AAAI 2023 Tutorial

Machine Learning for Causal Inference

https://aaai23causalinference.github.io/

Zhixuan Chu[1], Jing Ma[2], Jundong Li[2], Sheng Li[2]

[1] Ant Group, Hangzhou, China

[2] University of Virginia, Charlottesville, USA

Tuesday, February 7, 2023

Agenda

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causality-aided Machine Learning
5. Applications and Future Directions
6. Conclusions

1. **Background on Causal Inference**
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causality-aided Machine Learning
5. Applications and Future Directions
6. Conclusions

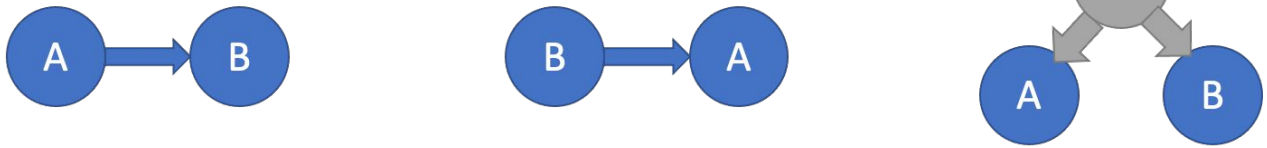
Causality

- *Causality* is also referred to as “causation”, or “cause and effect”
- Causality has been extensively discussed in many fields, such as statistics, philosophy, psychology, economics, education, and health care.

Figure: https://www.pinterest.com/pin/195906652514487365/

Correlation and Causation

- **Correlation does not imply causation**
- For two correlated events A and B, the possible relations might be: (1) A causes B, (2) B causes A, (3) A and B are consequences of a common cause, but do not cause each other, etc.



- Example of (3): As ice cream sales increase, the rate of drowning deaths increases sharply. The two events are correlated. However, increasing ice cream consumption and drowning deaths may not have causal relationships.

Causal Inference

- **Causal inference** is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect
- Two major tasks in causal inference
  - **Treatment Effect Estimation (This Tutorial)**: estimate the causal effects of an intervention on subjects, e.g., the effects of medication
  - **Causal Discovery**: infer causal structure from data, i.e., finding causal relations among variables

Experimental Study vs. Observational Study

- **Experimental Study**
  - Randomized Controlled Trial (RCT)
  - Assignment of control/treated is random
  - Gold-standard for studying causal relationships
  - Expensive and time-consuming, e.g., A/B testing

- **Observational Study**
  - Assignment is NOT random
  - Approaches: structural causal models, potential outcome framework
  - Simple, efficient and interpretable, e.g., nearest neighbor matching

About This Tutorial

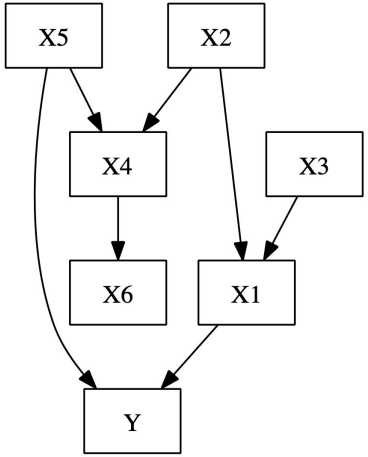
- Causal inference is an active research area with many research topics, this tutorial mainly focuses on the potential outcome framework in observational study
- Machine learning could assist causal inference at different stages. In this tutorial, we focus on how to design representation learning methods and graph neural networks for causal inference. Moreover, we will discuss causality-aided machine learning.

About This Tutorial

- **Schedule**
  - 8:30 AM - 9:15 AM: Background on Causal Inference (S. Li)
  - 9:15 AM - 10:00 AM: Representation Learning based Methods (Z. Chu)
  - 10:00 AM - 10:30 AM: Coffee Break
  - 10:30 AM - 11:10 AM: Graph Neural Networks based Methods (J. Li)
  - 11:10 AM - 11:40 AM: Causality-aided Machine Learning (J. Ma)
  - 11:40 AM - 12:00 AM: Applications, Future Directions, and Closing Remarks (S. Li)
- **Website:** https://aaai23causalinference.github.io/

Causal Inference Paradigms

- **Graphical Causal Models**
  - Causal graphs are probabilistic graphical models to encode assumptions about data-generating process *[Pearl, 2009]*
  - Related approach: structural equation modeling (SEM)
- **Potential Outcome Framework**
  - An approach to the statistical analysis of cause and effect based on the potential outcomes *[Rubin, 2005]*
  - Also known as Rubin causal model (RCM), or Neyman–Rubin causal model



An Example of Causal Graph

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.
[Rubin, 2005] Donald Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2005.

Potential Outcome Framework (1)

- **Unit**: A unit is the atomic research object in the causal study
- **Treatment**: An action that applies to a unit

  In the binary treatment case (i.e., $W = 0$ or 1), *treated* group contains units received treatment $W = 1$, while *control* group contains units received treatment $W = 0$
- **Outcome**: response of units after treatment/control, denoted as Y
- **Treatment Effect**: The change of outcome when applying the different treatments on the units

An Illustrative Example

- **Task:** Evaluate the treatment effects of several different medications for one disease, by exploiting the observational data, such as the electronic health records (EHR)
- **Observational data** may include: (1) demographic information of patients, (2) specific medication with the specific dosage taken by patients, and (3) the outcome of medical tests
- **Units**: patients
- **Treatments**: different medications
- **Outcome**: recovery, blood test results, or others

Potential Outcome Framework (2)

- **Potential Outcome**: For each unit-treatment pair, the outcome of that treatment when applied on that unit is the potential outcome. $Y(W=w)$
- **Observed Outcome**: Outcome of treatment that is actually applied. In binary case,
- **Counterfactual Outcome**: Potential outcome of the treatments that the unit had not taken. In binary case,
- A unit can only take one treatment. Thus, counterfactual outcomes are not observed, leading to the well-known "*missing data*" problem

Potential Outcome Framework (3)

- **Treatment Effects** can be defined at the population, treated group, subgroup and individual levels
- *Population Level*: Average Treatment Effect (**ATE**)

$$\text{ATE} = \mathbb{E}[\mathbf{Y}(W=1) - \mathbf{Y}(W=0)]$$

- *Treated group*: Average Treatment Effect on the Treated Group (**ATT**)

$$\text{ATT} = \mathbb{E}[\mathbf{Y}(W=1)|\mathbf{W}=1] - \mathbb{E}[\mathbf{Y}(W=0)|\mathbf{W}=1]$$

- *Subgroup*: Conditional Average Treatment Effect (**CATE**)

$$\text{CATE} = \mathbb{E}[\mathbf{Y}(W=1)|X=x] - \mathbb{E}[\mathbf{Y}(W=0)|X=x]$$

- *Individual*: Individual Treatment Effect (**ITE**)

$$\text{ITE}_i = Y_i(W=1) - Y_i(W=0)$$

Assumptions

- **Assumption 1**: Stable Unit Treatment Value Assumption (**SUTVA**)

  The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

- This assumption emphasizes that:
  - **Independence of each unit**, i.e., there are no interactions between units. In our example, one patient's outcome will not affect other patients' outcomes
  - **Single version for each treatment**. For instance, one medicine with different dosages are different treatments under the SUTVA assumption

Assumptions

- **<u>Assumption 2</u>**: **Ignorability**

  *Given the background variable, **X**, treatment assignment W is independent of the potential outcomes, i.e.,*

- Following our example, this assumption implies that:
  - If two patients have the same background variable ***X***, their potential outcome should be the **same** whatever the treatment assignment is.
  - Analogously, if two patients have the same background variable value, their ***treatment assignment mechanism*** should be same whatever the value of potential outcomes they have

Assumptions

- **<u>Assumption 3</u>: Positivity**

  *For any set of values of **X**, treatment assignment is not deterministic:*

  $$P(W = w|X = X) > 0 \quad \forall w \text{ and } x.$$

- If treatment assignment for some values of X is deterministic, the outcomes of at least one treatment could never be observed. It would be unable and meaningless to estimate causal effects
- It implies “common support” or “overlap” of treated and control groups
- The ignorability and the positivity assumptions together are also called ***Strong Ignorability*** or ***Strongly Ignorable Treatment Assignment***

A Naive Solution

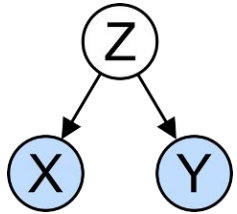
- The core problem in causal inference is: how to estimate the average potential treated/control outcomes over a specific group?
- One naive solution is to calculate the difference between the average treated and control outcomes, i.e.,

$$\hat{\mathrm{ATE}} = \frac{1}{N_T}\sum_{i=1}^{N_T} Y_i^F - \frac{1}{N_C}\sum_{i=1}^{N_C} Y_j^F$$

- However, this solution is not reasonable due to the existence of confounders

Confounders

- **Confounders**: Variables that affect both treatment assignment and outcome
- In the medical example, **age** is a confounder
  - Age would affect the recovery rate
  - Age would also affect the treatment choice



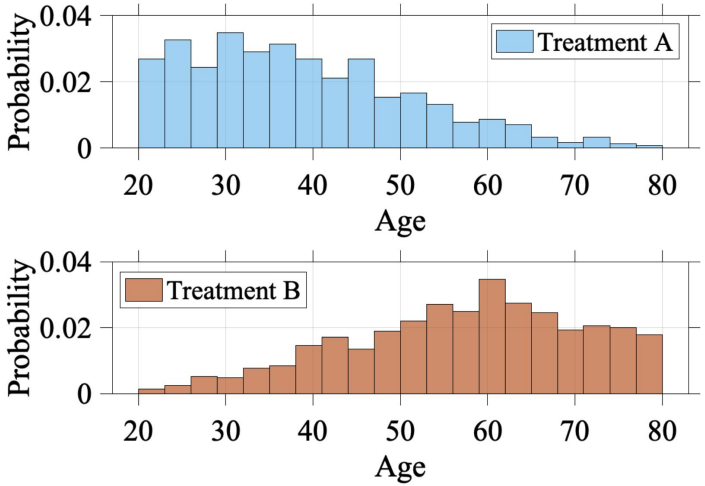
Recovery Rate / Treatment / Age	Treatment A	Treatment B
Young	234/270 = 87%	81/87 = **92%**
Older	55/80 = 69%	192/263 = **73%**
Overall	289/350 = **83%**	273/350 = 78%

Table 1. An example to show the spurious effect of confounder variable *Age*.

Simpson's paradox due to confounder

Selection Bias

- **Selection Bias**: The distribution of the observed group is not representative to the group we are interested in
- Confounder variables affect units' treatment choices, leading to selection bias
- Selection bias makes counterfactual outcome estimation more difficult



Classical Causal Inference Methods

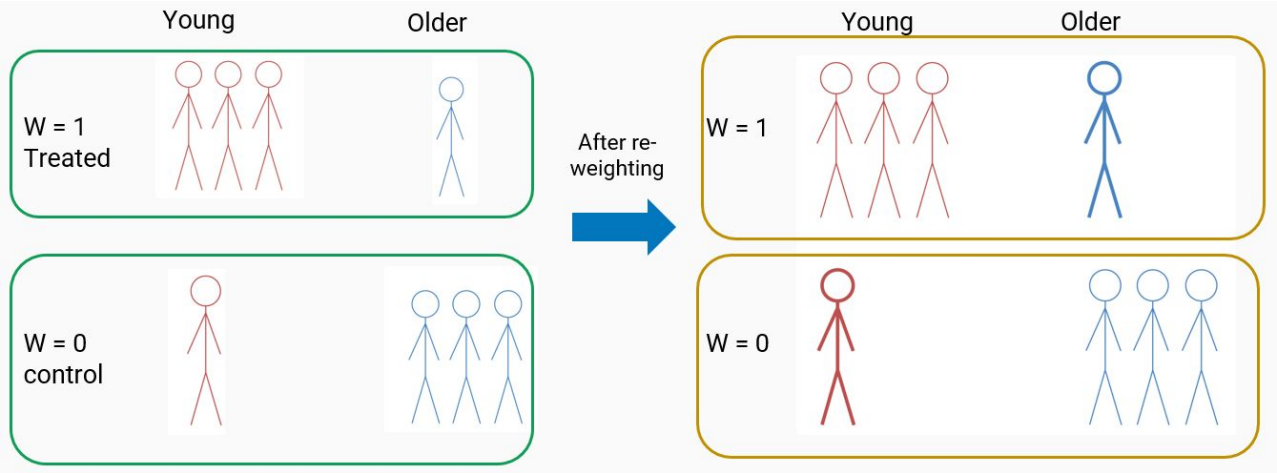
- Causal inference has been an active research area in statistics in the past several decades
- Categorization of Classical Methods
  - Re-weighting methods
  - Stratification methods
  - Matching methods
  - Tree-based methods

Re-weighting Methods

- **Challenge of Selection Bias**: due to different distributions of treated and control groups
- Sample re-weighting is a simple way to overcome the selection bias problem
- **Key Idea**: By *assigning appropriate weight to each sample* in the observation dataset, a ***pseudo-population*** is created on which the distributions of the treated group and control group are similar

Sample Re-weighting Methods

- Intuition example: **Age** (Yong/older) as the confounder
  - Young people: 75% chance of receiving treatment
  - Older people: only a 25% chance of receiving treatment



http://www.rebeccabarter.com/blog/2017-07-05-ip-weighting/

Stratification Methods

- **Stratification** adjusts the selection bias by splitting the entire group into subgroups, where within each subgroup, the treated group and the control group are similar under some measurements
- Stratification is also named as ***subclassification*** or ***blocking***
- ATE for stratification is estimated as

$$\hat{\tau}^{\text{strat}} = \sum_{j=1}^{J} q(j) \left[\bar{Y}_t(j) - \bar{Y}_c(j)\right]$$

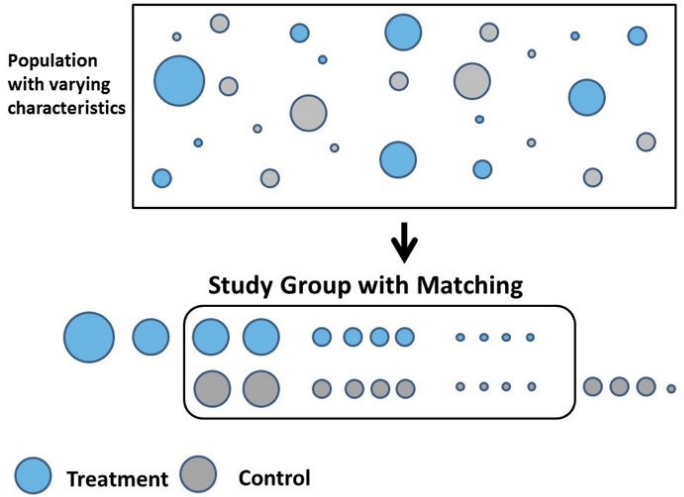
Matching Methods

- Matching methods estimate the counterfactuals and meanwhile reduce the estimation bias brought by the confounders
- Potential outcomes of the i-th unit estimated by matching are:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 1; \end{cases}$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1; \end{cases}$$

where *J(i)* is the matched neighbors of unit i in the opposite treatment group



Propensity Score Matching (PSM)

- Propensity scores denote conditional probability of assignment to a particular treatment given a vector of observed covariates.

$$e(x) = Pr(W = 1|X = x)$$

- Based on propensity scores, the distance between two units is

$$D(\mathbf{x}_i, \mathbf{x}_j) = |e_i - e_j|$$

- Alternatively, linear propensity score based distance metric

$$D(\mathbf{x}_i, \mathbf{x}_j) = |\text{logit}(e_i) - \text{logit}(e_j)|$$

P. Rosenbaum, and D. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

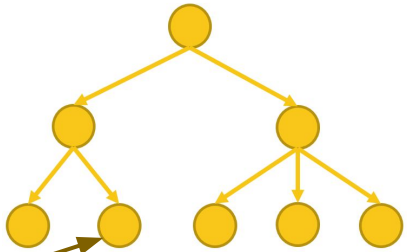
Tree-based Methods

- **Classification And Regression Trees (CART)**
  - Recursively partition the data space
  - Fit a simple prediction model for each partition
  - Represent every partitioning as a decision tree
- Leaf specific effect:

$$\mu(w, x \mid \Pi) \equiv \mathbb{E}\Big[Y_i(w) \mid X_i \in \ell(x \mid \Pi) \Big]$$

$$\tau(x \mid \Pi) \equiv \mu(1, x \mid \Pi) - \mu(0, x \mid \Pi)$$

A specific leaf node



P. Rosenbaum, and D. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.

Causal Forest

- Single tree is noisy -> using forest
- Forests = nearest neighbor methods + adaptive neighborhood metric
  - k-nearest neighbors: seek the k closest points to x according to some prespecified distance measure
  - Tree-based methods: closeness is defined with respect to a decision tree, and the closest points to x are those that fall in the same leaf

S. Wager, and S. Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113.523 (2018): 1228-1242.

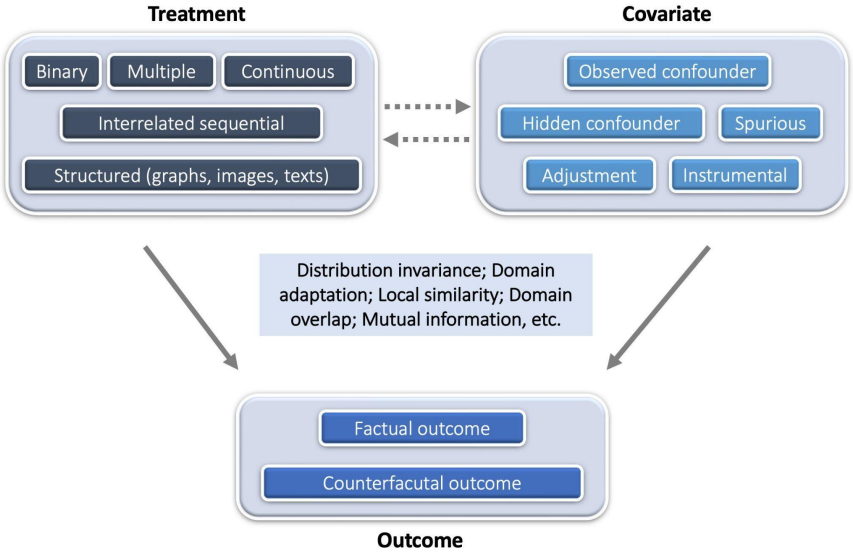
Why ML is Helpful for Causal Inference?

