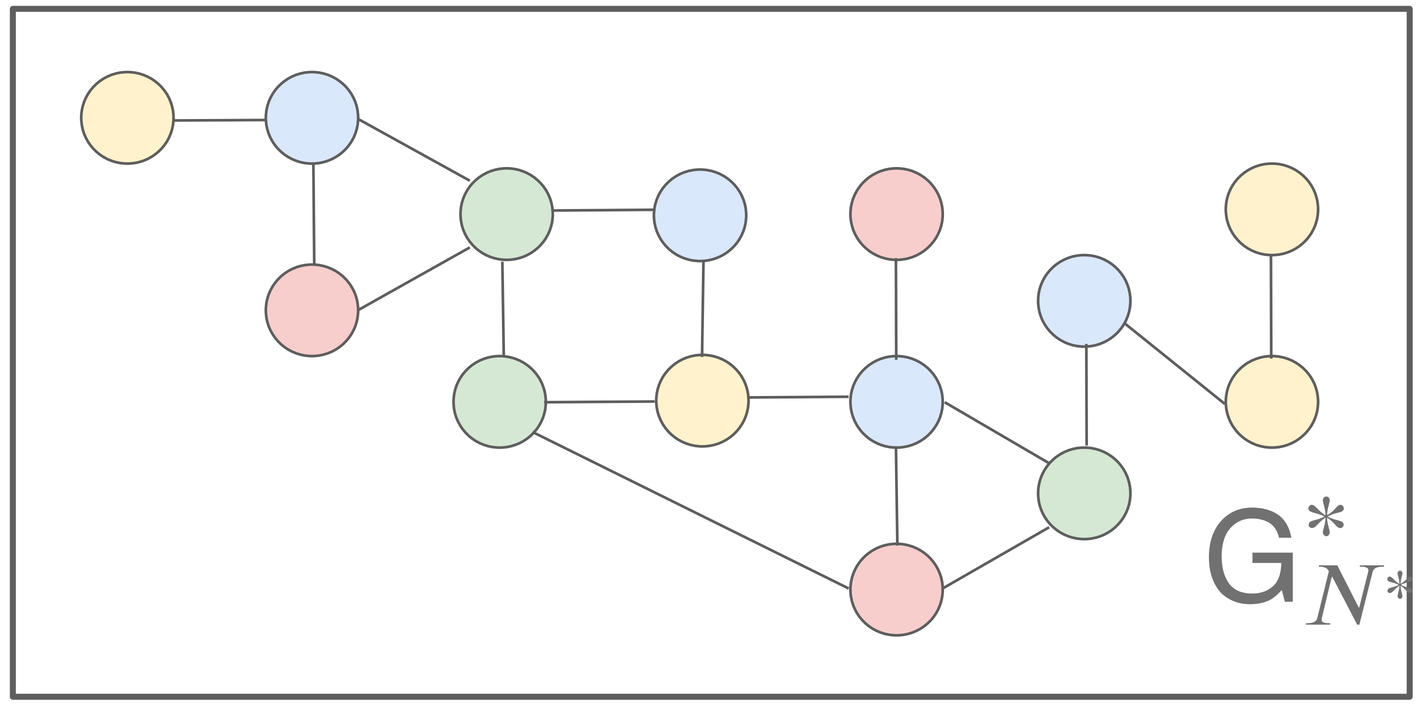


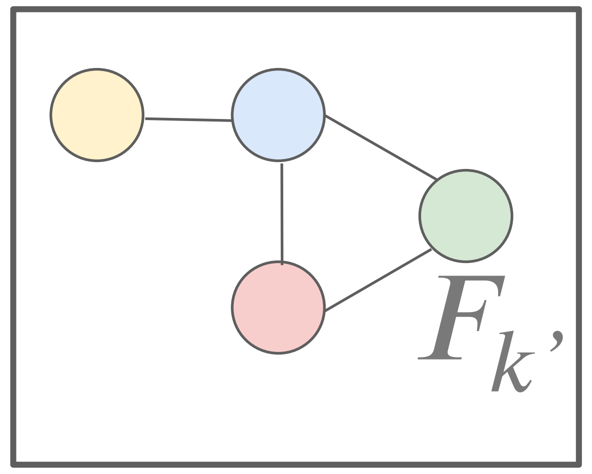
Graph Classification Task Example



Bevilacqua, Zhou, R., ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Graph Representation from Subgraph Densities





Induced subgraph of $G^*_{N^*}$

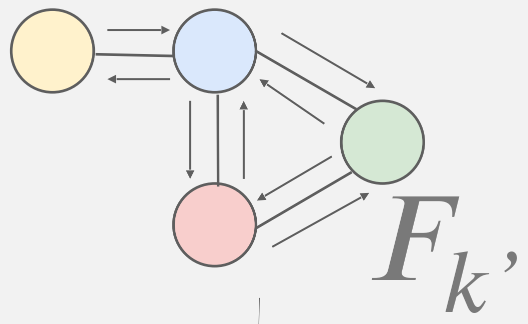
New graph representation

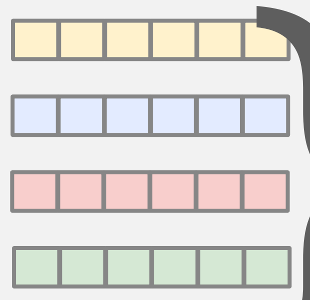
Induced subgraph density

$$\Gamma_{\text{GNN}}\left(\mathrm{G}^*_{N^*}\right) = \sum_{F_{k'} \in \mathscr{F}_{\leq k}} \mathrm{t}_{\text{ind}}\left(F_{k'}, \mathrm{G}^*_{N^*}\right) \mathrm{READOUT}_{\Gamma}(\mathrm{GNN}(F_{k'}))$$

GNN-representation of $F_{k'}$

$\mathrm{GNN}(F_{k'})$:



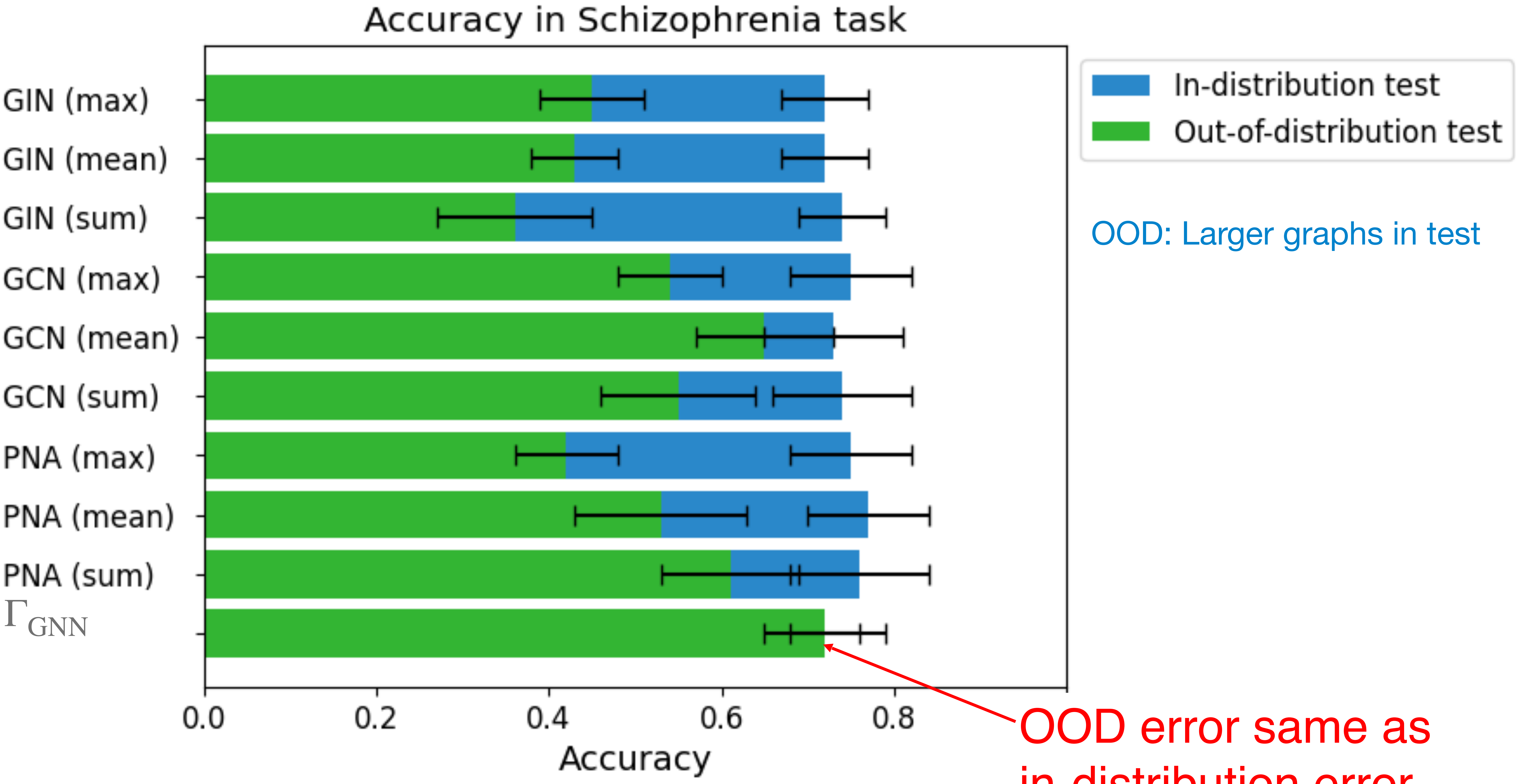
$\mathrm{READOUT}_{\Gamma}\left(\left\{ \right.\right.$  $\left.\left.\right\}\right)$

Proof of approximate
counter-factual invariance in the paper

Bevilacqua, Zhou, R., ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

OOD Error in Schizophrenia Graph Classification Task

- Can subgraph density representation Γ_{GNN} extrapolate OOD?



OOD: Larger graphs in test

OOD error same as in-distribution error

Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

End Detour

Link Prediction

Is Temporal Graph Learning Causal?

Not necessarily.