- **Machine Learning**
  - Various learning tasks, e.g., regression, classification, clustering
  - Various settings: multi-view, multi-task, transfer learning, etc.
  - Feature learning by shallow and deep models
- **Connections between Causal Inference and Machine Learning**
  - Matching in representation space
  - Covariate shift and group balancing
  - Counterfactual inference could be modeled as a regression problem

1. Background on Causal Inference
2. **Representation Learning based Methods**
3. Graph Neural Networks based Methods
4. Causality-aided Machine Learning
5. Applications and Future Directions
6. Conclusions

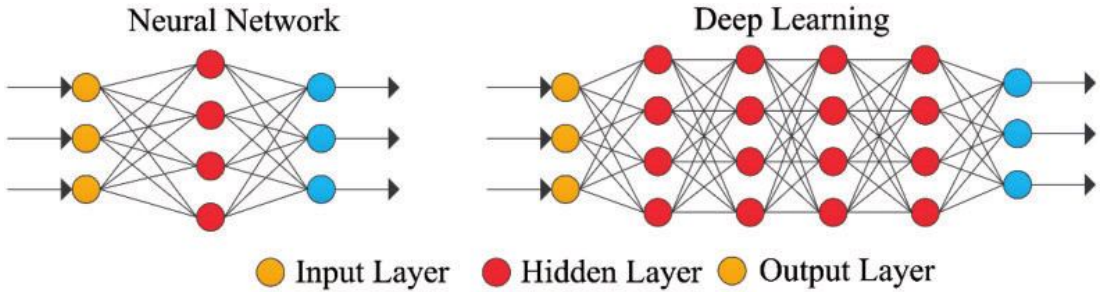
Representation Learning based Methods

- Traditional treatment effect estimation methods may not well handle **large-scale and high-dimensional heterogeneous data**
- Advanced machine learning approaches -> extraordinary performance
- New topics and new research questions from the core components of the treatment effect estimation task:
  - Treatment
  - Covariates
  - Outcome



Representation Learning

- Deep learning architecture is composed of an input layer, hidden layers, and an output layer
  - The output of each intermediate layer can be viewed as a representation of the original input data
  - Ability to deliver high-quality features and enhanced learning performance
  - Examples: Feed forward NN, CNN, Auto Encoder, VAE, GAN, etc.

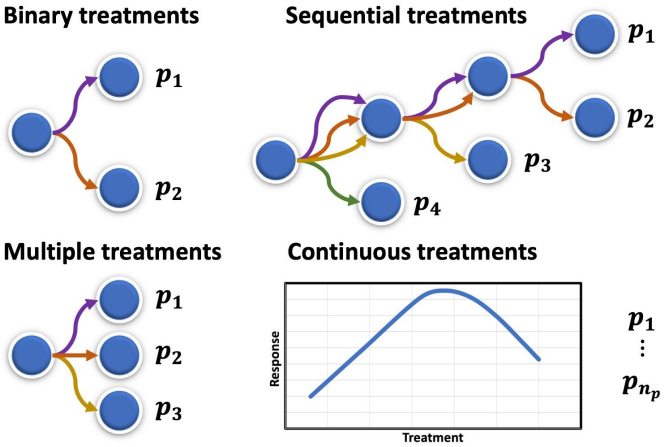


Treatment

- Treatment: How could we deal with different types of treatments?
  - (1) Binary
  - (2) Multiple
  - (3) Continuous scalar treatments
  - (4) Interrelated sequential treatments
  - (5) Structured treatments (e.g., graphs, images, texts)

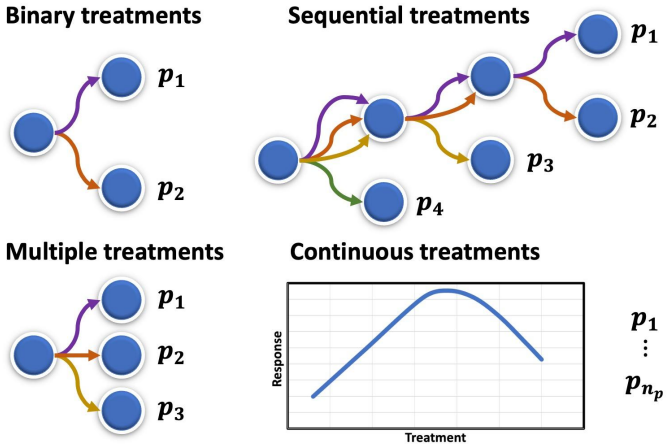
Binary, Multiple, Continuous, Sequential Treatments

- **A unified terminology**
  - Suppose that the observational data contain n units
  - Each unit goes through one potential path, including several treatment stages
  - In each potential path, the unit i can sequentially choose one of the two or multiple treatments T at each stage S, and finally, the corresponding outcome Y could be observed at the end of the path
  - Let $\{t_s^i; t_s = 1, ..., n_{t_s},\ i = 1, ..., n,\ \text{and}\ s = 1, ..., n_s\}$ denote the treatment assignment for unit i at stage s



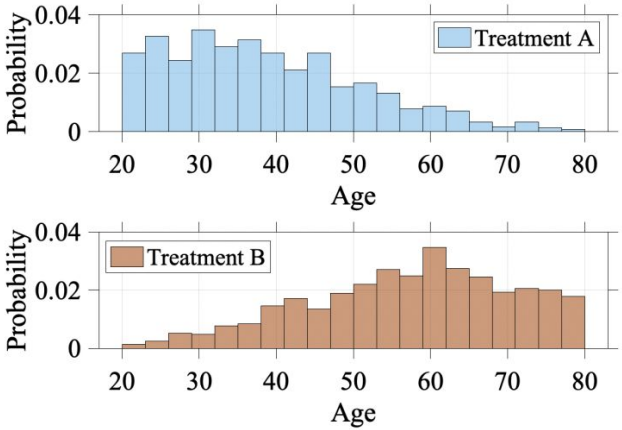
Binary, Multiple, Continuous, Sequential Treatments

- Exist several potential paths
- However, only one of the potential outcomes is observed at the end of the path according to the actual treatment assignments
- This observed outcome is called the **factual outcome**, and the remaining unobserved potential outcomes are called **counterfactual outcomes**

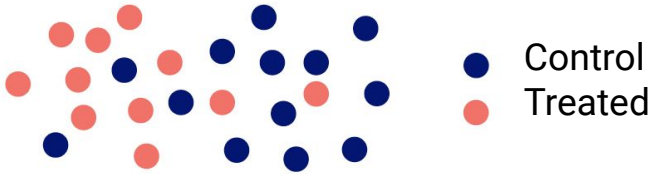


Binary Treatments

- Motivation
  - Selection Bias

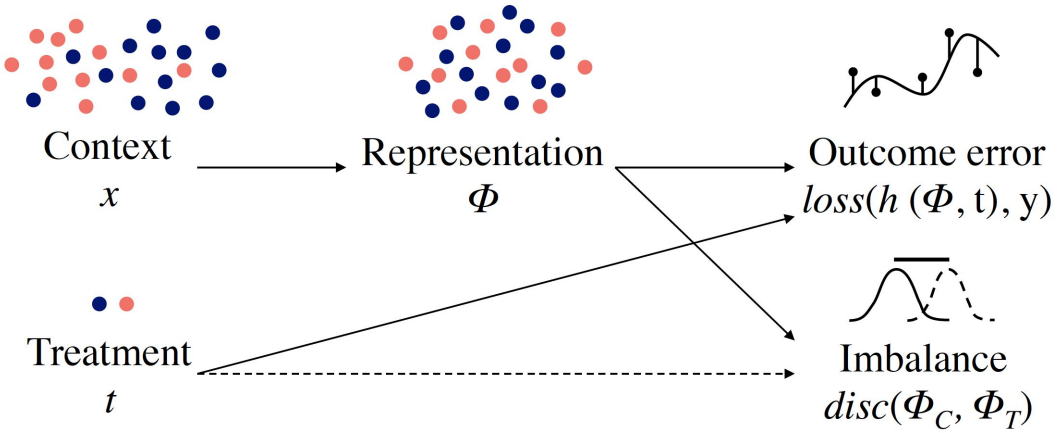


  - Counterfactual inference <-> Domain adaptation



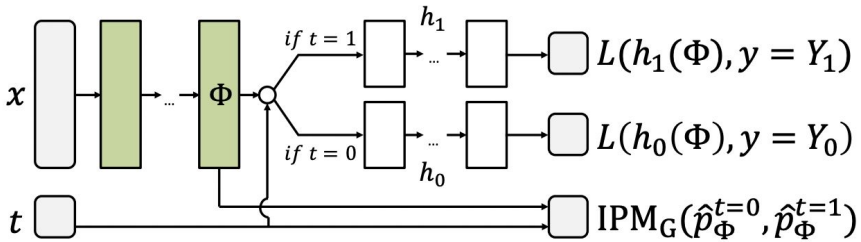
Binary Treatments

- Balancing the two groups in the representation space



Binary Treatments

- Counterfactual Regression



- Objective Function

Factual loss

$$\min_{\substack{h,\Phi \\ \|\Phi\|=1}} \frac{1}{n}\sum_{i=1}^{n} w_i \cdot L\left(h(\Phi(x_i), t_i), y_i\right) + \lambda \cdot \Re(h)$$

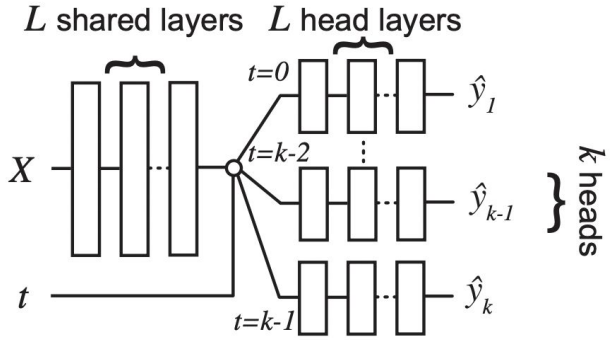
Discrepancy

$$+ \alpha \cdot \mathrm{IPM}_G\left(\{\Phi(x_i)\}_{i:t_i=0}, \{\Phi(x_i)\}_{i:t_i=1}\right)$$

U. Shalit, F. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Multiple Treatments

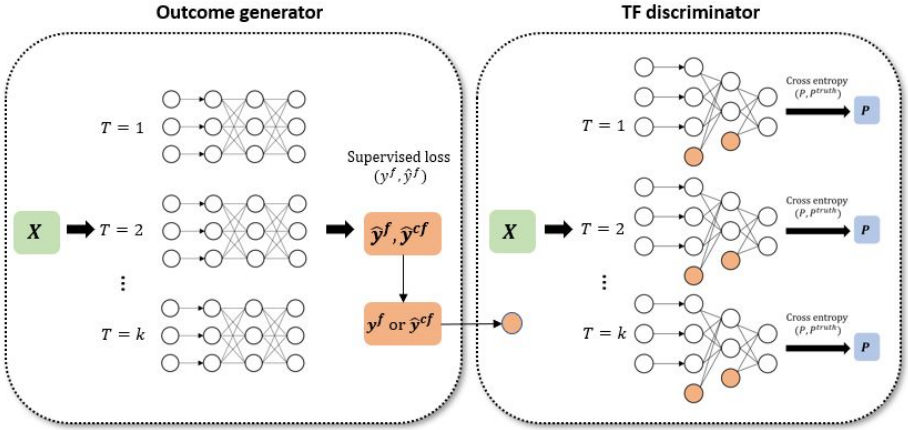
- Binary treatment models can be extended to multiple treatment models
  - Augment every sample within a minibatch with its closest matches by propensity score from the other treatment options
  - Use pairwise discrepancy distance to get balanced representations
  - Map to the common wasserstein barycenter



Schwab, P., Linhardt, L. and Karlen, W., 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Multiple Treatments

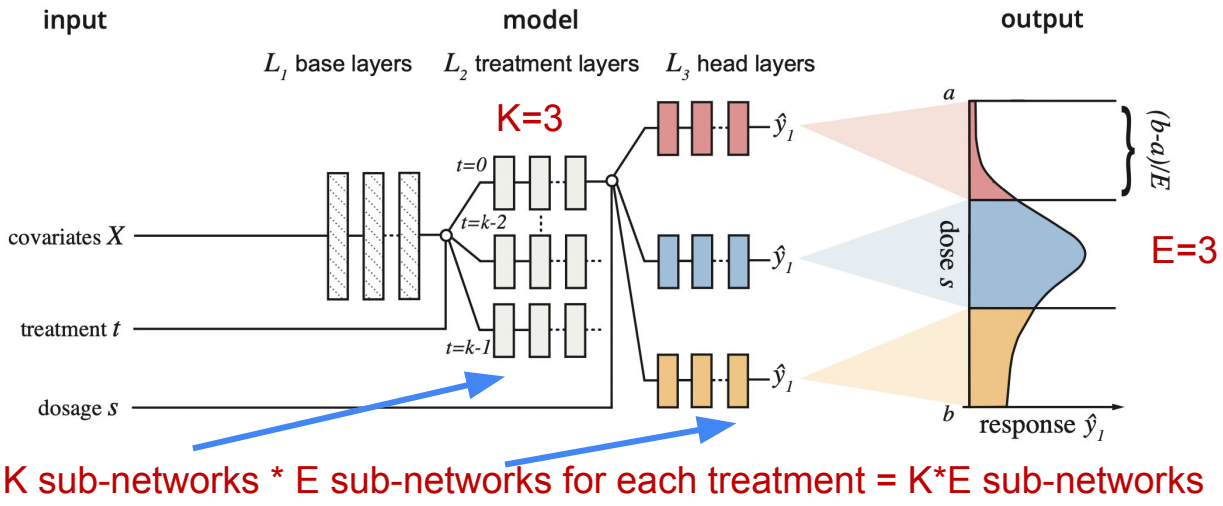
- Multi-task adversarial learning method
  - Outcome generator
  - True or false discriminator (TF discriminator)
  - These two models are trained together in a zero-sum game



Chu, Z., Rathbun, S.L. and Li, S., 2022, April. Multi-task adversarial learning for treatment effect estimation in basket trials. In *Conference on Health, Inference, and Learning* (pp. 79-91). PMLR.

Continuous Treatments

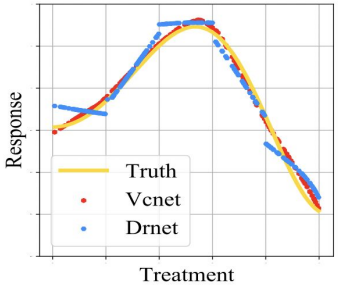
- Each head layer is assigned a dosage stratum that subdivides the range of potential dosages into partitions with equal width

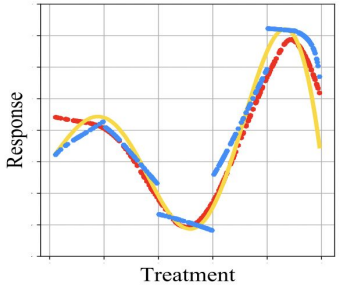


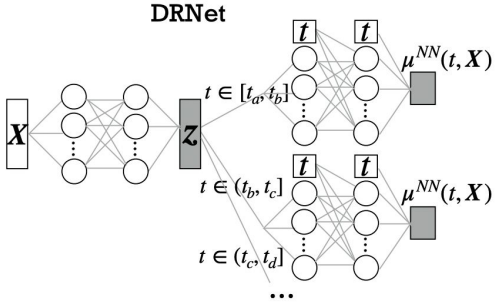
Schwab, P., Linhardt, L., Bauer, S., Buhmann, J.M. and Karlen, W., 2020, April. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 04, pp. 5612-5619).

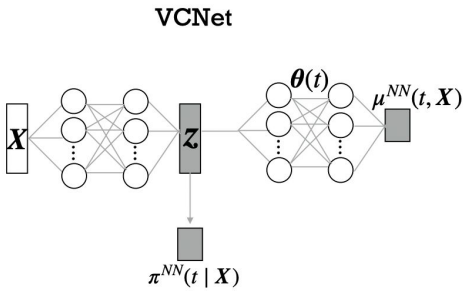
Continuous Treatments

- Continuity of dose-response curve





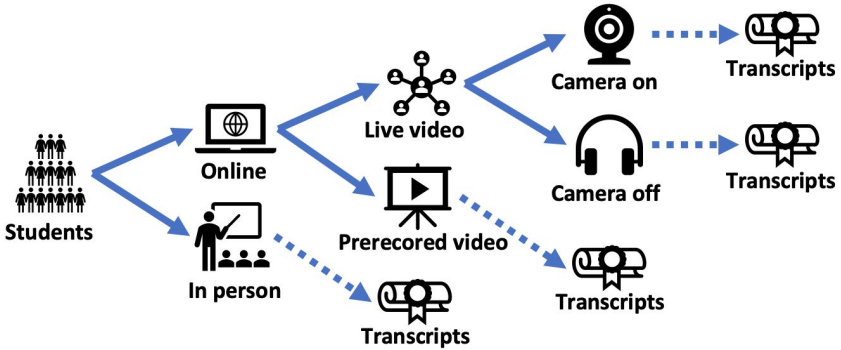




- Varying coefficient neural network (VCNet)
- Neural network with parameter θ(t) instead of a fixed θ $\mu^{\mathrm{NN}}(t, \boldsymbol{x}) = f_{\boldsymbol{\theta}(t)}(\boldsymbol{z})$

Nie, L., Ye, M., Liu, Q. and Nicolae, D., 2021. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *ICLR 2021*.

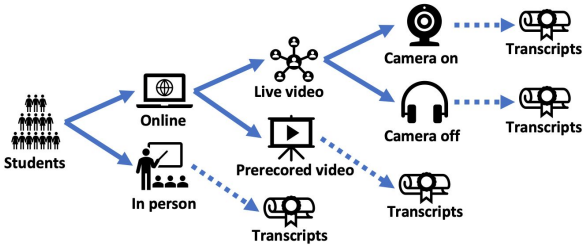
Sequential Treatments

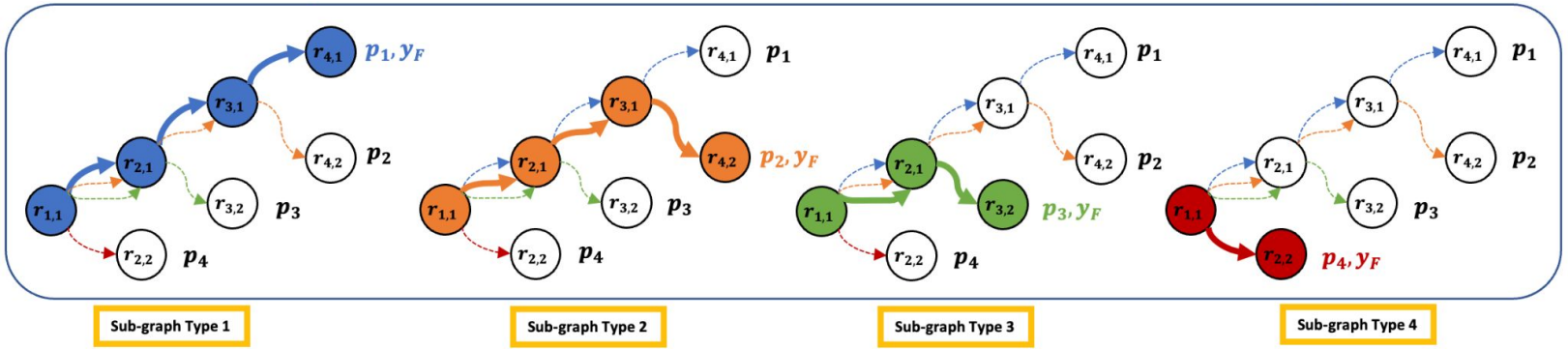


- There exists the sequential selection bias
- Selection bias will accumulate and accumulate over multiple stages
- The estimation of counterfactual outcomes is more challenging

Sequential Treatments

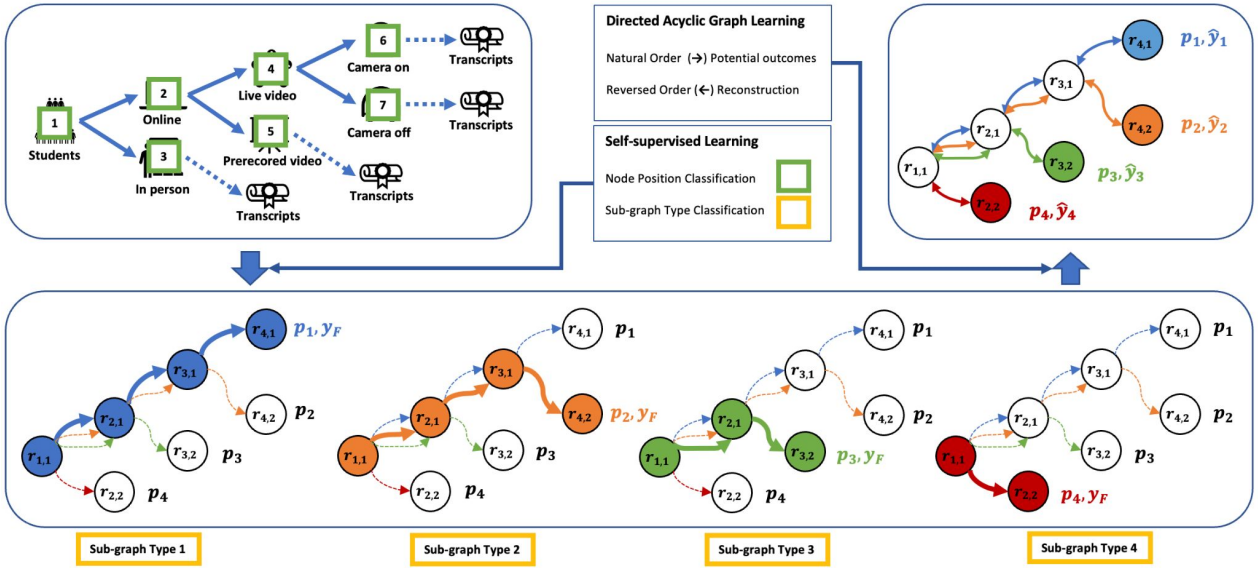
- To transform the Causal Effect Estimation Framework to the Heterogeneous Graph
  - Construct a heterogeneous graph with a large number of sub-graphs
  - Each sub-graph represents one unit and all the potential paths





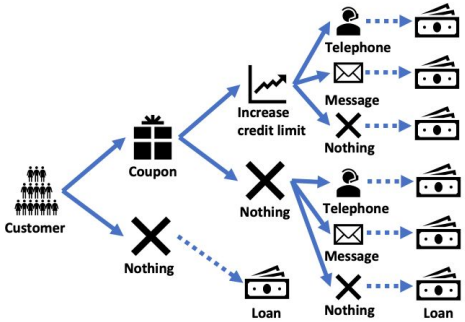
Sequential Treatments

- To preserve all information in observational data and selection bias
  - Potential Path Propagation and Completion ←Self-supervised Learning
- To infer the Potential Outcomes at the end of paths
  - Bidirectional processing based on Directed Acyclic Graph Learning
  - Inference the outcome and reconstruct the feature vectors



Sequential Treatments

- Real industrial application data



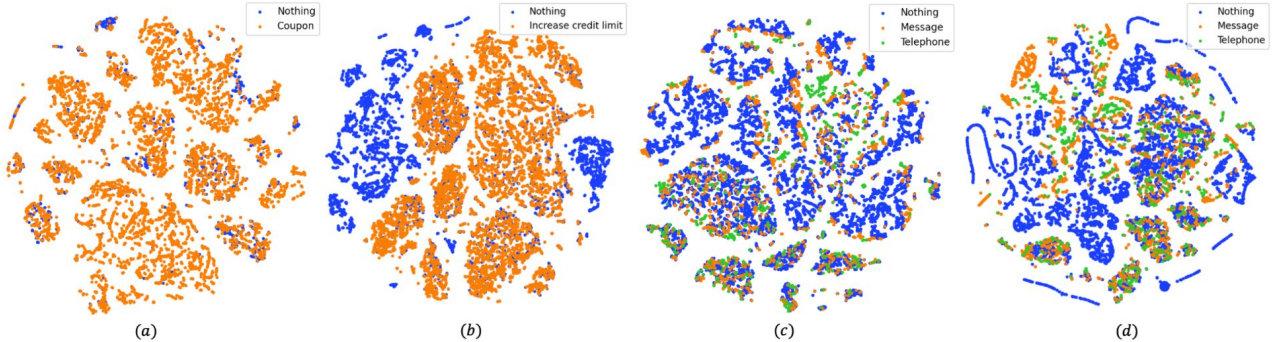


Figure 5: T-SNE visualizations for the feature distributions of sampled users along different paths. We can observe the significant bias among different intervention paths, such as in (a) get coupon or not, (b) increase credit limit or not, (c) contact via telephone, message, or not in the "increase credit limit" group, and (d) contact via telephone, message, or not in the "credit limit unchanged" group.

Structured Treatments

- Treatments are naturally structured
  - Medical prescriptions (text)
  - Protein structures (graph)
  - Computed tomography scans (image)
- Extending this idea directly to structured treatments
  - Computationally expensive
  - Not be able to make use of treatment features or learn treatment representations
- Generalized Robinson decomposition (GRD)
  - Treatments can be vectorized as a continuous embedding

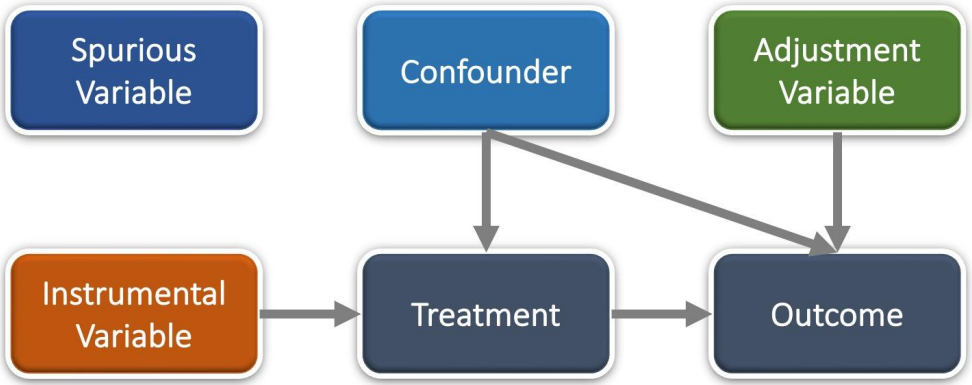
Kaddour, J., Zhu, Y., Liu, Q., Kusner, M.J. and Silva, R., 2021. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, *34*, pp.24841-24854.

Covariate

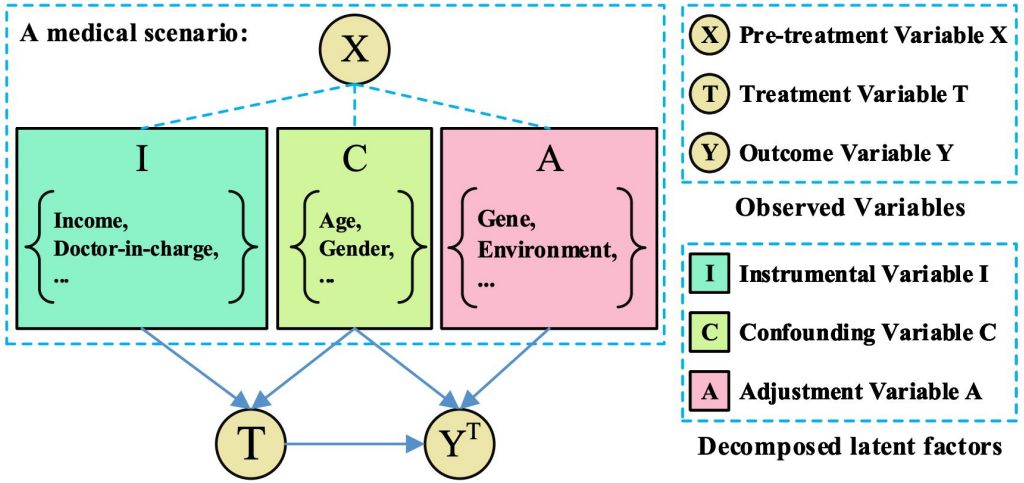
- Covariate: How could we handle the different types of covariates
  - Confounders
  - Adjustment variables
  - Instrumental variables
  - Spurious variables
- Potential solutions:
  - Feature selection
  - Feature representation disentanglement

Covariate

- Observed variable:
  - Pre-treatment variable
  - Treatment variable
  - Outcome variable



Example



Wu, A., Kuang, K., Yuan, J., Li, B., Wu, R., Zhu, Q., Zhuang, Y. and Wu, F., 2020. Learning decomposed representation for counterfactual inference. *arXiv preprint arXiv:2006.07040*.

Deep Adaptive Variable Selection Propensity Score

- Although including all the confounders is important, this does not mean that including more variables is better
- Conditioning on an instrumental variable
  - Treatment assignment ✓
  - Outcome ✗
  - Increase both bias and variance
- Conditioning on an adjustment variable
  - Outcomes ✓
  - Treatment assignment ✗
  - Unnecessary to remove bias, but can reduce variance
- Conditioning on spurious (irrelevant) variables
  - Treatment assignment ✗
  - Outcomes ✗
  - May introduce more bias into model

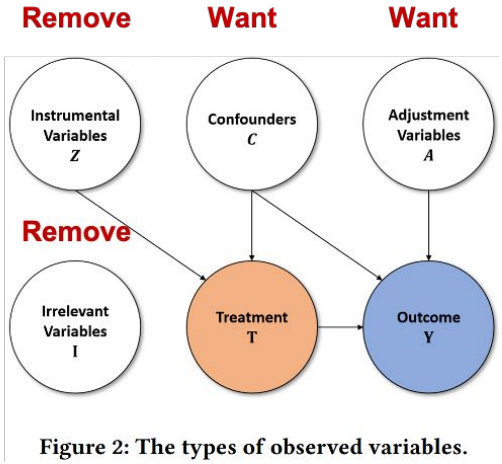
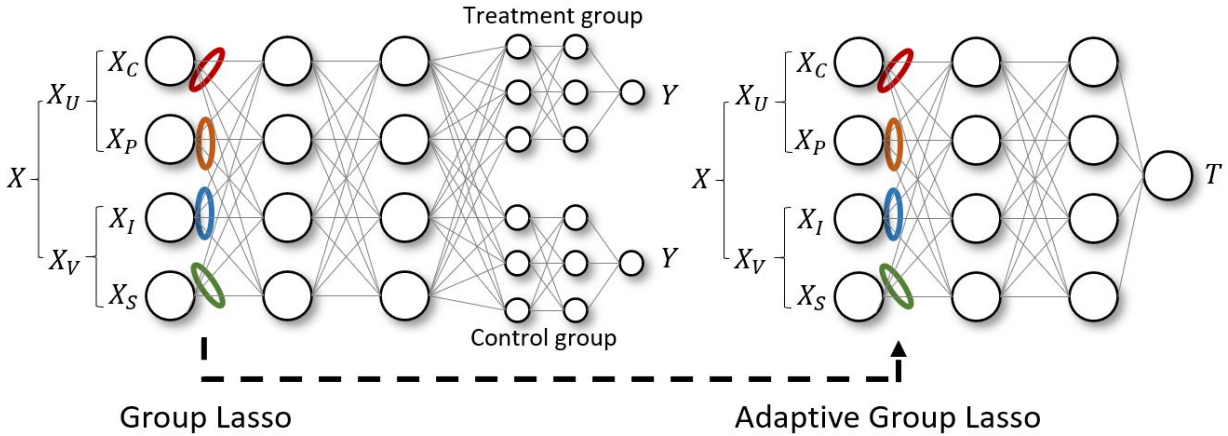


Figure 2: The types of observed variables.