Theoretically,
Temporal Graph Learning is
Equivalent to Static Graph Learning

Temporal Graph Representation Learning is **Observational**

(Gao & R., 2021) describes the theory of temporal graph representation learning

- ***Equivalence*** *between two temporal graph representation learning frameworks:*
  1. **Time-and-graph representations**
  2. **Time-then-graph representations**

- In general, time-and-graph and time-then-graph are equally expressive

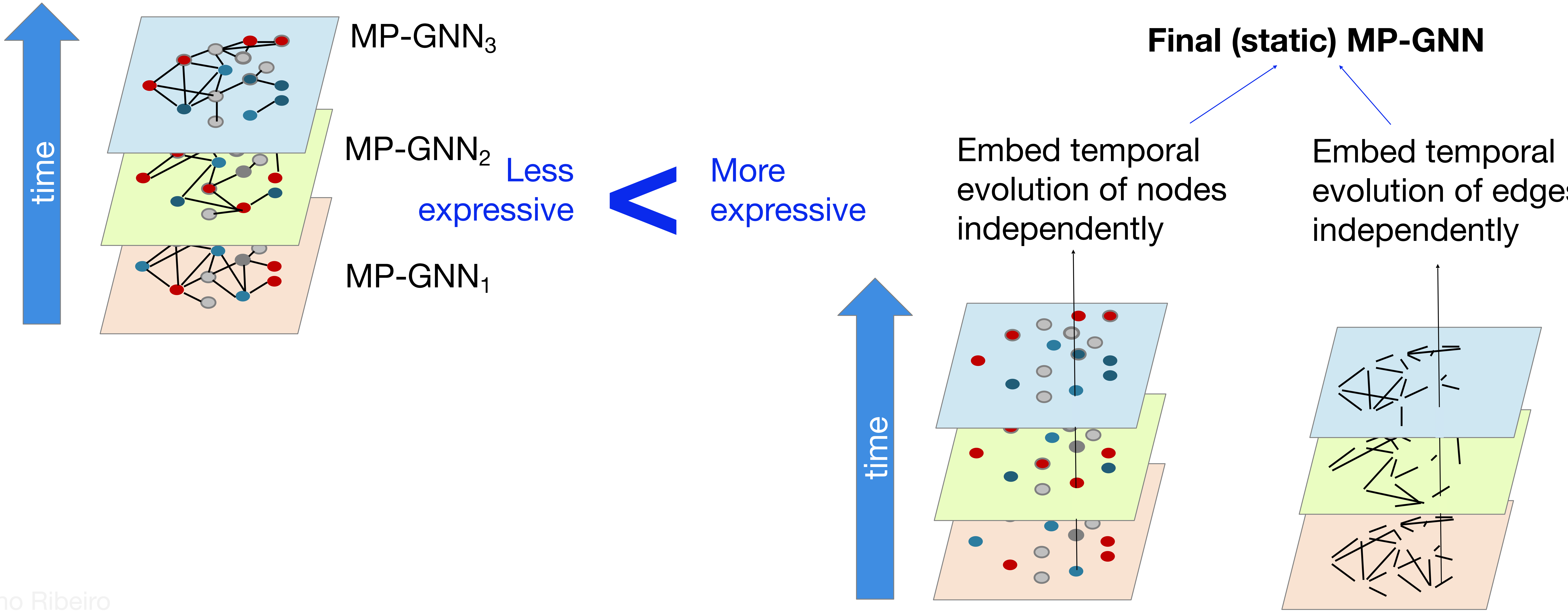
- Using Message Passing GNNs (MP-GNNs), **time-then-graph** are *more expressive* than **time-and-graph**

PURDUE UNIVERSITY

Time-then-graph more expressive than Time-and-graph (when using MP-GNNs)

- **Time-and-graph**
  - Encodes how node embeddings evolve over time
  - Majority of existing works

- **Time-then-graph**
  - Embedding encodes time evolution of nodes and edges independently
  - Impose permutation-equivariance via final static graph

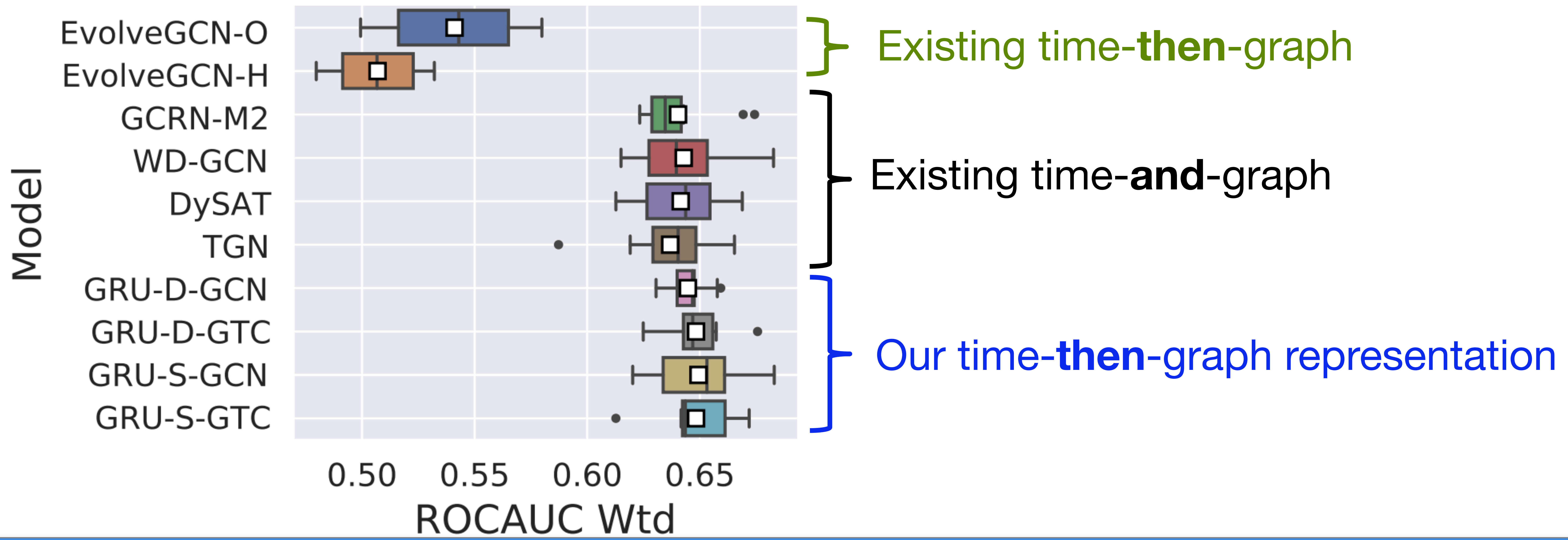


Bruno Ribeiro

Example: COVID-19 Observational Predictions

PURDUE UNIVERSITY

Task: Predict if a node will get infected
- **Input:** Temporal graph and epidemic evolution (discretized in time)
- **Output:** Probability a node gets infected in next step



Temporal-GNNs predictions can be purely observational:
Modeling infections without modeling how virus spreads over contact network

Take-home: Temporal GNNs not enough to predict causal effects on graphs

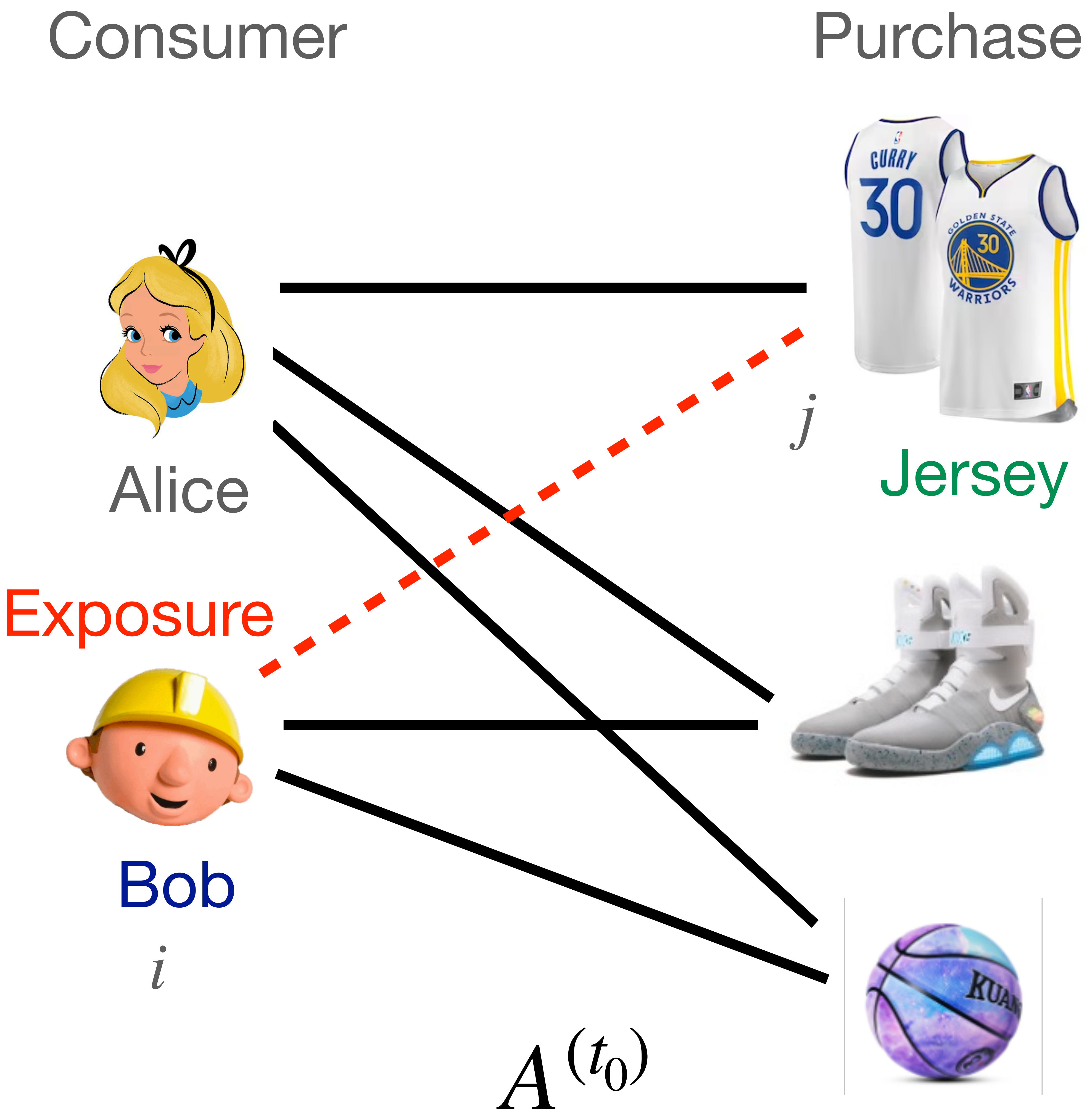
Bruno Ribeiro

Causality & Link Prediction

Link prediction as an exposure

- At time t_0 we expose Bob to Jersey
- We will define this intervention exposure as