- To improve the estimation of propensity score by **selecting out confounders and adjustment variables**, while **removing instrumental and spurious variables**

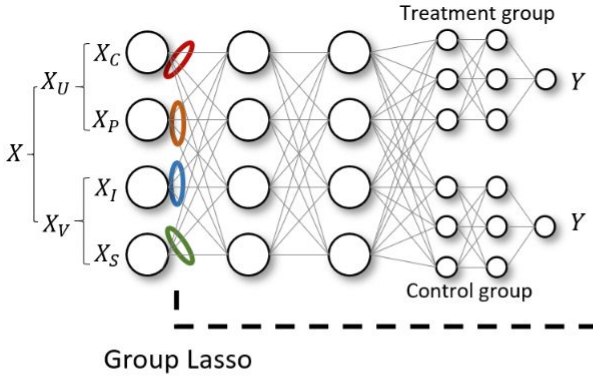
Feature Selection

- Combine the representation learning and variable selection to estimate the propensity score
- Automatically select confounders and adjustment variables and remove instrumental and spurious variables
  - Outcome prediction with group LASSO
  - Propensity score estimation with adaptive group LASSO



Zhixuan Chu, Mechelle Claridy, Jose Cordero, Sheng Li, and Stephen L Rathbun. Estimating propensity scores with deep adaptive variable selection. *In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 2023.

Outcome prediction with group LASSO



The estimator for outcome prediction with group lasso is thus defined by:

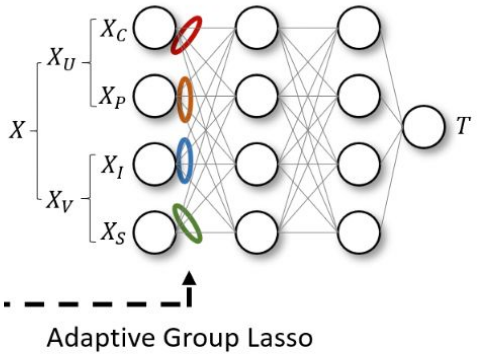
$$\widehat{\beta}_n = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\beta(x_i)) + \lambda_n q(\beta) \right\}, \tag{3.2}$$

where $\ell(y_i, f_\beta(x_i))$ denotes the log probability density (mass) function of y_i given $f_\beta(x)$. The penalty function is

$$q(\beta) = \sum_{c=1}^{n_{\mathcal{C}}} ||\beta_{c(\mathcal{C})}|| + \sum_{p=1}^{n_{\mathcal{P}}} ||\beta_{p(\mathcal{P})}|| + \sum_{i=1}^{n_{\mathcal{I}}} ||\beta_{i(\mathcal{I})}|| + \sum_{s=1}^{n_{\mathcal{S}}} ||\beta_{s(\mathcal{S})}||, \tag{3.3}$$

- Impose a group LASSO penalty -> **get the initial weight estimates for each covariate**
  - Select covariates predictive of the outcome (i.e., confounder and adjustment variables)
  - Remove covariates independent of the outcome (i.e., instrumental and spurious variables)

Propensity score with adaptive group LASSO

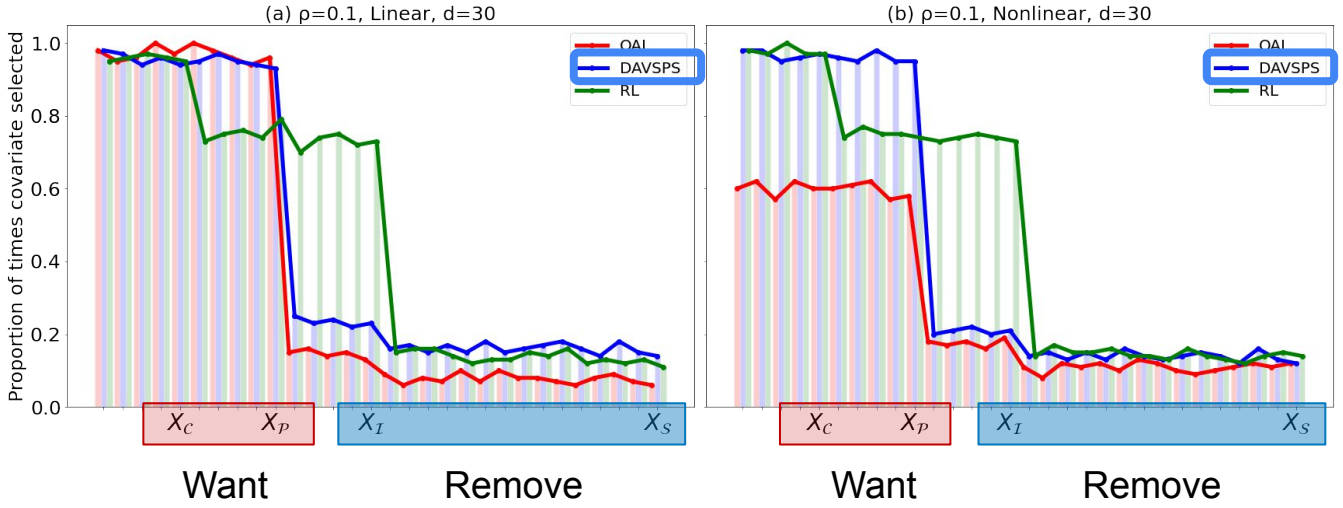


$$q(\alpha) = \underbrace{\sum_{c=1}^{n_{\mathcal{C}}} \frac{||\alpha_{c(\mathcal{C})}||}{||\widehat{\beta}_{c(\mathcal{C})}||^{\gamma}} + \sum_{p=1}^{n_{\mathcal{P}}} \frac{||\alpha_{p(\mathcal{P})}||}{||\widehat{\beta}_{p(\mathcal{P})}||^{\gamma}}}_{||\widehat{\beta}_{c(\mathcal{C})}||^{-\gamma} \text{ and} ||\widehat{\beta}_{p(\mathcal{P})}||^{-\gamma} \text{ bounded}, X_{\mathcal{C}} \text{ and } X_{\mathcal{P}} \text{ are included}} + \underbrace{\sum_{i=1}^{n_{\mathcal{I}}} \frac{||\alpha_{i(\mathcal{I})}||}{||\widehat{\beta}_{i(\mathcal{I})}||^{\gamma}} + \sum_{s=1}^{n_{\mathcal{S}}} \frac{||\alpha_{s(\mathcal{S})}||}{||\widehat{\beta}_{s(\mathcal{S})}||^{\gamma}}}_{||\widehat{\beta}_{i(\mathcal{I})}||^{-\gamma} \text{ and } ||\widehat{\beta}_{s(\mathcal{S})}||^{-\gamma} \text{ inflated to infinity}, X_{\mathcal{I}} \text{ and } X_{\mathcal{S}} \text{ are removed}}$$

- Adopt a deep neural network with adaptive group LASSO to estimate the propensity score
  - A penalty function with different regularization strengths according to different types of covariates
  - The weighted penalty is **based on initial weight estimates**

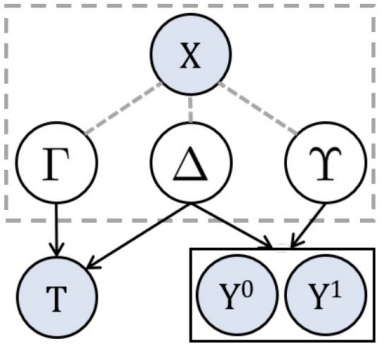
- The weights for instrumental and spurious variables **are inflated to infinity**
- While the weights for confounders and adjustment variables **are bounded**

Feature Selection



Feature Representation Disentanglement

- Decompose covariates into three latent factors
  - Instrumental factors Γ
  - Confounding factors Δ
  - Adjustment factors Υ
- Random variable X follows an unknown joint probability distribution Pr(X | Γ, Δ, Υ)
- Treatment T follows Pr(T | Γ, Δ)
- Outcome Y follows Pr (Y | Δ, Υ)
- Selection bias is induced by factors Γ and Δ



Hassanpour, N. and Greiner, R., 2020. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*.

Objective Function

- Factual Loss: $\mathcal{L}\left[\, y,\, h^{t}\left(\Delta(x), \Upsilon(x)\right)\,\right]$
- Re-Weighting Function: $\omega\left(t, \Delta(x)\right)$
- Imbalance Loss: $\texttt{disc}\left(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\right)$
- Cross Entropy Loss: $-\log\left[\,\pi_0\left(t \,|\, \Gamma(x), \Delta(x)\right)\,\right]$

$$
\begin{aligned}
J(\Gamma, \Delta, \Upsilon, h^0, h^1, \pi_0) \;=\; & \frac{1}{N}\sum_{i=1}^{N} \omega\left(t_i, \Delta(x_i)\right) \cdot \mathcal{L}\left[\, y_i,\, h^{t_i}\left(\Delta(x_i), \Upsilon(x_i)\right)\,\right] \\
& + \alpha \cdot \texttt{disc}\left(\{\Upsilon(x_i)\}_{i:t_i=0}, \{\Upsilon(x_i)\}_{i:t_i=1}\right) \\
& + \beta \cdot \frac{1}{N}\sum_{i=1}^{N} -\log\left[\,\pi_0\left(t_i \,|\, \Gamma(x_i), \Delta(x_i)\right)\,\right] \\
& + \lambda \cdot \mathfrak{Reg}(\Gamma, \Delta, \Upsilon, h^0, h^1, \pi_0)
\end{aligned}
$$

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations (ICLR)*, 2020.

Disentangled Factors

- New challenges:
  - An open problem how to learn the underlying disentangled factors precisely
  - Previous methods may fail to obtain independent disentangled factors

- Potential solutions:
  - Incorporate MI minimization learning criteria to ensure the independence of these factors

Outcome

- Outcome: When estimating the factual and counterfactual outcomes, how could we overcome the selection bias among different treatment groups?
  - Distribution invariance
  - Domain adaptation
  - Local similarity
  - Domain overlap
  - Mutual information
  - And so on
- Three concerns !

Outcome (1)

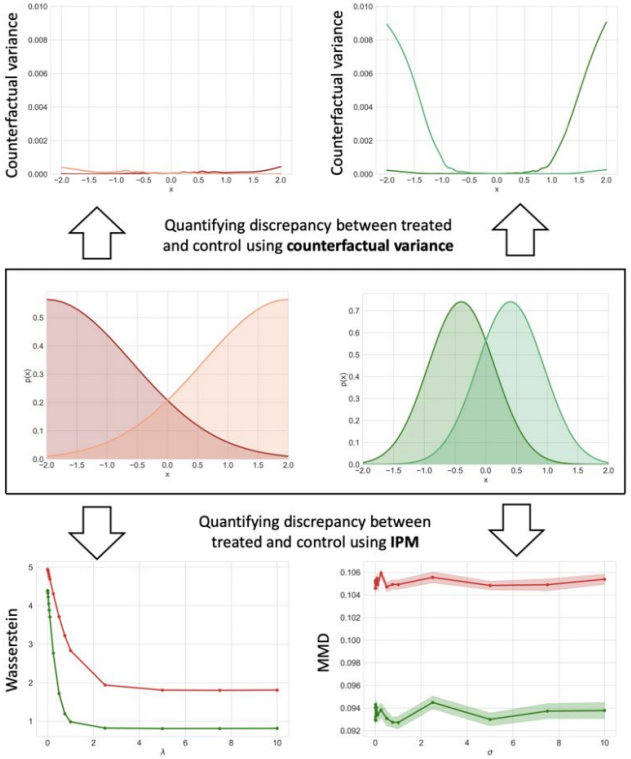
- For these domain adaptation methodologies
  - Domain-invariant representations
  - Treatment and control groups are indistinguishable in the representation space
- The impossibility theories uncover intrinsic limitations of learning invariant representations when it comes to shift in the support of domains
- Although the positivity assumption is the fundamental assumption in causal effect estimation, the positivity assumption is not guaranteed
  - High-dimensional data often contain information that is redundant or irrelevant for predicting the outcome but still helps to distinguish the treatment and control groups
  - Variables distributed differently across intervention groups are usually critical for prediction

Outcome (2)

- Optimal metric to measure the distance between the treatment and control groups remains unsettled
  - Use wasserstein or MMD to reduce distributional distance
  - Hard samples to learn representations that preserve local similarity information
  - Use counterfactual variance to measure the domain overlap
  - Utilize the mutual information between feature representations and treatment assignment

- The choice of distance metrics is highly dependent on
  - Characteristics of data distributions
  - Hyperparameters of regularization terms for imbalance mitigation

Outcome (2)



Overlap: “green” has greater counterfactual variance than “red”

No consensus among different metrics in terms of balancing data

Discrepancy: “red” has greater wasserstein and MMD than “green”

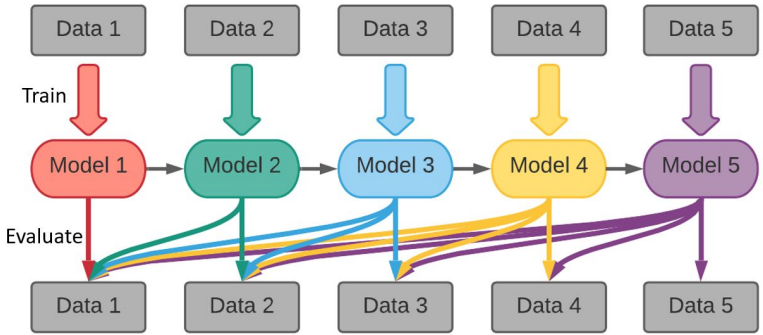
Zhang, Y., Bellot, A. and Schaar, M., 2020, June. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics* (pp. 1005-1014). PMLR.

Outcome (3)

- Regularizing representations to be domain-invariant is too strict
  - when domains (e.g., treatment and control groups) are partially overlapped
- The empirical risk minimization only on factual data outperforms domain-invariant representation learning algorithms
- Therefore, enforcing domain-invariant
  - Remove predictive information
  - Lead to a loss in predictive power
  - Regardless of which type of domain divergence metrics is employed
- This is a promising and urgent direction for the treatment effect estimation task
  - Relax the positivity assumption
  - Avoid the choice dilemma of domain divergence metrics
  - Overcome the loss of predictive information

Another Issue about Potential Outcome Estimation

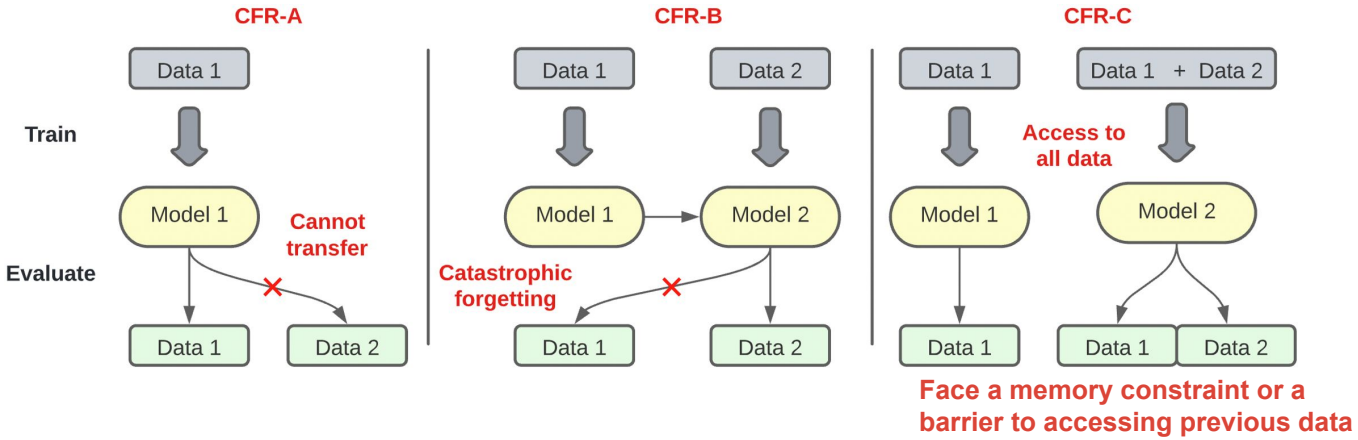
- Existing methods only focus on source-specific and stationary observational data
  - Assume that all observational data are already available and from the only one source
- This assumption is unsubstantial in practice due to two reasons:
  - Incrementally available from non-stationary data distributions
  - The realistic consideration of accessibility
- Extensibility
- Adaptability
- Accessibility