$$E^{(t_0)} = (\text{Bob, Jersey})$$



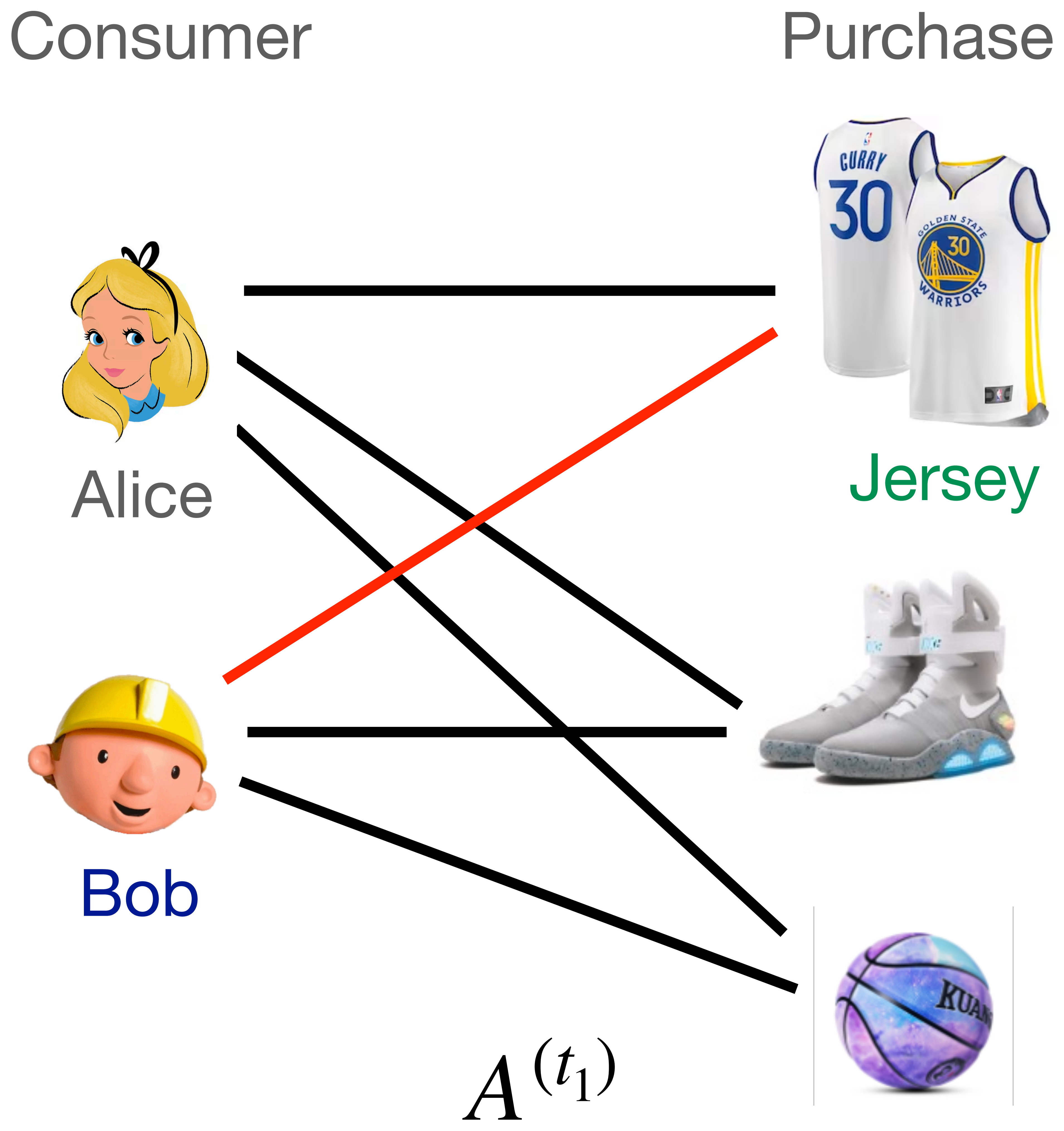
Recommendations as treatments
(Joachims et al., 2021)

Task: Link creation outcome

- At time t_1 we see if Bob **bought** Jersey

- The outcome of the exposure at t_1

$$A^{(t_1)}_{\text{Bob, Jersey}} \in \{0,1\}$$



Causal Identifiability

- $E^{(t_0)} = (i,j)$ is an exposure (intervention), $i, j \in V$

Edge at time t_1 Graph at time t_0

Exposure (intervention)

- Let $A^{(t_0)}$ be the graph at time t_0
- **Causal identifiability**: Can fit a predictive model for $A^{(t_1)}_{(i,j)}(E^{(t_0)} = (i,j)) \,|\, A^{(t_0)}$ on the available data?
- **Best Reply Model**
    - We will assume that outcome of an exposure is not strategic w.r.t. future outcomes

Graph Formation Process Key to Understand Effect of Exposures

Graph Formation Processes

Consider the formation process of an edge $A_{ij}(t)$ at time t

Simple Latent Factor Model

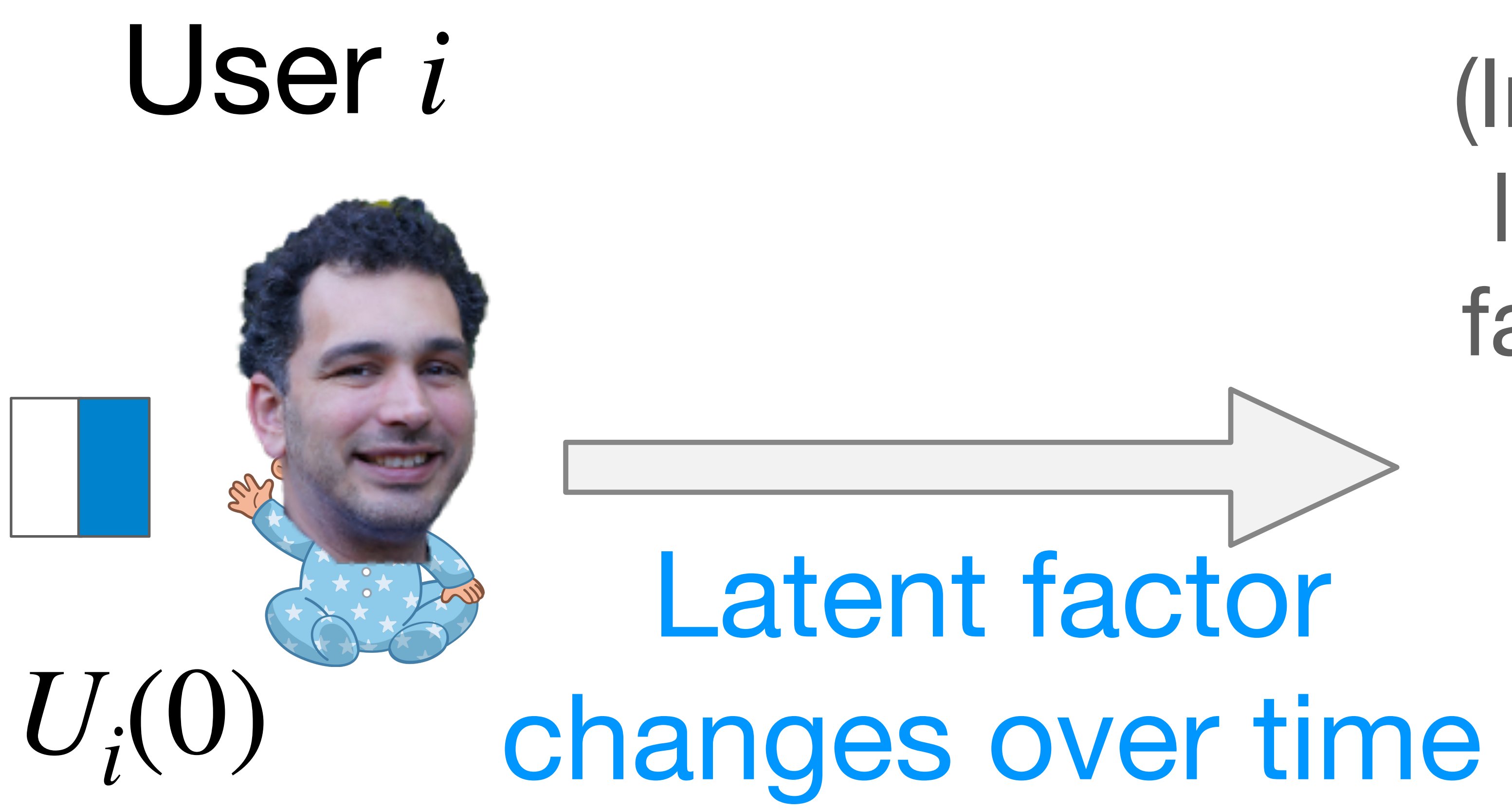
- $U_i(t) \sim g(t)$
- $U_j(t) \sim g(t)$
- $A_{ij}(t) \sim f(U_i(t), U_j(t))$