Continual learning on data 1,2,3,4, and 5

Continual Causal Inference with Incremental Observational Data

- Existing strategies to handle the new challenges :
  - Directly apply the previously trained model to new observational data (CFR-A);
  - Utilize newly available data to fine-tune the previously learned model (CFR-B);
  - Store all previous data and combine with new data to re-train the model from scratch (CFR-C);



Continual Causal Inference with Incremental Observational Data

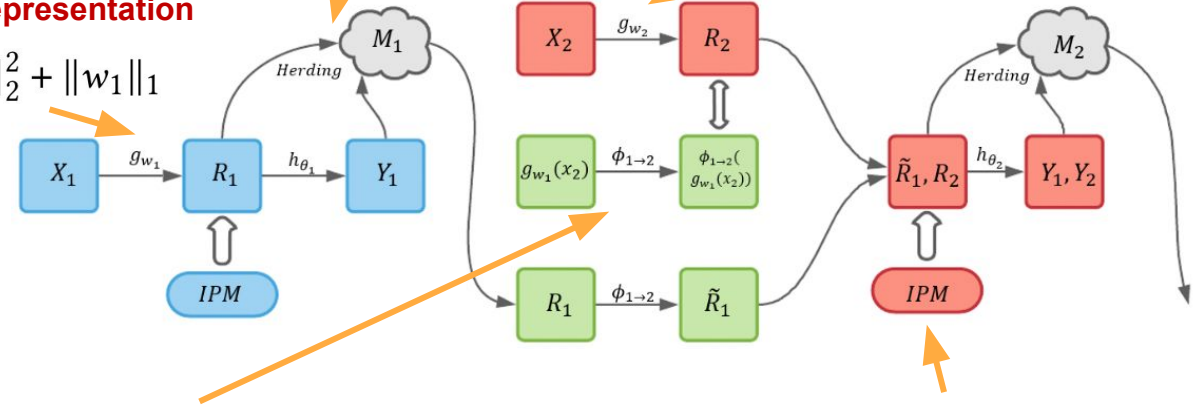
Learning selective representation

$$L_{w_1} = \|w_1\|_2^2 + \|w_1\|_1$$

Herding sampling algorithm

Feature Representation Distillation

$$L_{FD}(x) = 1 - cos\left(g_{w_1}(x), g_{w_2}(x)\right)$$



Feature Representation Transformation

$$L_{FT}(x) = 1 - cos\left(\phi_{1\to2}(g_{w_1}(x)), g_{w_2}(x)\right)$$

Balancing Global Feature Representation Space

$$\text{Wass}(P, Q) = \inf_{k \in \mathcal{K}} \int_{g(x)} \|k(g(x)) - g(x)\| P(g(x)) d(g(x))$$

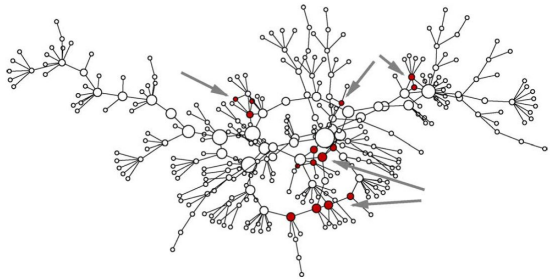
Agenda

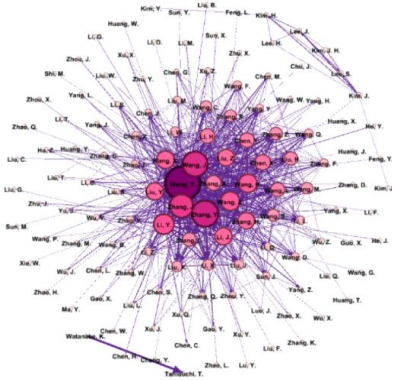
1. Background on Causal Inference
2. Representation Learning based Methods
3. **Graph Neural Networks based Methods**
4. Causality-aided Machine Learning
5. Applications and Future Directions
6. Conclusions

Graph Data

- Graphs have been extensively used for modeling many real-world systems with connected units

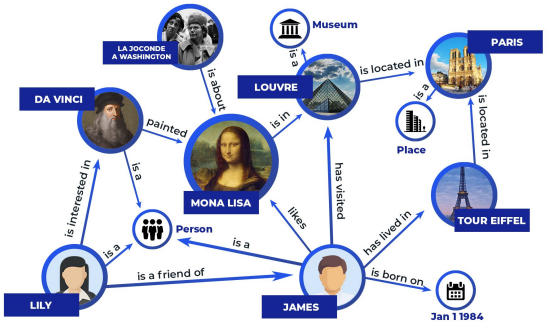
Social Network





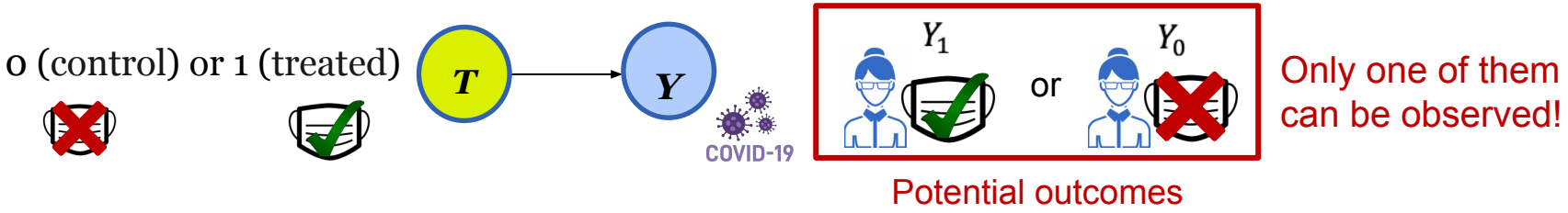
Cooperation Network

Knowledge Graph



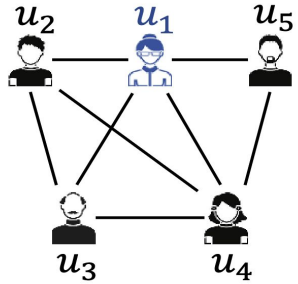
Causal Inference on Graphs

- Causal inference studies the *causal relations* rather than *statistical dependencies* between variables
- **Causal effect estimation:** assessing the causal effects of a treatment (e.g., wearing a mask) on an outcome (e.g., COVID-19 infection)



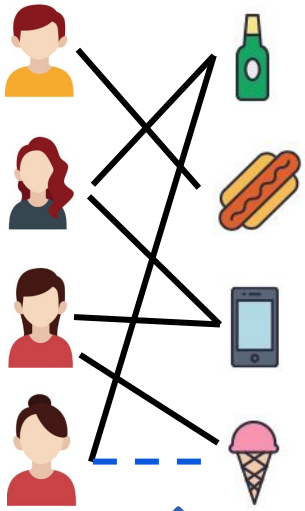
- Causal effect estimation on graphs

  Question: In a contact network, how does the face mask practice influence COVID-19 infection risk?

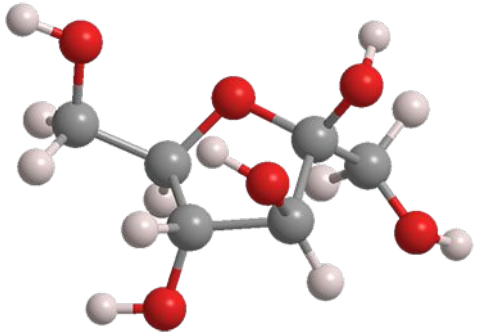


Applications of Causal Inference on Graphs

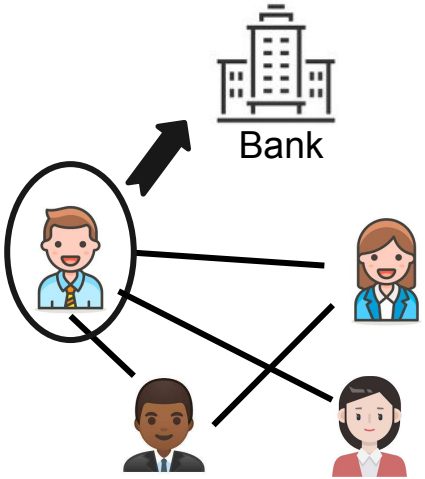
- Causal inference has a wide range of applications in graph data



How does an ad campaign motivate users' purchase?



How does a substructure influence the molecular property?

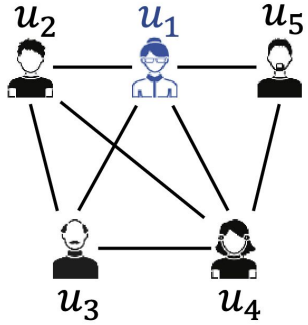


Which factors most impact the applicant's credit application result?

Causal Effect Estimation on Graphs

Problem definition:

- **Given**: observational data {***X, A, T, Y***}
  - node features ***X***
  - graph structure ***A***
  - treatment ***T***
  - observed outcomes ***Y***

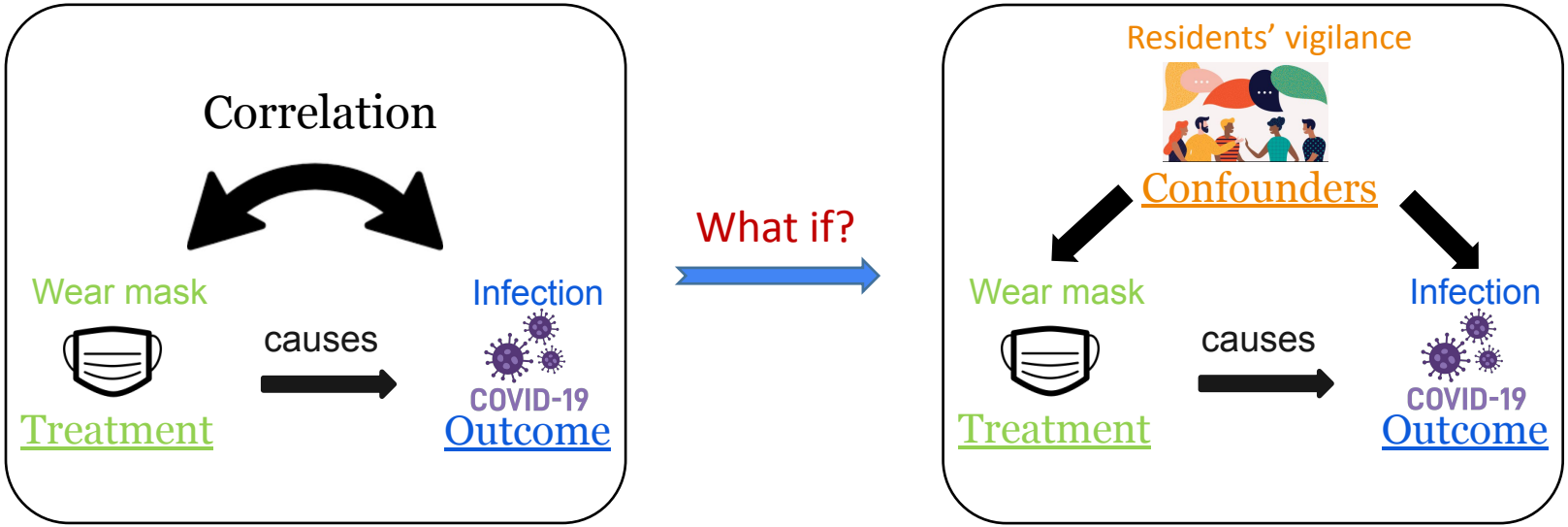


- **Goal**: Given a graph, a treatment assignment and the outcome, we aim to estimate the individual treatment effect (ITE) for each node i:

$$\text{ITE} = Y_{1,i} - Y_{0,i}$$

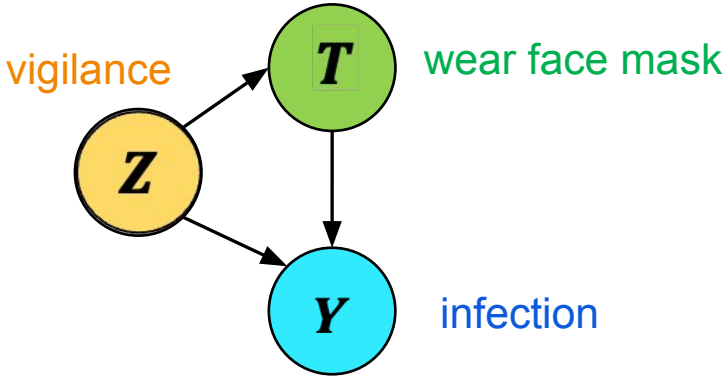
Hidden Confounders on Graphs

- The confounders are often unobserved
- Even cannot be fully reflected through unit features/covariates

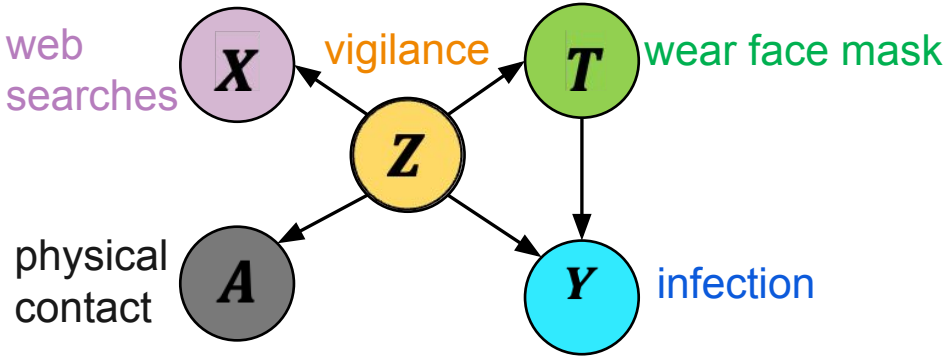


Graph as a Proxy for Confounders

- Hidden confounders ***Z*** causally affect treatment ***T*** and outcome ***Y***



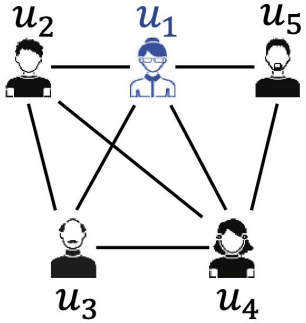
- Graph data (node features ***X*** and network structure ***A***) can be used as proxy variables for hidden confounders ***Z***



Motivation of Introducing Graph: similar nodes are connected more often than dissimilar nodes (**homophily**)

Key Idea of Leveraging Graph

- **Motivation**: Hidden confounders often lead to biased causal effect estimation
- **Key idea**: Capture hidden confounders through representation learning from graph data



Graph data can be proxy variables for hidden confounders

Confounder representations

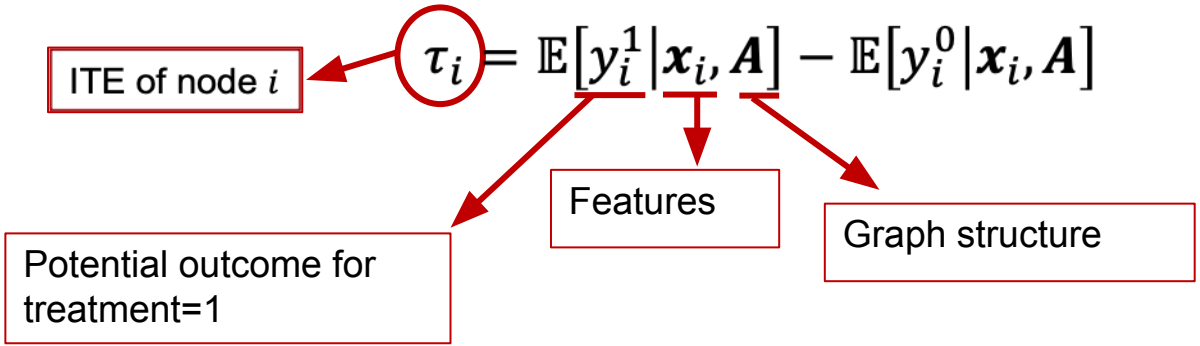
Effective deep learning techniques are utilized

ITE = $Y_{1,i}^t - Y_{0,i}^t$

Estimate ITE based on confounder representations

Problem Definition

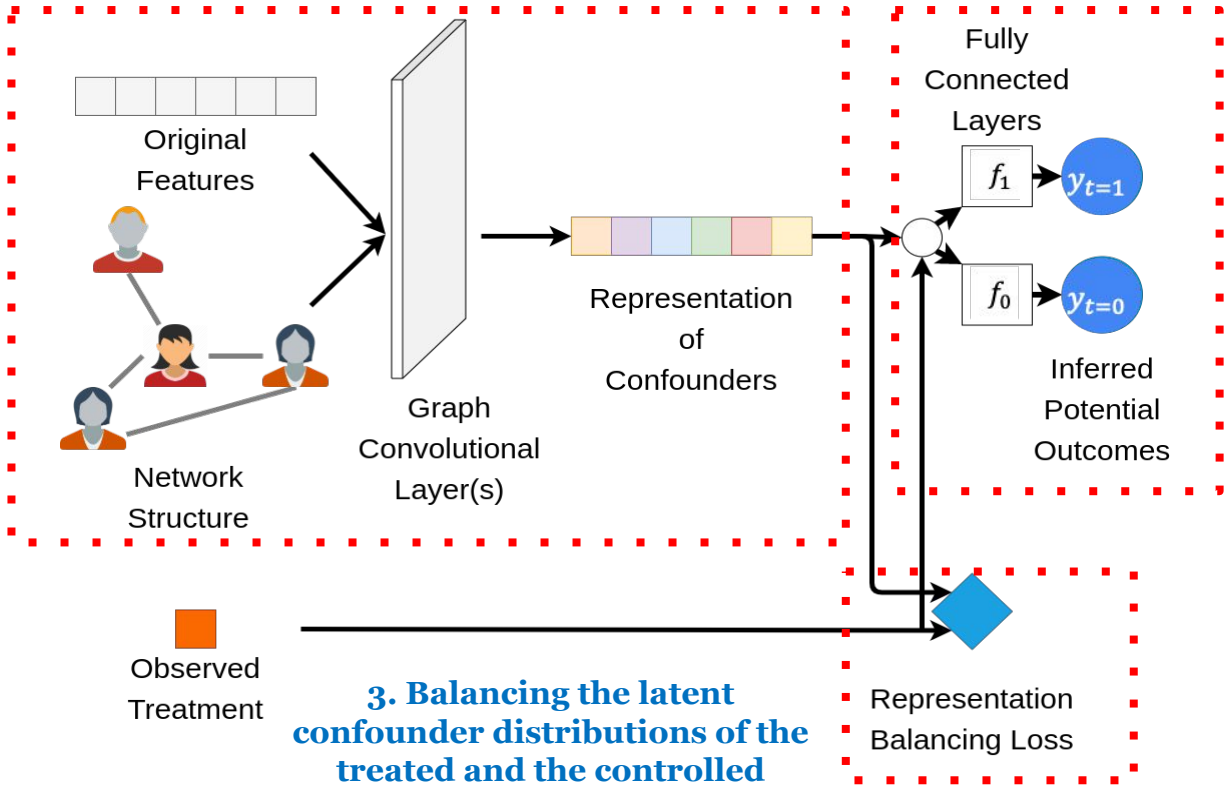
- Given the observational data $\{\boldsymbol{x}_i, t_i, y_{i,t_i}\}_{i=1}^{n}$ and graph adjacency matrix $\boldsymbol{A}$ among n instances, the goal is to estimate the ITE for each individual node on the graph



Network Deconfounder

1. Learning latent confounders with GNN

2. Outcome inference with fully connected layers

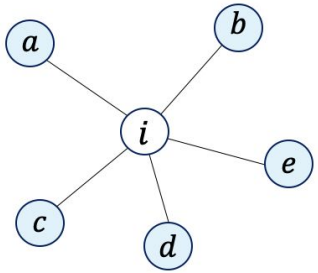


3. Balancing the latent confounder distributions of the treated and the controlled

[1] Guo, Ruocheng, et al. "Learning individual causal effects from networked observational data." *WSDM*. 2020.

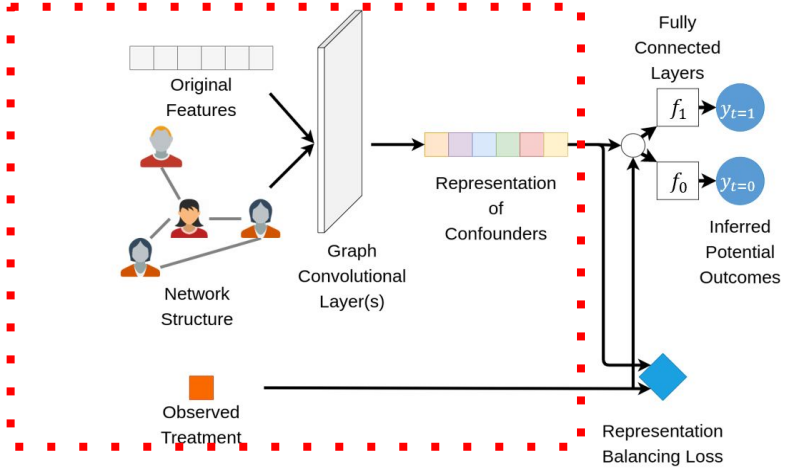
Network Deconfounder

- Learn latent confounders with observed features and graph information through graph neural networks (GNNs)



$$\boldsymbol{h}_i = g(\boldsymbol{x}_i, \boldsymbol{A}) = \sigma\left(\boldsymbol{W} \sum_{u \in N(i) \cup i} \frac{\boldsymbol{x}_j}{\sqrt{|N(j)||N(i)|}}\right)$$

$$= \sigma\left(\left(\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\right)_i \boldsymbol{W}\right) = \sigma((\widehat{\boldsymbol{A}}\boldsymbol{X})_i \boldsymbol{W})$$



$\boldsymbol{h}_i$: latent confounder representations of instance i

$\boldsymbol{W}$: weight matrix of GCN layer (parameters to be learned)

$\widehat{\boldsymbol{A}}$: normalized adjacency matrix with normalization trick, $\widehat{\boldsymbol{A}} = \widetilde{\boldsymbol{D}}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{D}}^{-\frac{1}{2}}$

$\widetilde{\boldsymbol{A}}$: adjacency matrix of graph with self-loop

[1] Guo, Ruocheng, et al. "Learning individual causal effects from networked observational data." *WSDM*. 2020.

Network Deconfounder

- Build a counterfactual outcome inference model with the supervision of observed factual outcomes and the representations of confounders

Minimize the Mean Squared Error on factual outcomes

$$\min \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_{i,t_i} - y_{i,t_i}\right)^2$$

Inferred potential outcome

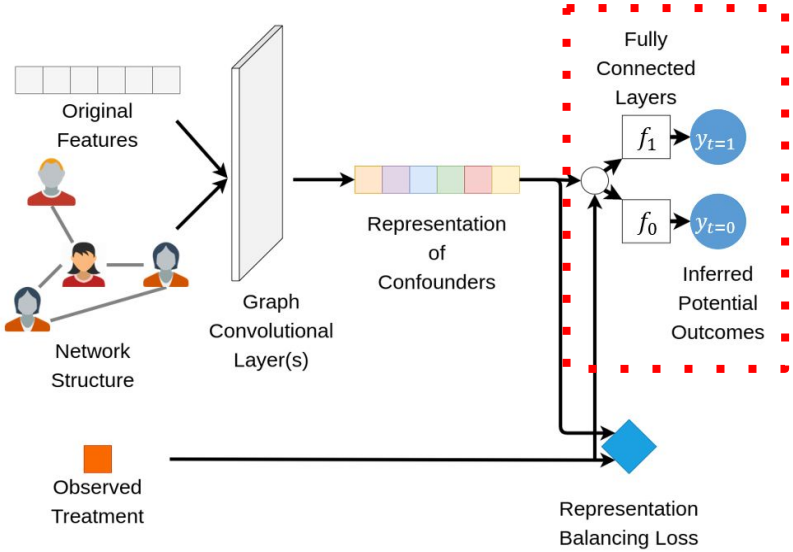
$$\hat{y}_{i,t_i} = f(g(\boldsymbol{x}_i, \boldsymbol{A}), t_i)$$

Outcome inference function

$$f(\hat{\boldsymbol{z}}_i, t_i) = \begin{cases} f_1(\hat{\boldsymbol{z}}_i) & \text{if } t_i = 1 \\ f_0(\hat{\boldsymbol{z}}_i) & \text{if } t_i = 0 \end{cases}$$

Fully connected layers for regression

$$f_1 = \boldsymbol{w}^1\sigma(\boldsymbol{W}_L^1 \ldots \sigma(\boldsymbol{W}_1^1\hat{\boldsymbol{z}}_i))$$
$$f_0 = \boldsymbol{w}^0\sigma(\boldsymbol{W}_L^0 \ldots \sigma(\boldsymbol{W}_1^0\hat{\boldsymbol{z}}_i))$$



[1] Guo, Ruocheng, et al. "Learning individual causal effects from networked observational data." *WSDM*. 2020.

Network Deconfounder

- Minimizing the error in the factual outcomes does not necessarily mean that the error in the counterfactual outcomes is also minimized
- Confront the challenge of distribution shift

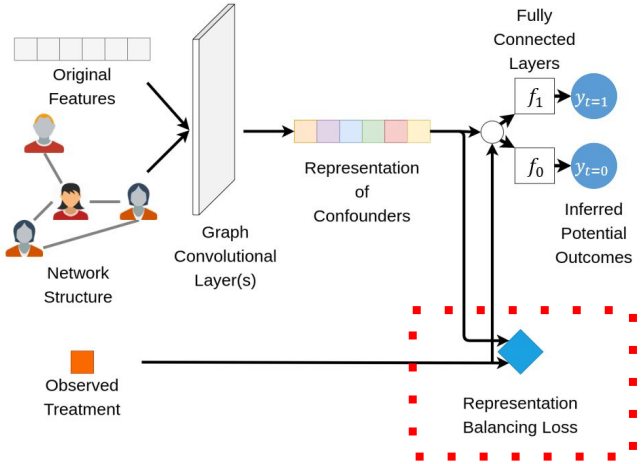
Solution: representation balancing for distribution shift

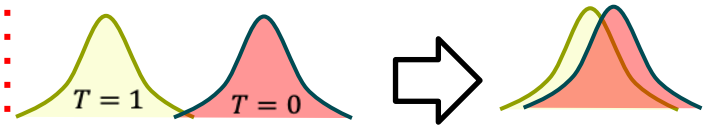
$$\min \rho_{\mathcal{Z}}(P, Q) = \inf_{k \in \mathcal{K}} \int_{\boldsymbol{h} \in \{\boldsymbol{h}_i\}_{i:t_i=1}} \|k(\boldsymbol{h}) - \boldsymbol{h}\| P(\boldsymbol{h}) d\boldsymbol{h}$$

integral probability metric (IPM) Wasserstein-1 distance

$\mathcal{K} = \{k | k: \mathbf{R}^d \to \mathbf{R}^d, s.t. Q(k(\boldsymbol{h})) = P(\boldsymbol{h})\}$

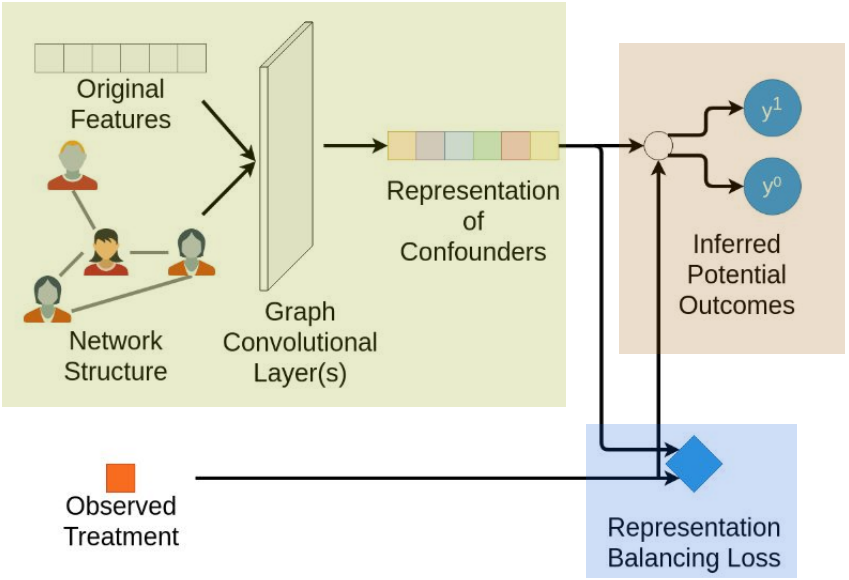
$\mathcal{K}$: the set of push-forward 1-Lipschitz functions that can transform the representation distribution of the treated $P(\boldsymbol{h})$ to that of the controlled $Q(\boldsymbol{h})$





[1] Guo, Ruocheng, et al. "Learning individual causal effects from networked observational data." *WSDM*. 2020.

Network Deconfounder



- Confounder representation learning

$$h_i = g(\boldsymbol{x}_i, \boldsymbol{A}) = \sigma((\widehat{\boldsymbol{A}}\boldsymbol{X})_i \boldsymbol{U})$$

GCN layers

- Representation balancing
  - Help reduce the biases in ITE estimation

distribution of confounder representations

Wasserstein-1 distance

$$\rho_{\mathcal{Z}}(P, Q) = \inf_{k \in \mathcal{K}} \int_{\mathbf{h} \in \{\mathbf{h}_i\}_{i:t_i=1}} ||k(\mathbf{h}) - \mathbf{h}|| P(\mathbf{h}) d\mathbf{h}$$

- Outcome prediction

$$f(\mathbf{h}_i, t) = \begin{cases} f_1(\mathbf{h}_i) \text{ if } t = 1 \\ f_0(\mathbf{h}_i) \text{ if } t = 0 \end{cases}$$

Loss function

$$\mathcal{L}(\{\mathbf{x}_i, t_i, y_i\}_{i=1}^n, \mathbf{A}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i^{t_i} - y_i)^2 + \alpha \rho_{\mathcal{Z}}(P, Q) + \lambda ||\boldsymbol{\theta}||_2^2,$$

balancing

regularization

Outcome prediction loss

[1] Guo, Ruocheng, et al. "Learning individual causal effects from networked observational data." *WSDM*. 2020.

Experiments

- Real-world social network BlogCatalog (node=user, edge=friendship)
- Simulated causal problem
  - Treatment: A user has more viewers from mobile devices (T=1) or desktops (T=0)
  - Outcome: the reviews a user receives
  - Confounder: user's post topics

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i\in[n]}(\tau_i - \hat{\tau}_i)^2}$$

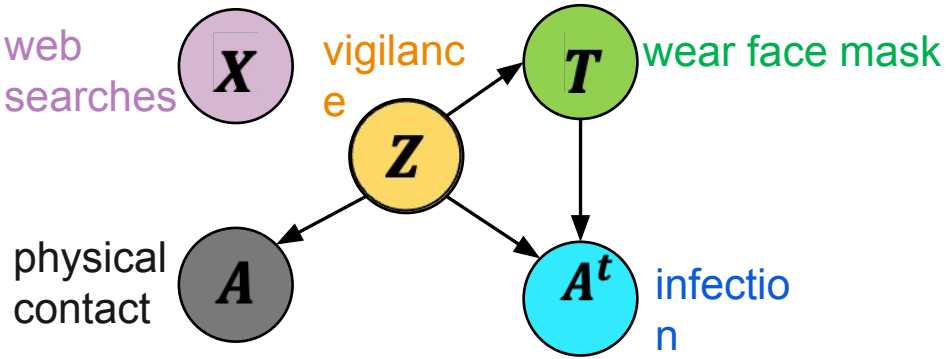
$$\epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}(\hat{\tau}_i) - \frac{1}{n}\sum_{i=1}(\tau_i)|$$

The smaller, the better

Controls the influence of neighbors on each node's confounders

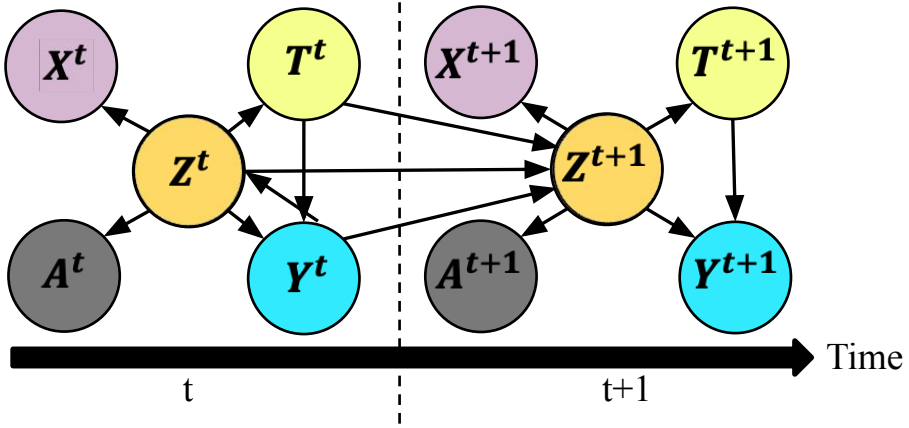
BlogCatalog						
κ_2	0.5		1		2	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
NetDeconf (ours)	**4.532**	**0.979**	**4.597**	**0.984**	**9.532**	**2.130**
CFR-Wass	10.904	4.257	11.644	5.107	34.848	13.053
CFR-MMD	11.536	4.127	12.332	5.345	34.654	13.785
TARNet	11.570	4.228	13.561	8.170	34.420	13.122
CEVAE	7.481	1.279	10.387	1.998	24.215	5.566
Causal Forest	7.456	1.261	7.805	1.763	19.271	4.050
BART	4.808	2.680	5.770	2.278	11.608	6.418