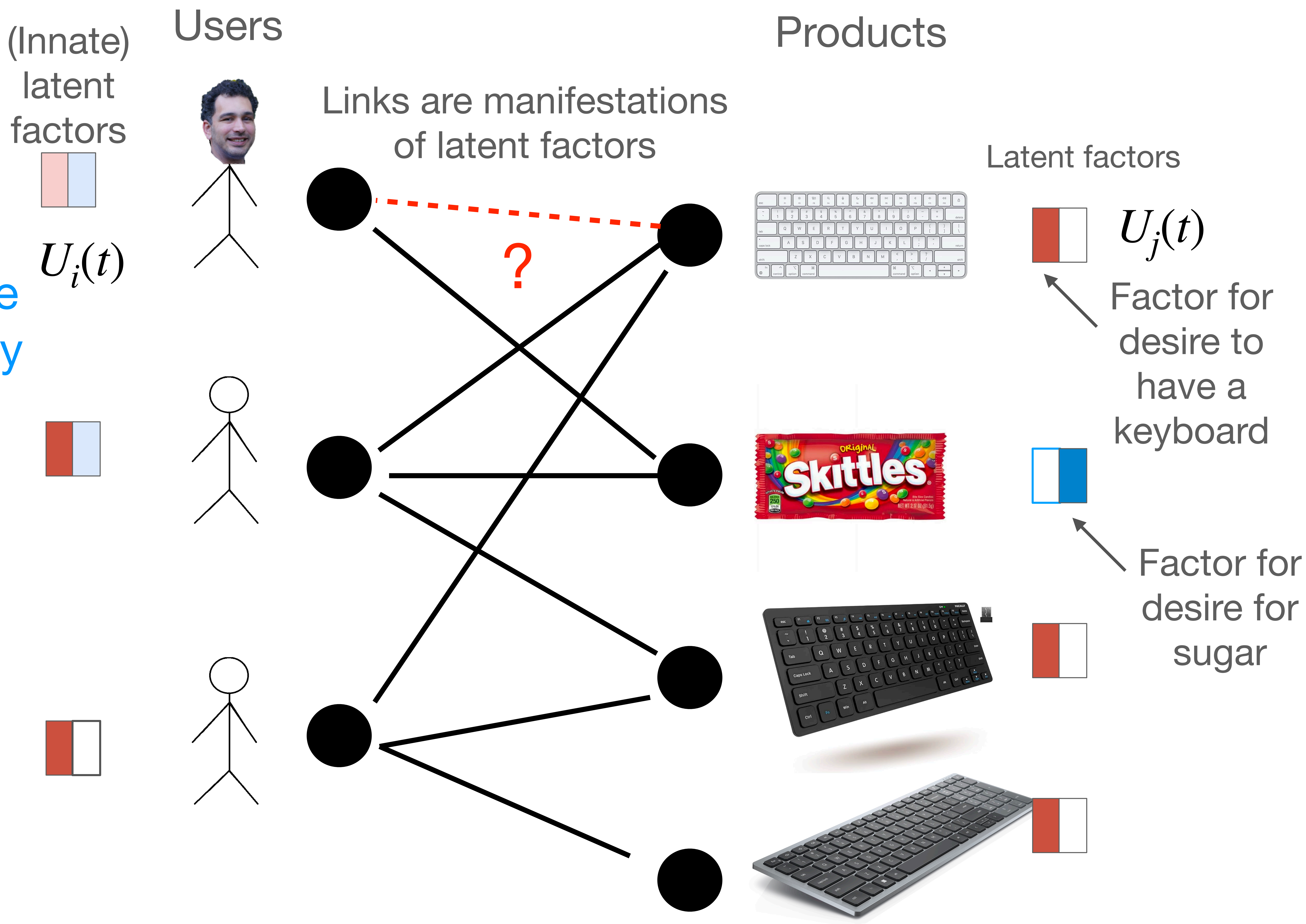
Simple Path-dependent Model

- $A_{ij}(t) \sim f(A(t - \Delta t), i, j)$

Most real graphs are both

Latent Factor Graph Formation





Latent Factor Model

- $U_i(t) \sim g(t)$
- $U_j(t) \sim g(t)$
- $A_{ij}(t) \sim f(U_i(t), U_j(t))$

Example: Probabilistic Factor Model

- Another common class of models are **factor models**
- Generally, these are consider only latent variables β_j and z_i that define how entity $j \in \{1,\ldots,m\}$ and element $i \in \{1,\ldots,n\}$ can interact
    - They are combined in the conditional distribution of causes,

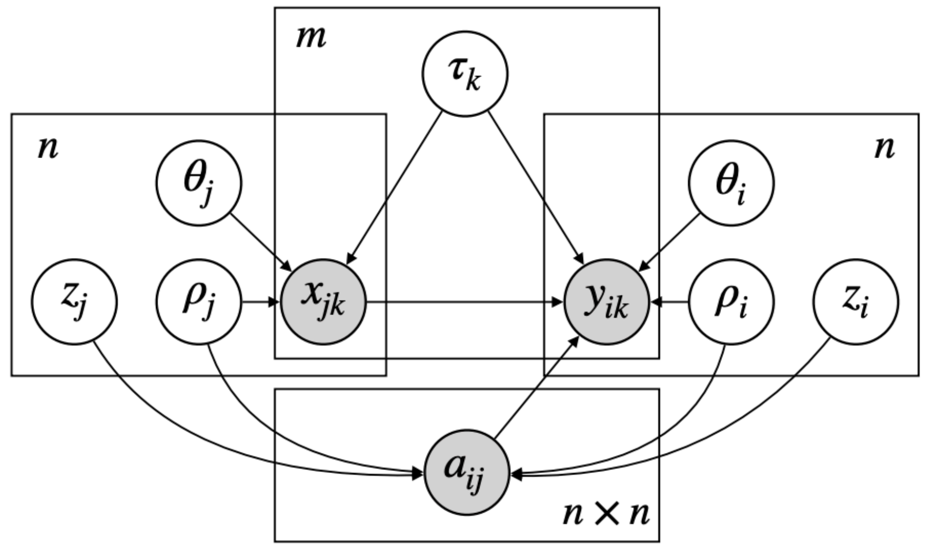
$$\beta_j \sim p(\beta)$$

$$z_i \sim p(z)$$

$$A_{ij} \sim p(a \,|\, z_i, \beta_j)$$

    - Given a dataset of edges (outcomes) estimate $p(\beta_{1:m}, z_{1:n} \,|\, A)$

More complex example:

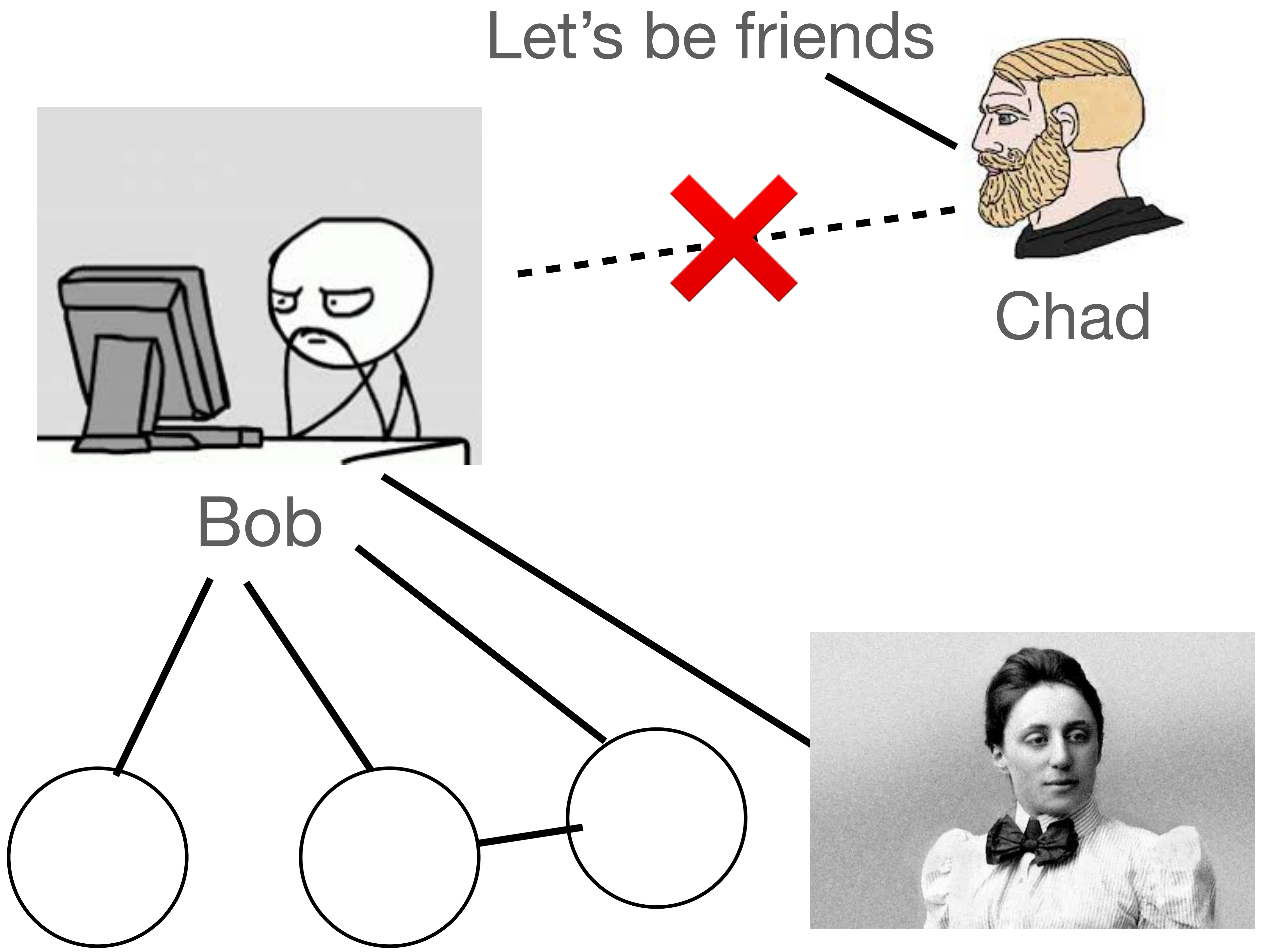


$x_{jk} \in \{0,1\}$: person j bought item k yesterday
$y_{ik} \in \mathbb{N}$: person i's consumption of item k today
$a_{ij} \in \{0,1\}$: person i is connected to person j
$\tau_k \in \mathbb{R}^P$: item k's attributes
$\theta_i \in \mathbb{R}^P$: person i's preferences for attributes
$z_i \in \mathbb{R}^D$: person i's traits that affect connections
$\rho_i \in \mathbb{R}^K$: person i's traits that drive connections and purchases

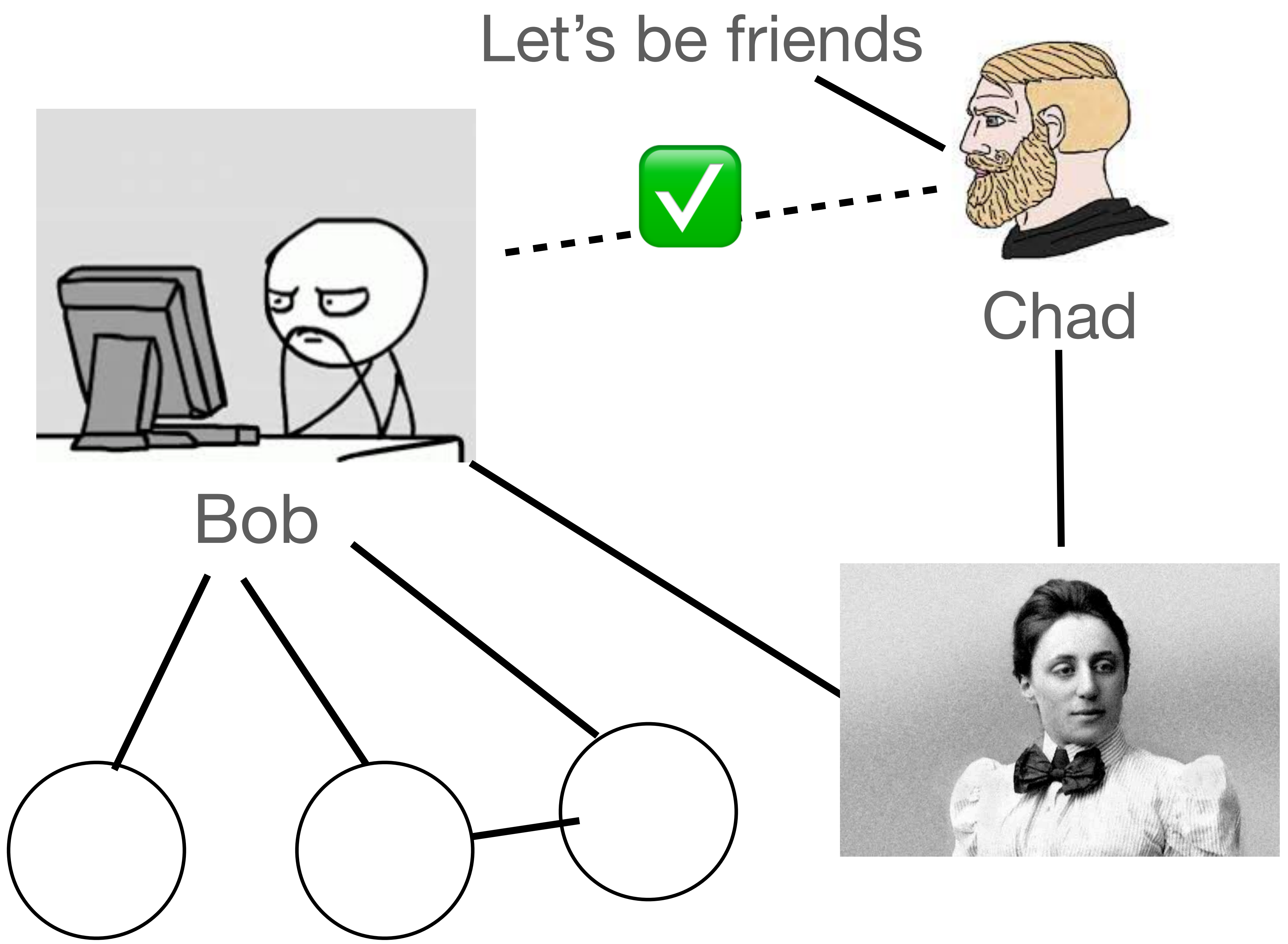
Poisson Influence Factorization
(Sridhar, De Bacco, Blei, 2022)

Path-dependent Graph Formation

- Path-dependency in graph evolution
  - **Graph evolution may depend on current state of graph**



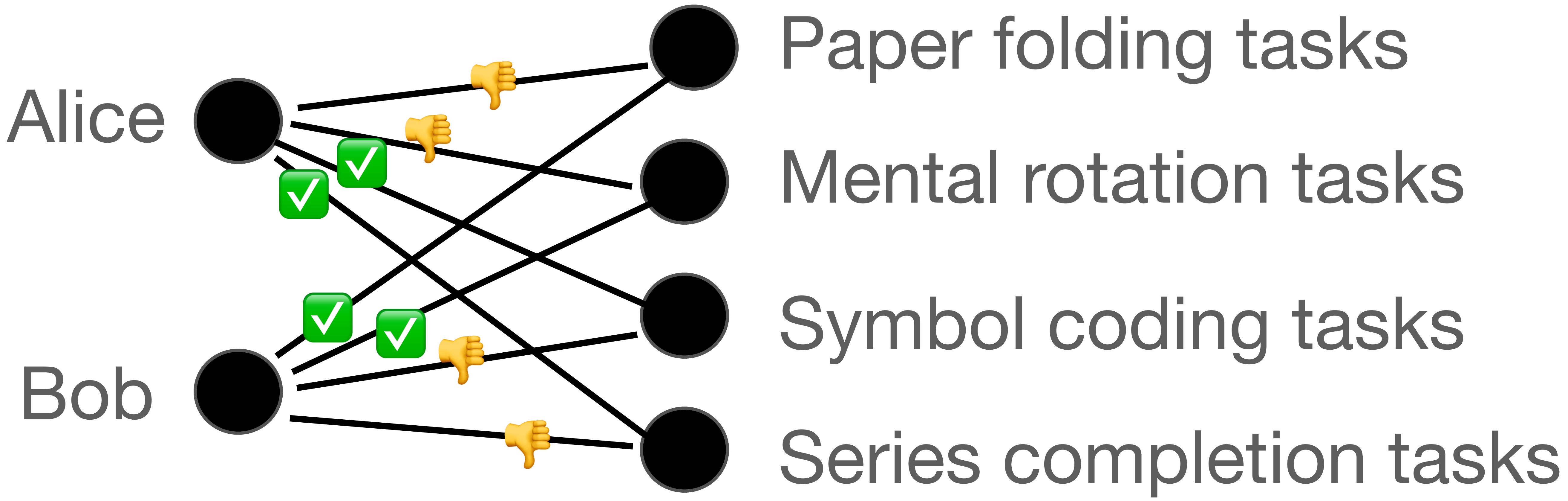
Scenario 1



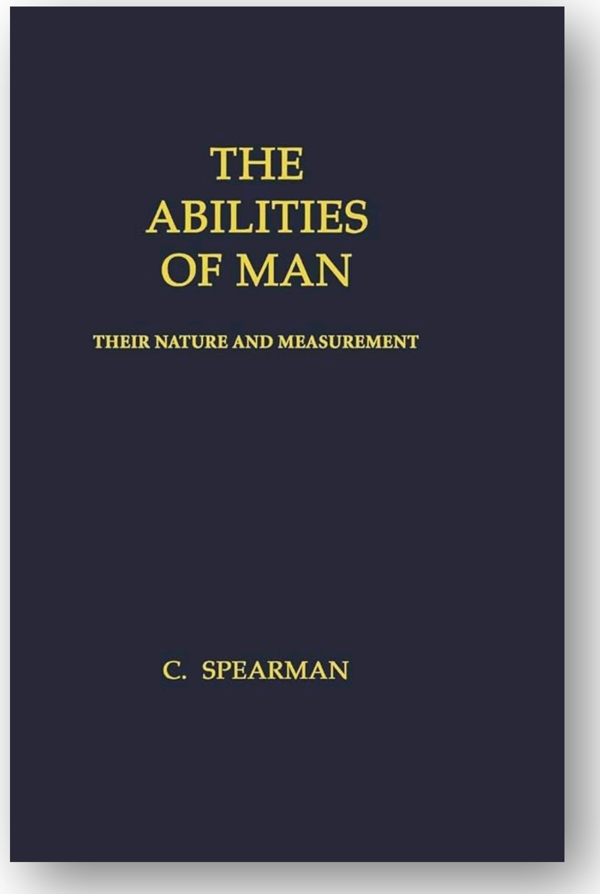
Scenario 2

Causal Interpretation of Latent Factors

- Factor models come from Spearman's *common factors of intelligence*
  - *Spearman* conjectures that **latent factors of intelligence** manifest as abilities to perform tasks



- In 1914 Woolley and Fischer's observed that *"boys are [innately] enormously superior [to girls] at spatial relations"*



- But Spearman (1927) disagreed with the conclusion: *"evidence [of this difference being] innate [rather than acquired] is still dubious"*.

Importance of Model on Predicting Exposures

Wikipedia

Under latent factors: Exposing Alice to spatial tasks does not improve her skills, but her skills may improve over time independently

Under path dependency: Exposing Alice to more spatial tasks may improve her skills

A Few Joint (Latent Factors + Path-dependent) Modeling Options

Causal identifiability under peer effects

(Goldsmith-Pinkham & Imbens, 2013) network formation process (e.g., Eq (5.1))

Strategic Network Formation Model whose next step adjacency matrix is $A_{ij}^{(t)} = \mathbf{1}(U_i(j) > 0) \cdot \mathbf{1}(U_j(i) > 0)$, where

$$U_i(j) = \alpha_0 + \alpha_x |X_i - X_j| + \alpha_\xi |\xi_i - \xi_j| + \alpha_d A_{ij}^{(t-1)} + \alpha_f F(A_{ij}^{(t-1)}, i, j) + \epsilon_{ij}$$

Covariates

Latent factors difference

Link between i,j at time $t-1$

Number of common neighbors in $A^{(t-1)}$

independent noise

Identification in social networks

(Graham, 2015)

Network formation process of link $i \to j$ at time t is

independent noise

$$A_{ij}^{(t)} = \mathbf{1}(\beta_0 A_{ij}^{(t-1)} + \gamma_0 F(A^{(t-1)}, i, j)) + M_{ij} - U_{ij} \geq 0)$$

with, for instance,

Common neighbors at time $t - 1$

$$M_{ij} = v_i + v_j - g(\xi_i, \xi_j)$$

Latent factor term

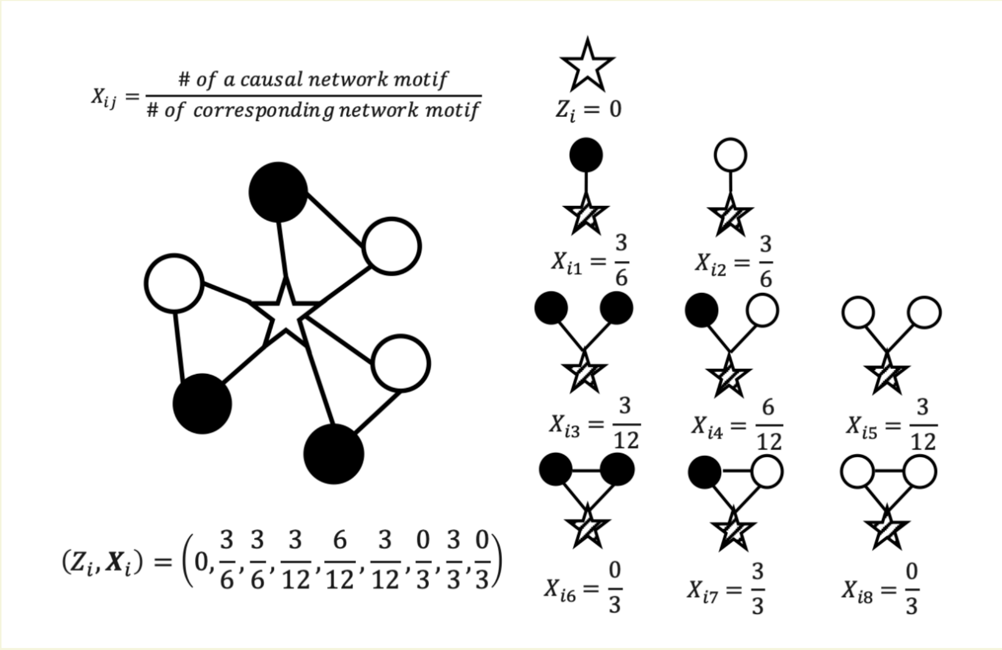
Could be replaced by more complex structural properties:

(Aronow, Samii, 2017)

(Yuan, Altenburger, Kooti, 2021)

(Leung, Loupos, 2023)

Example: rooted subgraphs



Overall Mechanism

For identifying the effect of an intervention (exposure) between $i, j \in V$

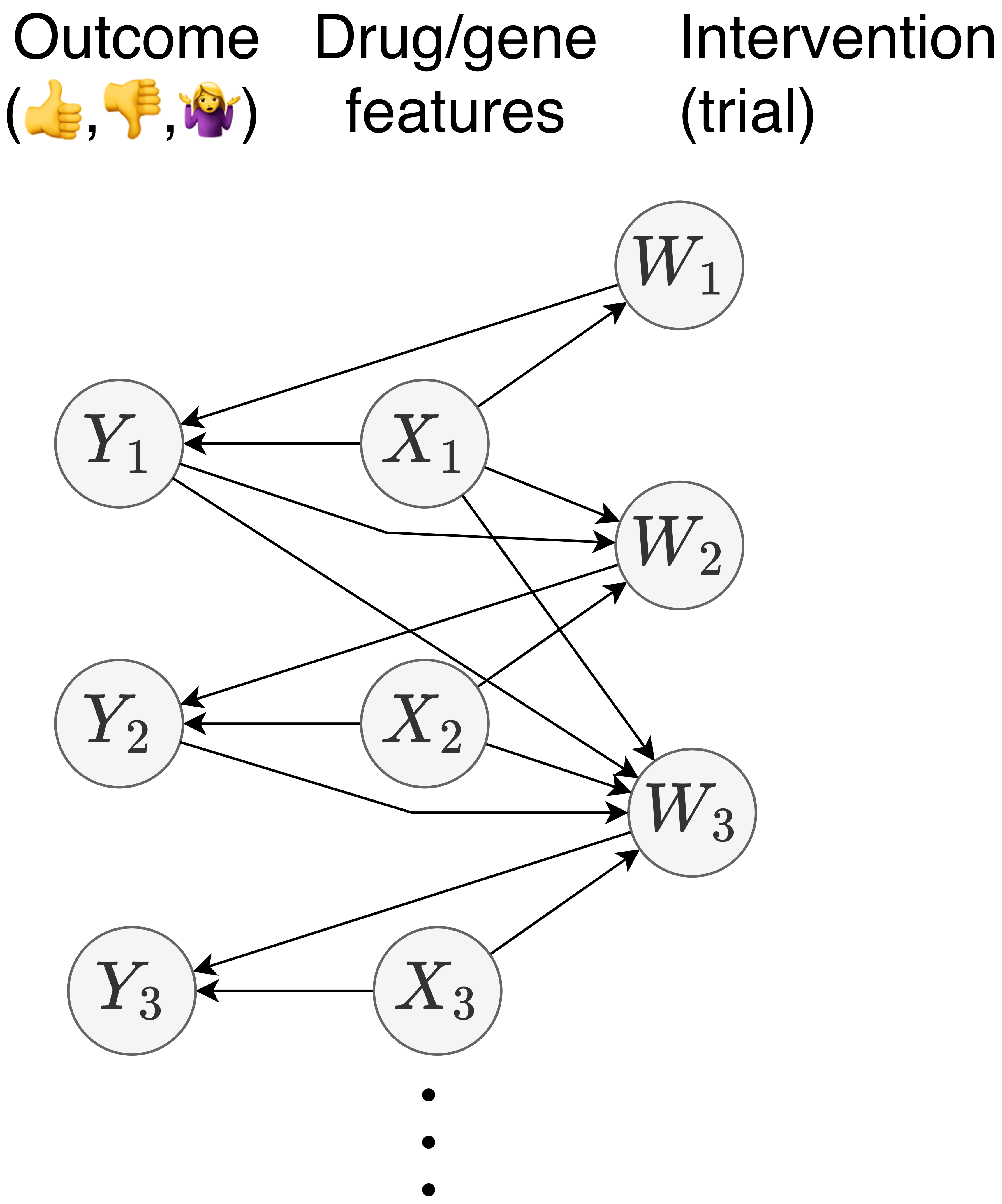
- $f(A^{(t_0)}, i, j)$: A structural characteristic of current graph $A^{(t_0)}$
- ξ_i, ξ_j : Some intrinsic factors of nodes

Then,

$$A_{ij}^{(t_1)} = g(\xi_i, \xi_j, f(A^{(t_0)}, i, j))$$

However, is
graph structure
enough for link prediction identifiability under cascading dependencies?

The Challenge of Cascading Dependencies



Someone intervening $W_j = 1$ may depend on features X_j and past success cases

Query: $P(Y_4 = y \,|\, X_4, \mathrm{do}(W_4 = 1))$

- May not be answerable with data due to cascading dependencies
- $Y_j \,|\, X_j, W_j$ depends on $Y_1, X_1, \ldots, Y_{j-1}, X_{j-1}$

“Universal” path-dependent graph formation

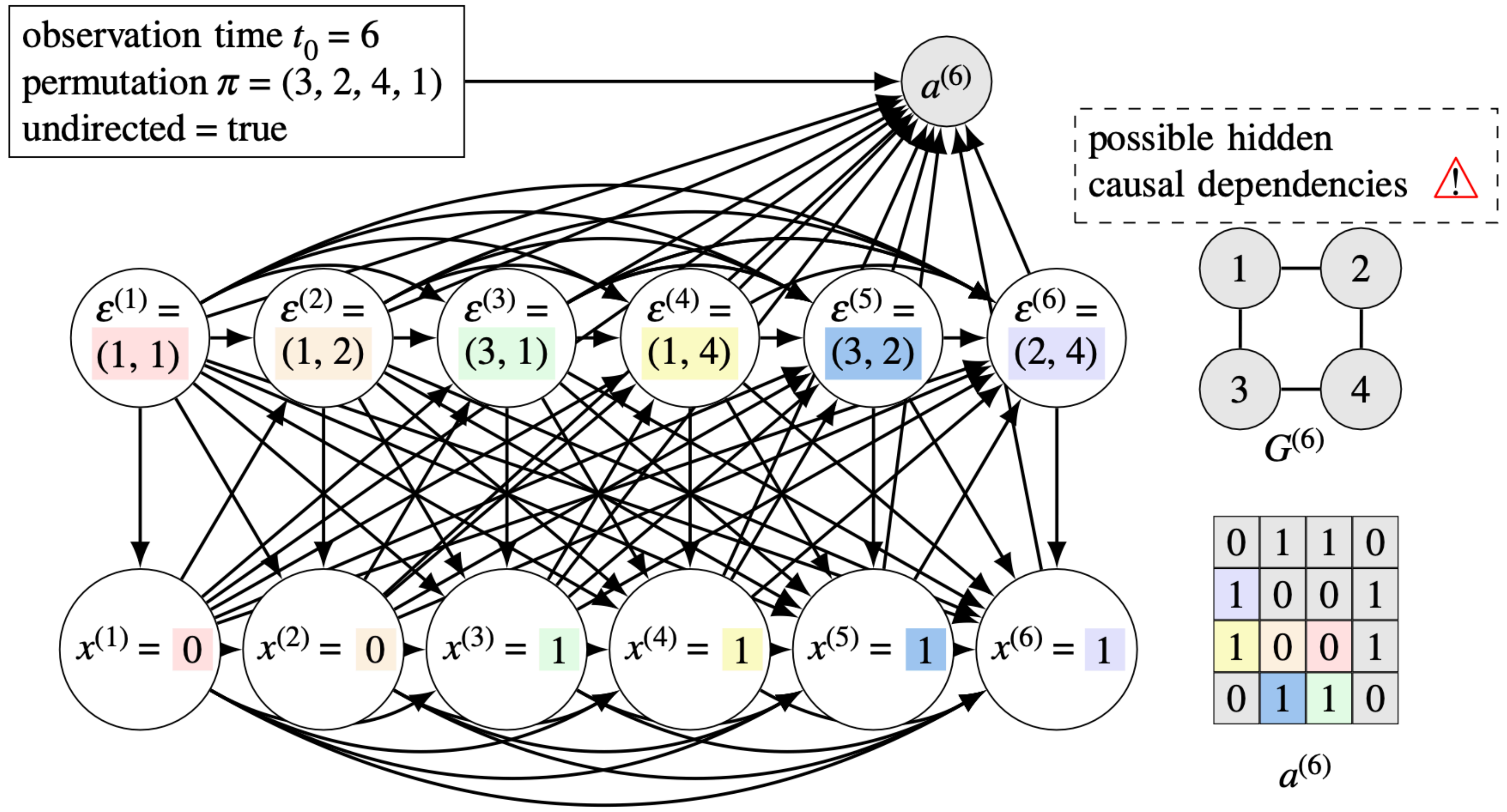
(Cotta, Bevilacqua, Ahmed, R., 2023)

Unobserved exposure (which pair is exposed next):

$$E^{(t)} = f_E^{(t)}\left((E^{(r)})_{r=1}^{t-1}, (A_{E^{(r)}}^{(r)})_{r=1}^{t-1}, U_E^{(t)} \right),$$

Edge from exposure: $$A_{E^{(t)}}^{(t)} = \begin{cases} f_A^{(t)}\left(U_A^{(t)} \right), & \text{if } t = 1, \\ f_A^{(t)}\left((E^{(r)})_{r=1}^{t}, (A_{E^{(r)}}^{(r)})_{r=1}^{t-1}, U_A^{(t)} \right), & \text{otherwise.} \end{cases}$$

$U_A^{(t)}$, $U_E^{(t)}$: independent exogenous variables



Causal Lifting

causal identifiability under cascading dependencies

(Cotta, Bevilacqua, Ahmed, R., 2023)

Defining invariances through groups

A group G is a set together with a binary operation $\star$ such that:

- Closure holds i.e., $\forall a, b \in \mathcal{G}, \;\; a \star b \in G$
- Associativity holds $(a \star b) \star c = a \star (b \star c) \;\; \forall a, b, c \in G$
- Identity element exists i.e., $\exists e \in \mathcal{G}$ s.t. $a \star e = e \star a = a \;\; \forall a \in G$
- Inverse exists for every element and $a \star a^{-1} = a^{-1} \star a = e \;\; \forall a \in G$

Credit: Bala Srinivasan

(Left) Group actions

For a group $\mathscr{G}$, binary operation $\star$, and with identity e, and a set X, a (left) group action is a function $\circ : \mathscr{G} \times X \to X$, such that

- $e \circ x = x, \;\; \forall x \in X$
- $g \circ (h \circ x) = (g \star h) \circ x, \;\; \forall g, h \in \mathscr{G}, \;\; \forall x \in X$

A function f is invariant to $\mathscr{G}$ (i.e. $\mathscr{G}$-invariant) if $f(x) = f(g \circ x), \forall g \in \mathscr{G}, \;\; \forall x \in X$

Credit: Bala Srinivasan

Causal Lifting (Cotta, Bevilacqua, Ahmed, R., 2023)

- **Associational** lifting: Let $\mathscr{G}$ be a group and $\circ$ is the left action of $\mathscr{G}$ onto supp(X)
  E.g., (Kimmig et al., 2014)

$$P(Y \mid X = x) = P(Y \mid X = g \circ x), \quad \forall g \in \mathscr{G}$$

- Definition 3.2 (**Interventional** lifting):

$$P(Y(X = x)) = P(Y(X = g \circ x)), \quad \forall g \in \mathscr{G}$$

- Definition 3.3 (**Counterfactual** lifting):

$$P(Y(X = x) \mid X = x') = P(Y(X = g \circ x) \mid X = x'), \quad \forall g \in \mathscr{G}$$

or

$$P(Y(X = x) \mid X = x', Y = y') = P(Y(X = g \circ x) \mid X = x', Y = y'), \quad \forall g \in \mathscr{G}$$

Sufficient invariances for identification in causal link prediction under cascading dependencies

Assumption 1: Gap Ignorability (informal)

We say that the universal SCM satisfies time gap ignorability if the mechanism $f_X^{(t_1)}$ is invariant to the SCM intermediate states between the time the intervention probe is performed t_0 and the instant before we see its effect in t_1

- Otherwise, we need to account for the intermediate states in the interval (t_0, t_1).
  - Difficulty if violated: ? (guess = Hard)

Assumption 2: Time Exchangeability (informal)

We say that the universal SCM satisfies time exchangeability if the mechanism $f_X^{(t_1)}$ is invariant to the order in which edges and nonedges have been generated

- Otherwise, we need a temporal graph
  - Difficulty if violated: ? (guess = Easy)

Assumption 3: Non-link Ignorability (informal)

We say that the universal SCM satisfies non-link ignorability if the mechanism $f_X^{(t_1)}$ is invariant to which pairs of nodes were generated as non-links or were not exposed yet at time t_0

- This is needed since the graph structure does not encode which pairs have been exposed
  - Difficulty if violated: ? (guess = Easy)

Assumption 4: Identifier Exchangeability (informal)

We say that the universal SCM satisfies identifier exchangeability if the mechanism $f_X^{(t_1)}$ is invariant to permutations of the node identifiers

- This is needed to define the data as a *graph* in a machine learning model.
  - Difficulty if violated: ? (guess = Hard)

If Assumptions 1-4 hold, then…

- Theorem 4.6 (Invariances for interventional lifting in link prediction).
  - Under Assumptions 1-4 in our Universal SCM then **causal lifting** can be used to obtain an **equivalent SCM** using just the **observed graph** A:
    - *Where* $W_{O_{IJ}}$ *is shared by all nodes structurally identical to the pair IJ*

Can be obtained by a special type of graph neural network

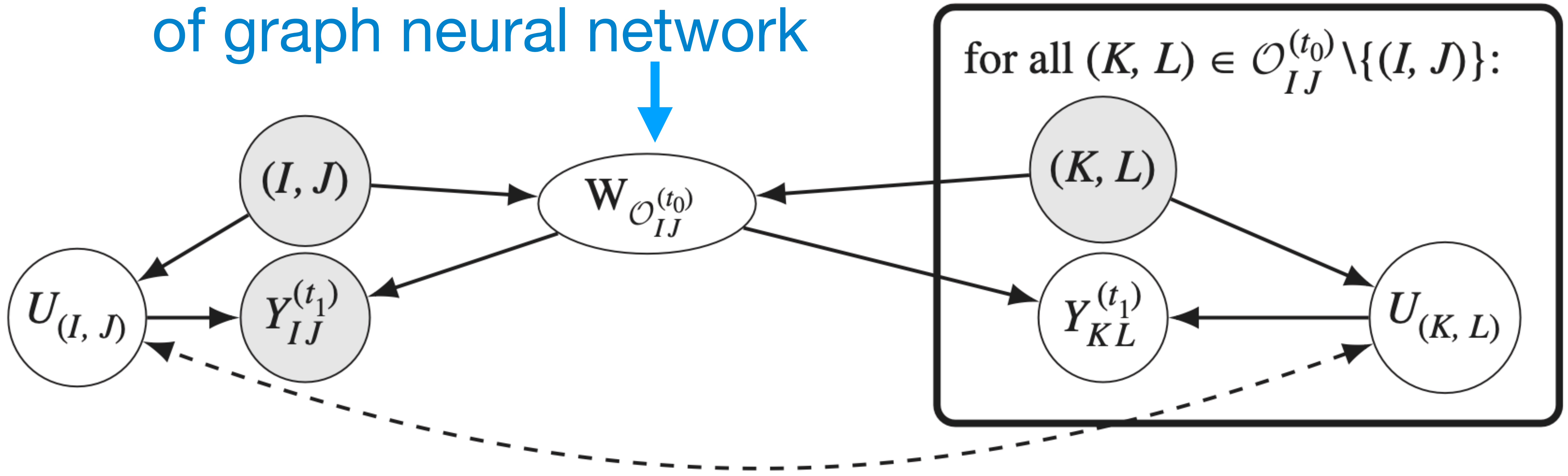
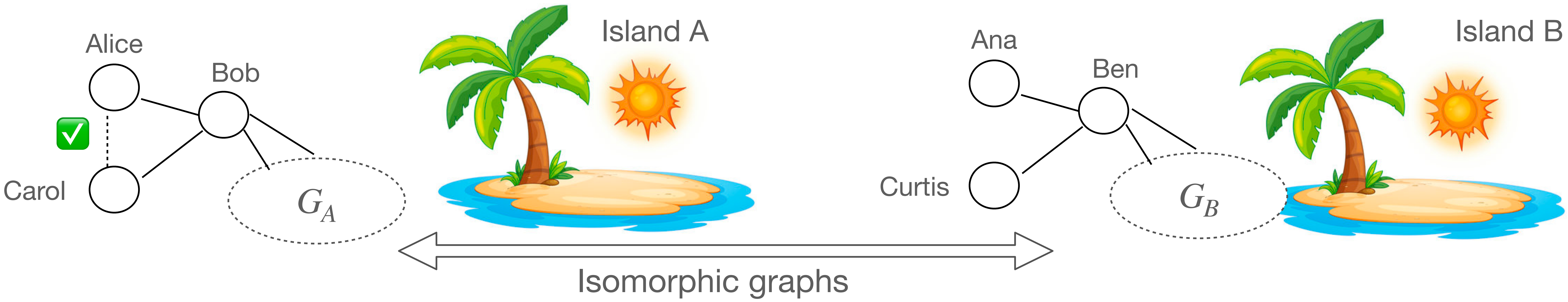


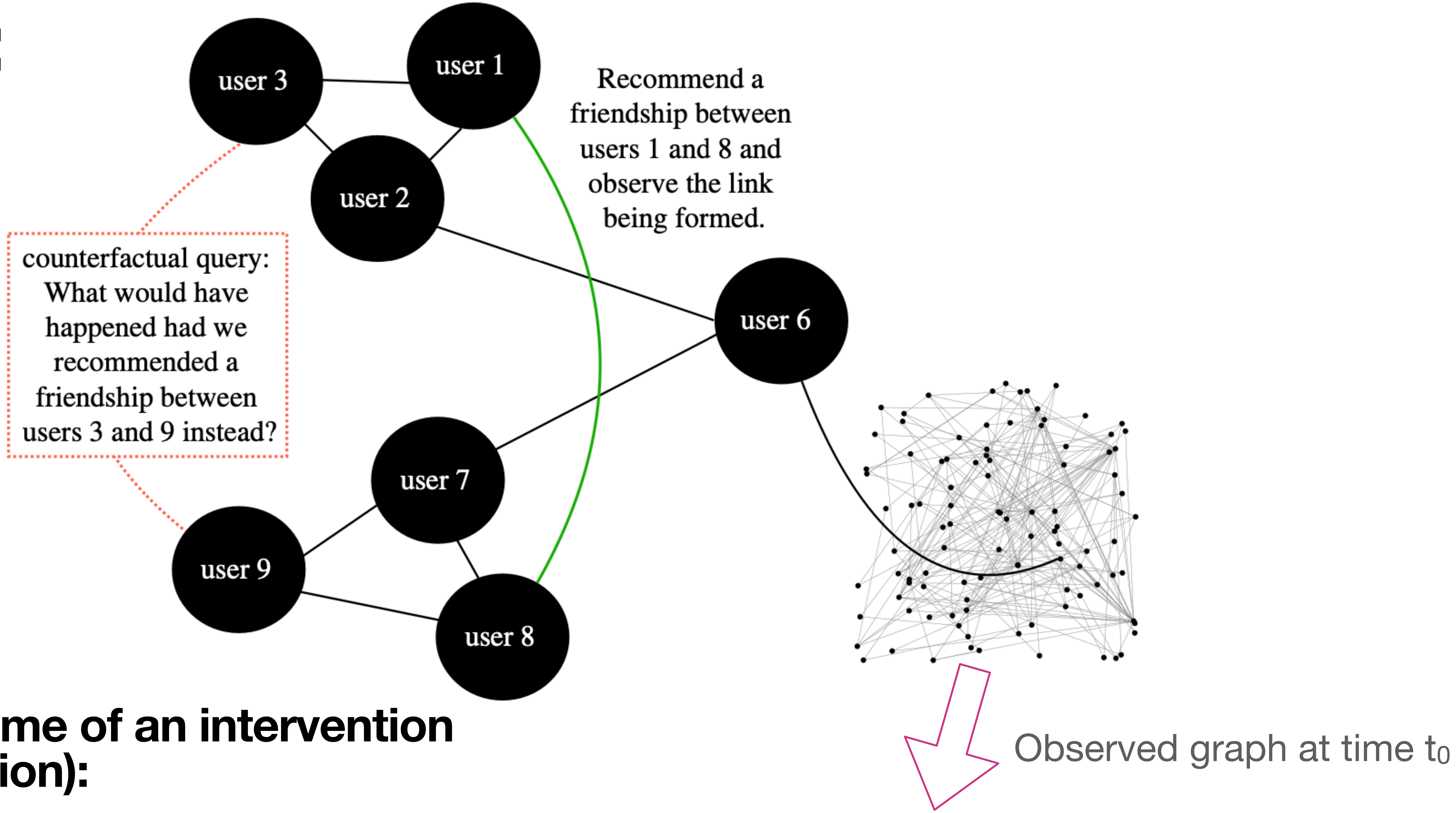
Figure 3. (Theorem 4.6(i)) Causal DAG of an equivalent data generating process of a probe in (i, j) (left) and in its orbit (right). As usual, we represent observed and unobserved variables with grey and white nodes, respectively.