Hidden Confounders in Dynamic Graphs



- Graph data (node features X and network structure A) can be used as proxy variables for hidden confounders Z

Static environment



- Historical data (previous confounders Z^t, treatment T^t, outcome Y^t) can influence current confounders Z^{t+1};

Dynamic environment

Problem Definition

- Given the time-evolving networked observational data $\{\boldsymbol{X}^t, \boldsymbol{A}^t, \boldsymbol{T}^t, \boldsymbol{Y}^t\}_1^P$ across P time stamps, the goal is to learn the ITE τ_i^t for each instance i at each time stamp t

$$\tau_i^t = \mathbb{E}\left[Y_{1,i}^t \middle| \boldsymbol{X}_i^t, \boldsymbol{A}^t, \mathcal{H}^t\right] - \mathbb{E}\left[Y_{0,i}^t \middle| \boldsymbol{X}_i^t, \boldsymbol{A}^t, \mathcal{H}^t\right]$$

ITE of node i at time stamp t

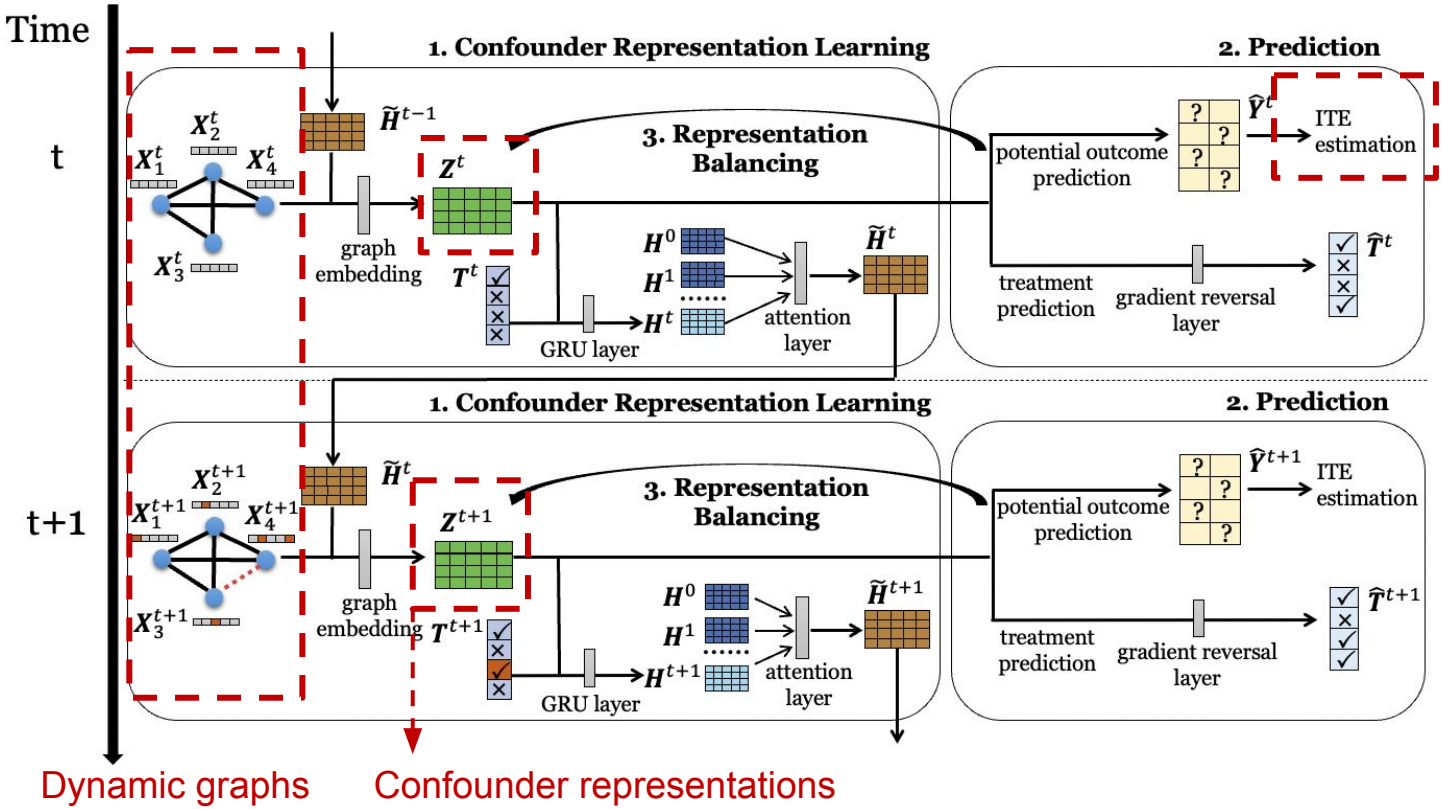
Potential outcome for treatment=1

Features

Graph structure

Historical data

Deconfounding in Dynamic Network (DNDC)



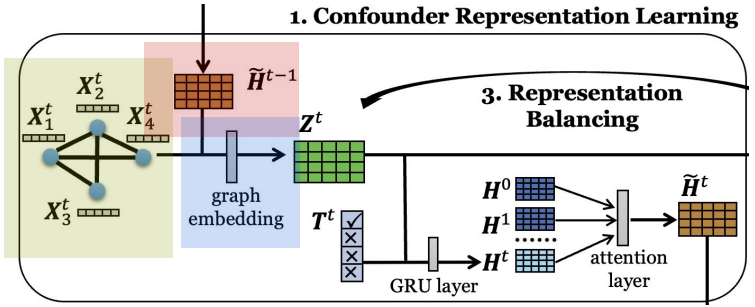
[1] Jing Ma, et al. Deconfounding with Networked Observational Data in a Dynamic Environment. WSDM 2021

Deconfounding in Dynamic Network (DNDC)

- Confounder representation learning

Graph neural network | History embedding | Graph structure

$$\boldsymbol{Z}_i^t = g(([\boldsymbol{X}^t, \tilde{\boldsymbol{H}}^{t-1}])_i, \boldsymbol{A}^t)$$

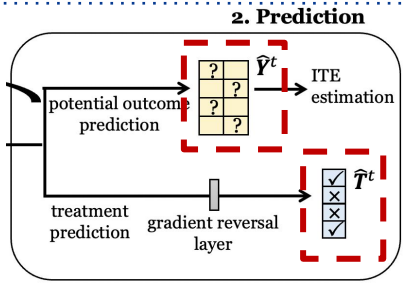


- Prediction for potential outcomes and treatment

$$\hat{Y}_{1,i}^t = f_1(\boldsymbol{Z}_i^t)$$

$$\hat{Y}_{0,i}^t = f_0(\boldsymbol{Z}_i^t)$$

ITE =



- Overall loss

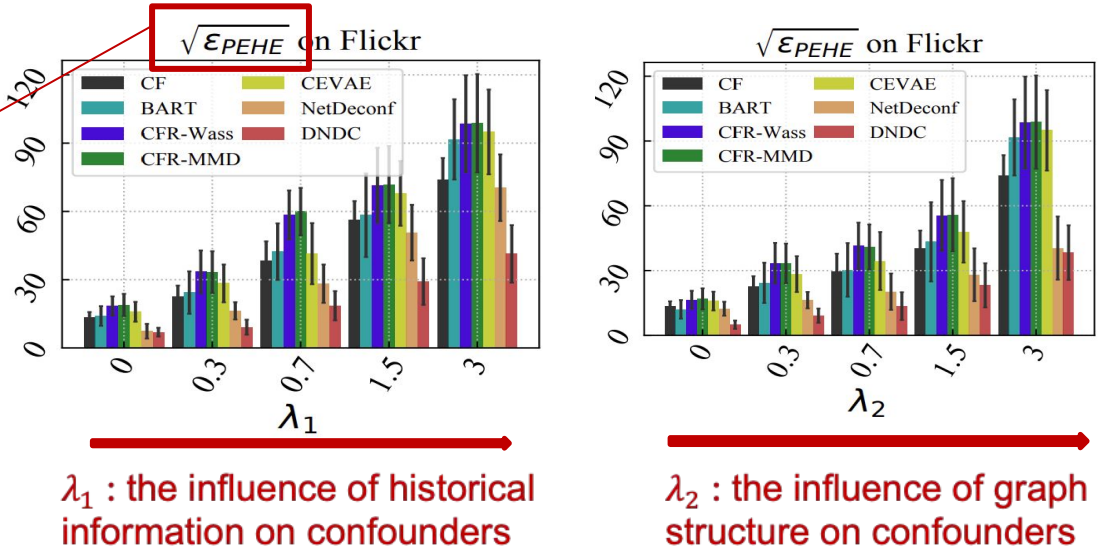
$$\mathcal{L} = \begin{cases} \mathcal{L}_Y & \text{outcome prediction loss} \\ \mathcal{L}_T & \text{treatment prediction loss} \\ \mathcal{L}_B & \text{balancing loss} \end{cases}$$

[1] Jing Ma, et al. Deconfounding with Networked Observational Data in a Dynamic Environment. WSDM 2021

Experiments

- Observations:
  - When $\lambda_1 \uparrow$, DNDC is better as it leverages historical information
  - When $\lambda_2 \uparrow$, our method is better as it leverages graph information

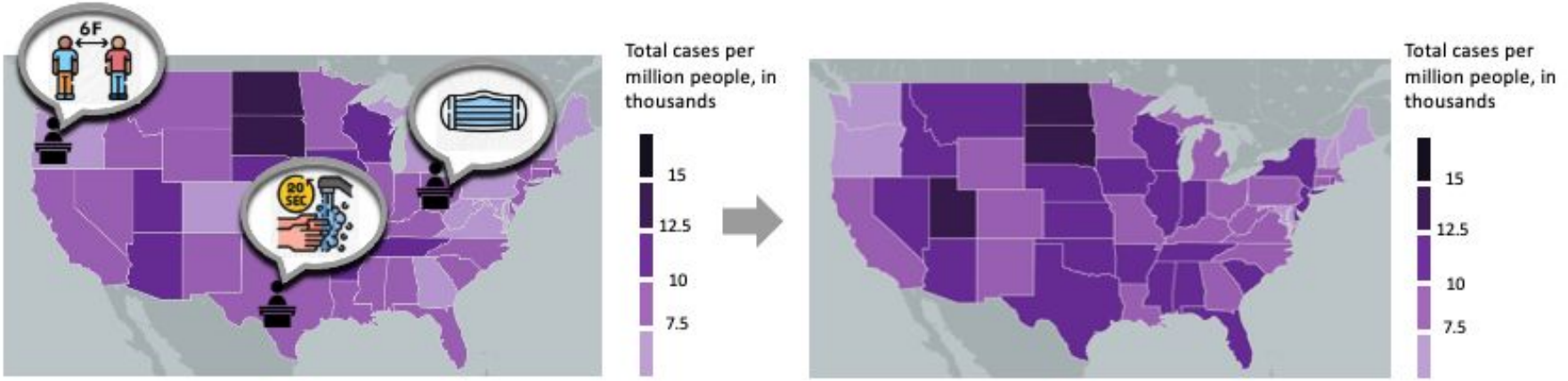
The smaller, the better



λ_1 : the influence of historical information on confounders

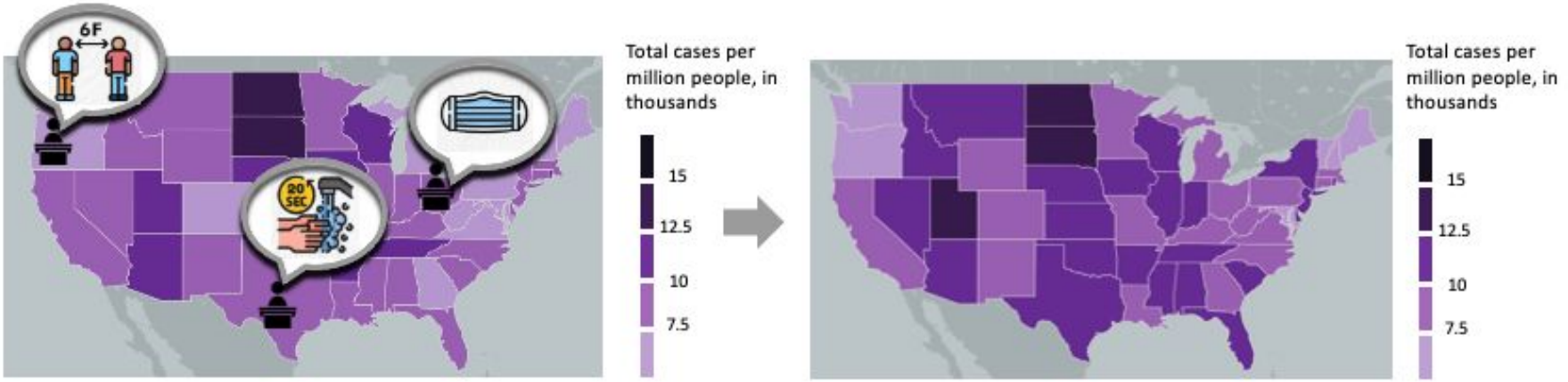
λ_2 : the influence of graph structure on confounders

Evaluation: A Case Study on COVID-19 Policies

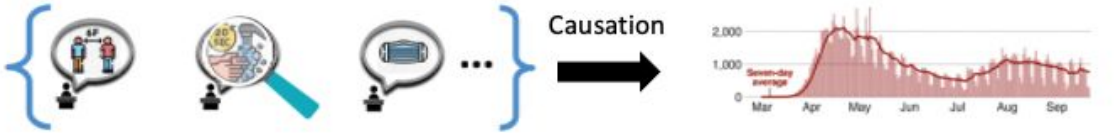


- The outbreak of COVID-19 has been affecting public health since 2019
- Various non-pharmaceutical **public policies** have been announced to limit impact of COVID-19 across the US (e.g., social distancing, mask requirement, travel restrictions)

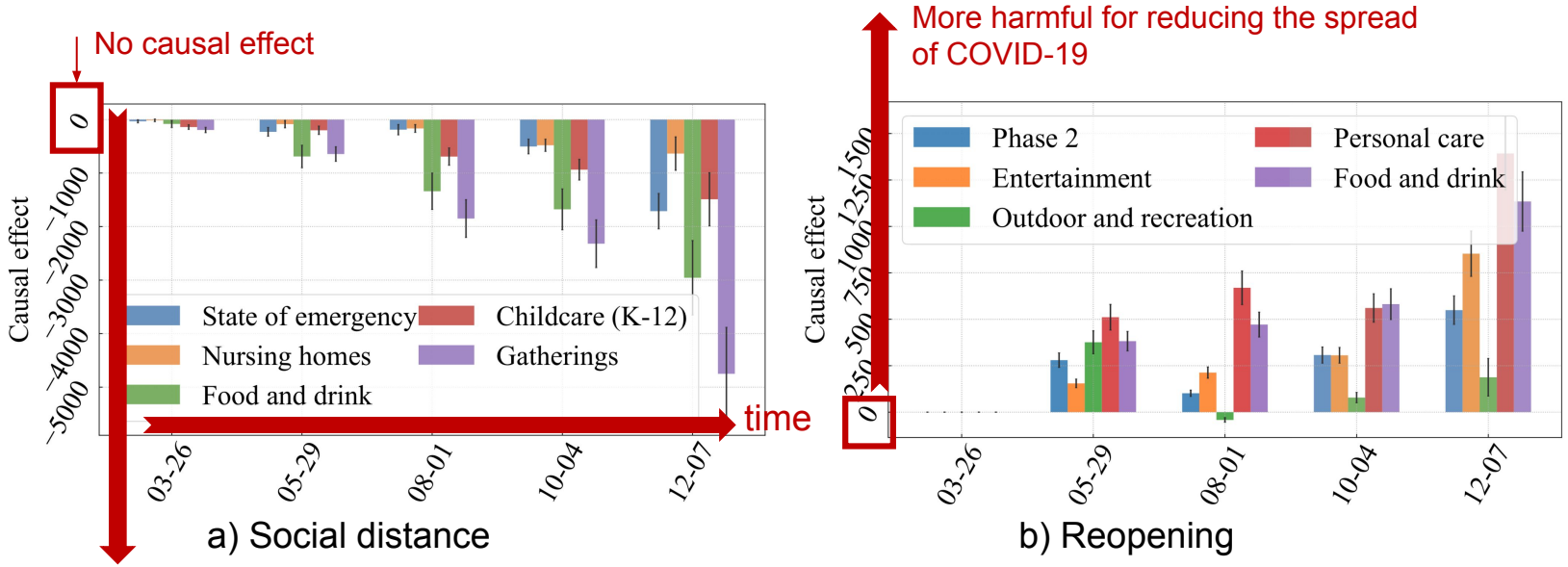
Evaluation: A Case Study on COVID-19 Policies