How Causal Lifting + Assumptions 1-4 = Identifiability via GNNs



- Consider two deserted islands
  - Assume the same **structural causal model** generated the two social networks
    - Also, for now, the social graphs of Islands A and B will be **isomorphic**
  - Assume we suggest **Alice to Carol** and she accepts ✅
  - In island B, under assumptions 1-4 the suggestion of **Ana to Curtis** will have a similar outcome (in distribution)

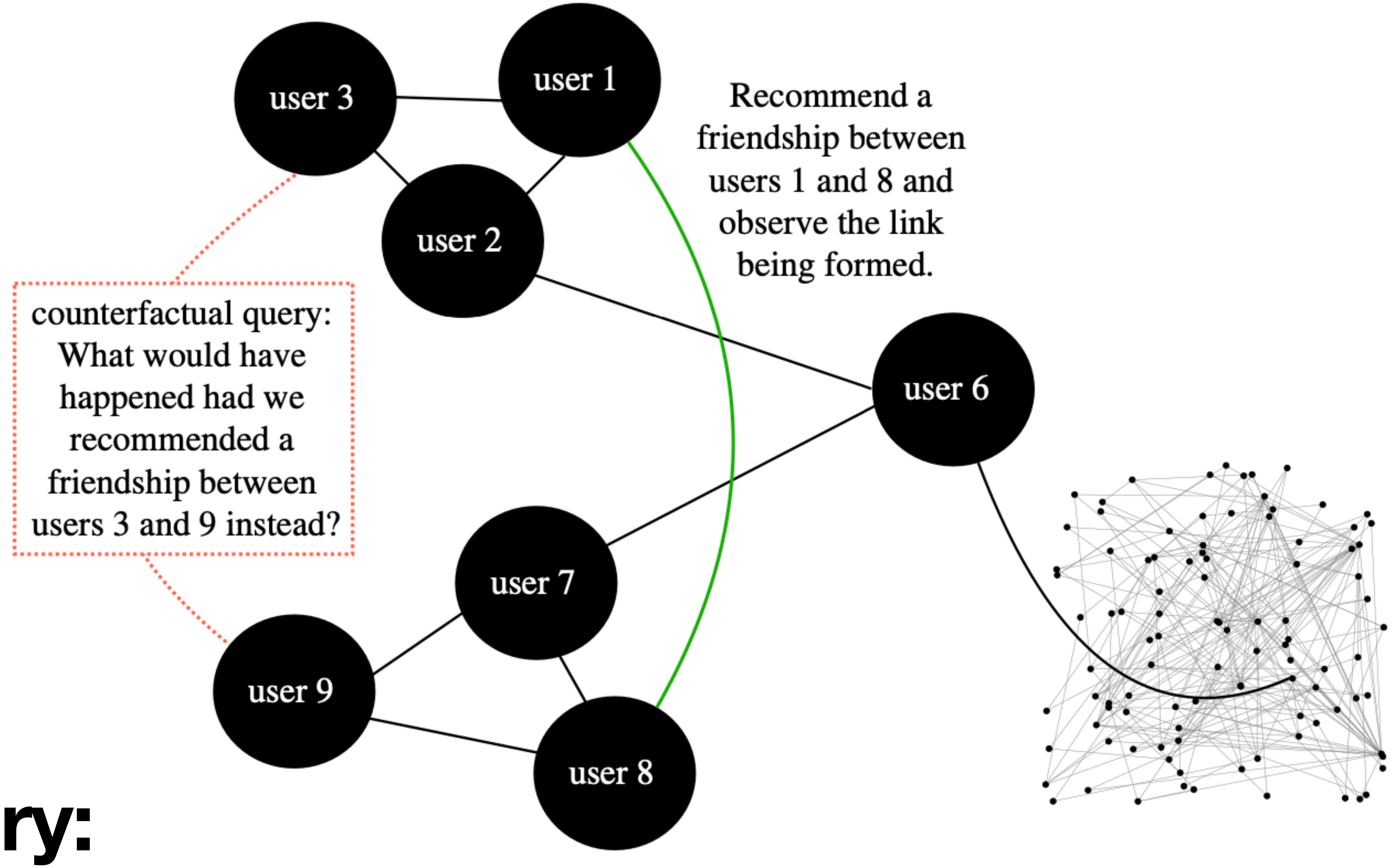
Example:



Observe outcome of an intervention (recommendation):

$$\underbrace{Y_{IJ}^{(t_1)}}_{\text{outcome of probe in } (8,1)} := A_{\mathcal{E}^{(t_1)}}^{(t_1)} \underbrace{\left(\mathcal{E}^{(t_1)} = (8,1)\right)}_{\text{probe in } (8,1)} \mid G^{(t_0)},$$

Identifiable



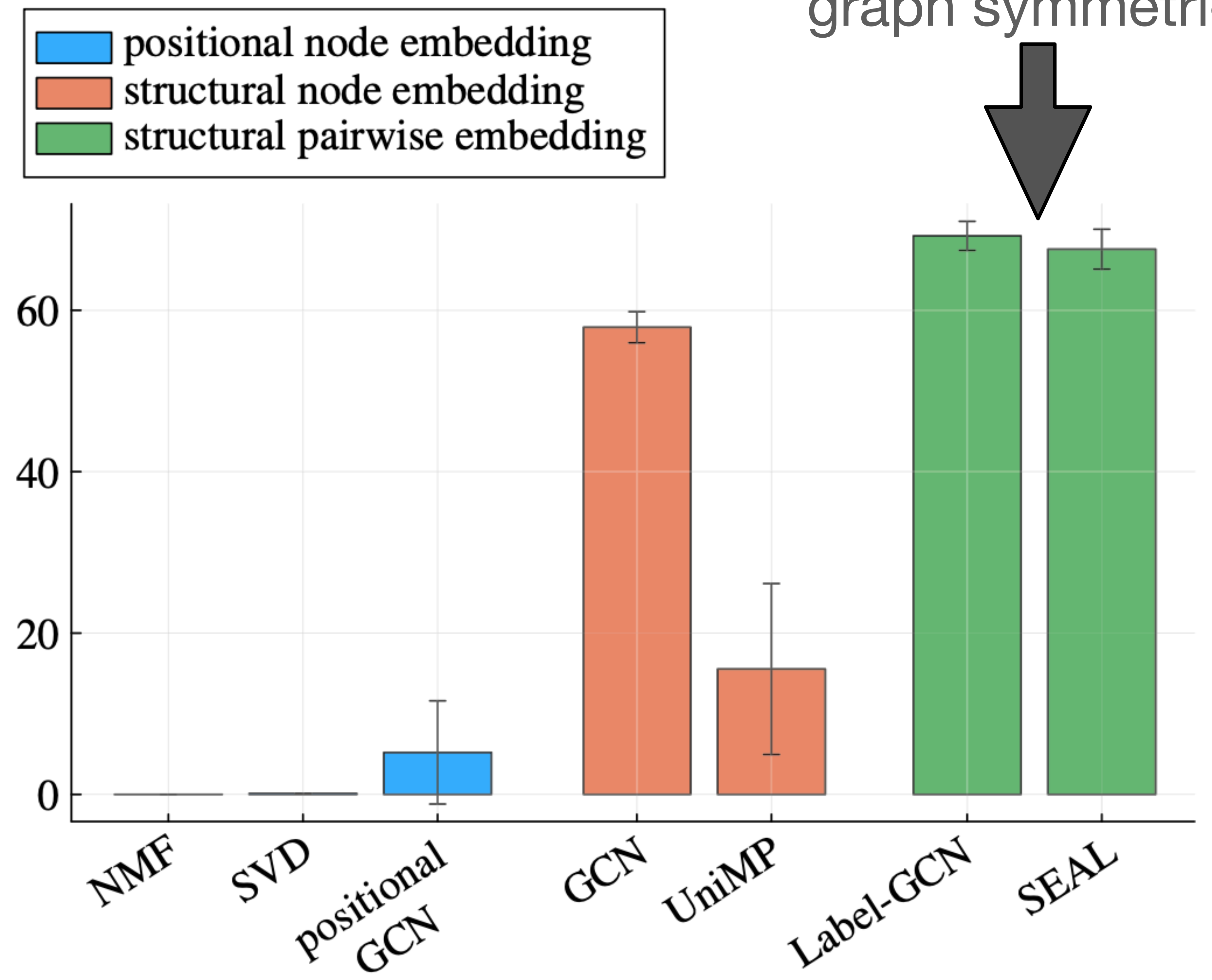
Counterfactual query:

$$P\Big(\underbrace{A^{(t_1)}_{\mathcal{E}^{(t_1)}}(\mathcal{E}^{(t_1)} = (3,9))}_{\text{what would have happened had we probed in (3,9) instead?}} \mid A^{(t_1)}_{\mathcal{E}^{(t_1)}}(\mathcal{E}^{(t_1)} = (8,1)), G^{(t_0)}\Big) \equiv P\Big(Y^{(t_1)}_{39} \mid Y^{(t_1)}_{81}\Big)$$

Example:

- Recommendations for Amazon purchases
  - In training we consider the subgroup of male users in recommendations.
  - At test time, our counterfactual queries are about female users.

Correctly encodes graph symmetries



Positional node embeddings

Structural pairwise embeddings

Summary

- Graph tasks that are used for decision-making are likely causal
  - Link prediction (for decision-making) is often a causal task
  - Temporal graph learning often not enough for decision-making
  - Causal lifting + invariances in graph formation process can tame cascading dependencies in causal graph learning