- To help future policy makers, a natural question is: *which policy is more effective to control the impact of COVID-19 in a given context?*
- Specifically, we study the causal effect of different public policies (treatment) on the outbreak dynamics (outcome)



Assess Causal Effects of COVID-19 Policies



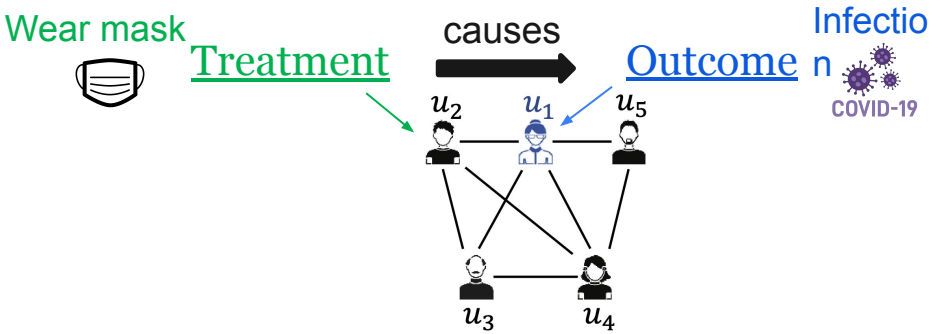
Observations:

- The causal effect estimation is consistent with epidemiological literature
- The estimation results can bring more insights for future decision making

[1] Jing Ma, et al. Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US. WWW 2022

Graph Interference

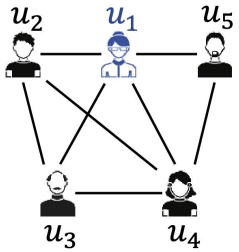
- **Interference**: the treatment of an individual may causally affect the outcome of other individuals
  - **Example**: whether a person wears a face mask in public may influence the infection risk of other people
- Traditional causal inference assumes interference does not exist, but interference is ubiquitous in networked data

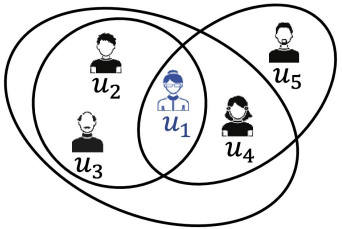


Hypergraph and High-order Interference

- In hypergraphs, each hyperedge can connect an arbitrary number of nodes, in contrast to an ordinary edge which connects exactly two nodes

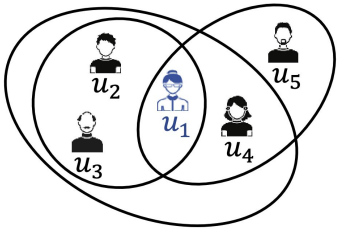
Ordinary graph

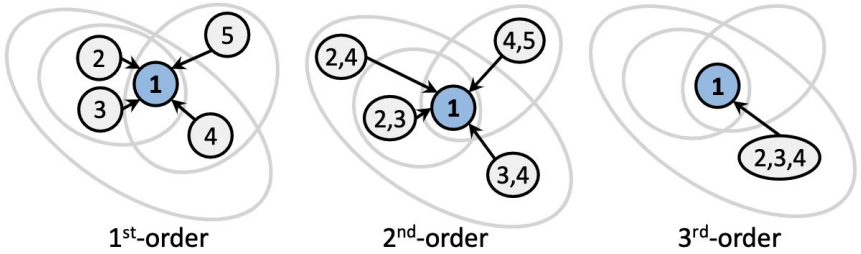




Hypergraph

- Example of high-order interference: The interaction between u_2 and u_3 may also influence the exposure of the virus to u_1; i.e., $u_2 \times u_3 \rightarrow u_1$





1st-order 2nd-order 3rd-order

[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Causal Inference under Interference

- **Given**: observational data $\{\boldsymbol{X}, \boldsymbol{H}, \boldsymbol{T}, \boldsymbol{Y}\}$, denoting features, hypergraph, treatment, and observed outcomes
- **Goal**: estimate the ITE for each individual (node) *i*:

$$\tau(\mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H}) = \mathbb{E}[Y_i^1 - Y_i^0 | X_i = \mathbf{x}_i, T_{-i} = \mathbf{T}_{-i}, X_{-i} = \mathbf{X}_{-i}, H = \mathbf{H}]$$
$$= \mathbb{E}[\Phi_Y(1, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H}) - \Phi_Y(0, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H})]$$

Potential outcome when T_i=1 or T_i=0

Treatment of other nodes

Assumptions

- **Assumption** 1. For any node i, given the node features X_i, the potential outcomes are independent with the treatment assignment and summary of neighbors

Assumption 1: (Unconfoundedness) no unobserved confounders exist

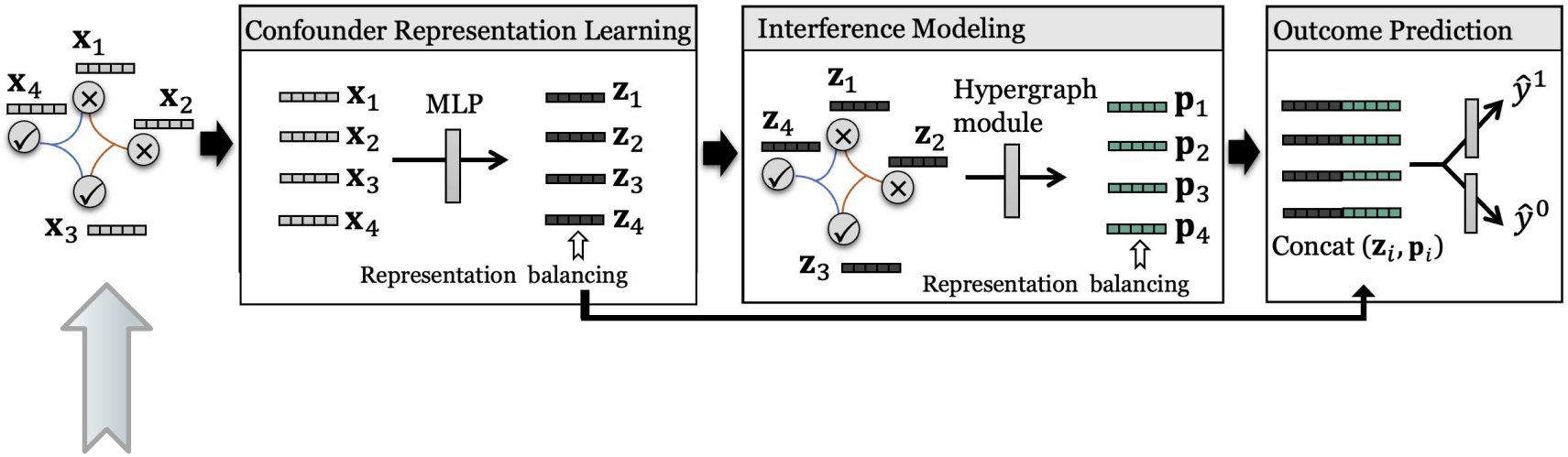
- **Assumption** 2. For any node i, any values of H, X_{-i}, and T_{-i}, if the output of a summary function $o_i = \mathrm{SMR}(H, T_{-i}, X_{-i})$ is determined, then the values of the potential outcomes with feature X_i are also determined

a function which characterizes all the "environmental" information related to node *i*.

Assumption 2: (Expressiveness of summary function)

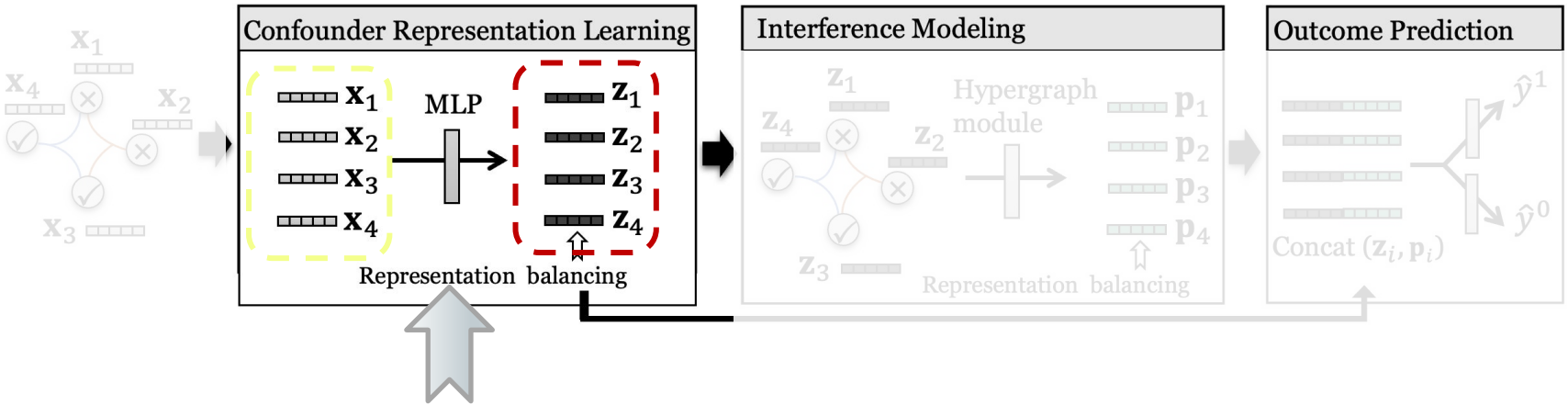
Theory (Identifiability): the defined ITE can be identifiable from observational data under the assumptions

Learning Causal Effects on Hypergraph



[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Learning Causal Effects on Hypergraph



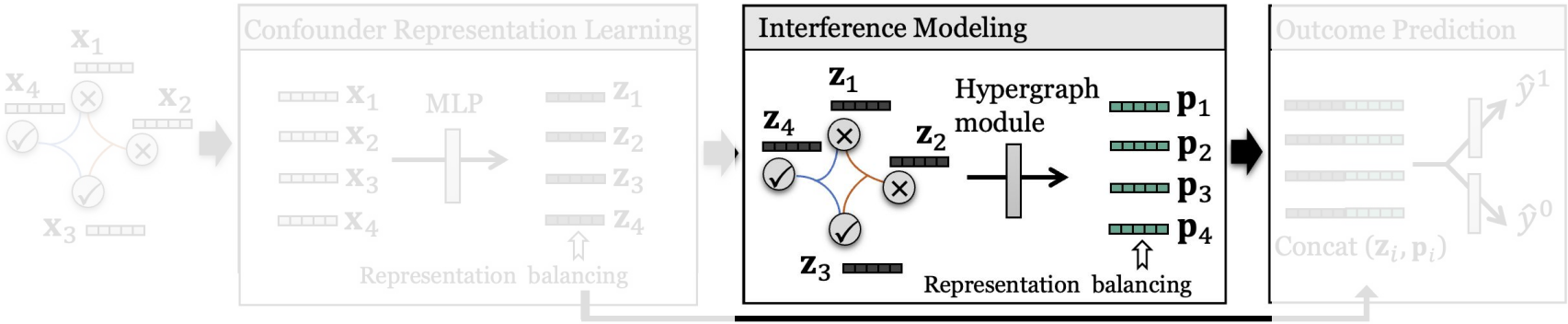
Confounder representation learning: encode the node features into a latent space to capture the confounders

$$\mathbf{z}_i = \mathrm{MLP}(\mathbf{x}_i)$$

Assumption 2 (Expressiveness of summary function)

[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Learning Causal Effects on Hypergraph



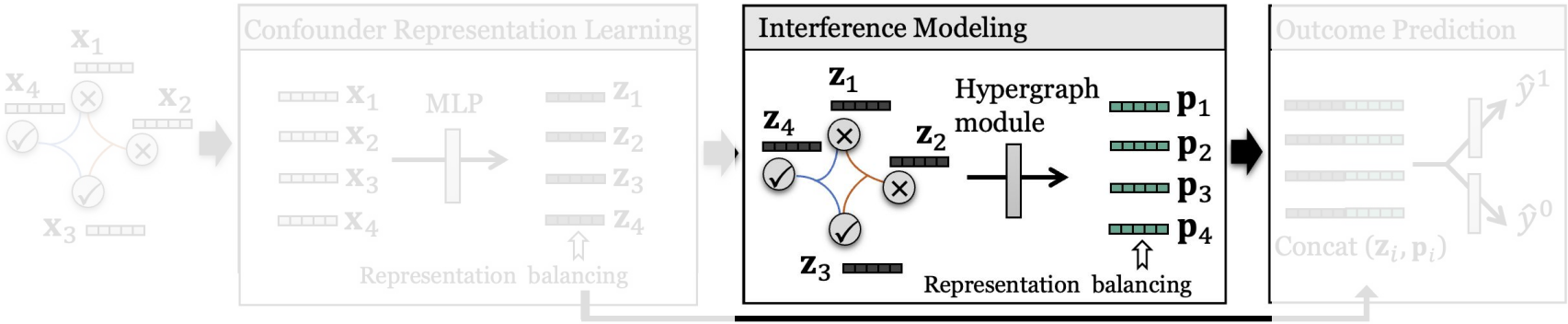
Interference Modeling: capture the high-order interference for each individual through representation learning

- Propagate the neighboring treatment assignment and confounder representations with a hypergraph module

Assumption 2 (Expressiveness of summary function)

[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Learning Causal Effects on Hypergraph



Interference Modeling: capture the high-order interference for each individual through representation learning

- Hypergraph convolutional network is applied in the hypergraph module:

$$\mathbf{P}^{(l+1)} = \text{LeakyReLU}\left(\mathbf{L}\mathbf{P}^{(l)}\mathbf{W}^{(l+1)}\right)$$

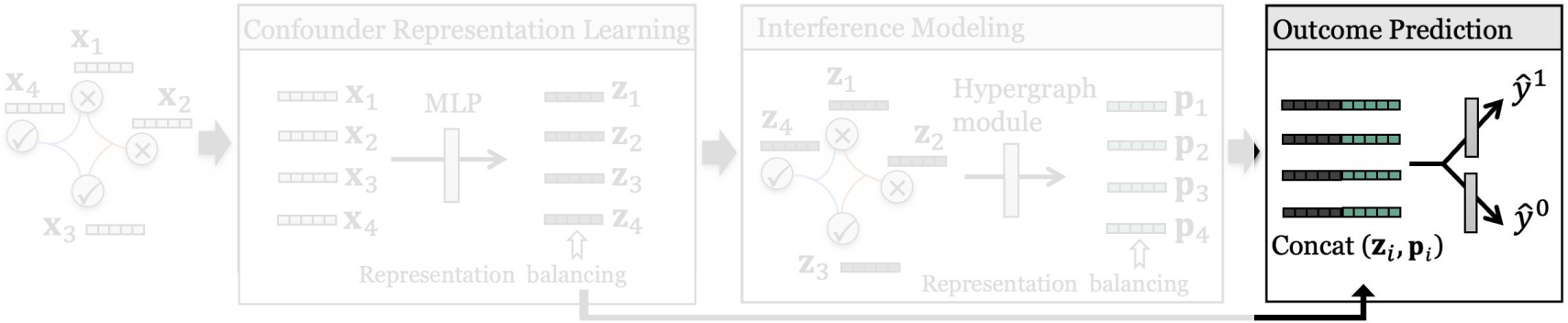
Interference representation in $l+1$-th layer

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{H}\mathbf{B}^{-1}\mathbf{H}^{T}\mathbf{D}^{-1/2}$$

vanilla Laplacian matrix for the hypergraph

[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Learning Causal Effects on Hypergraph



Outcome Prediction: predict the potential outcomes based on learned representations

$$\widehat{y}_i^1 = f_1([\mathbf{z}_i \| \mathbf{p}_i]),\ \widehat{y}_i^0 = f_0([\mathbf{z}_i \| \mathbf{p}_i])$$

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \alpha \mathcal{L}_b + \lambda \|\Theta\|^2$$

Model parameter regularization

Outcome prediction loss

Balancing loss

[1] Jing Ma, et al. Learning Causal Effects on Hypergraphs. KDD 2022

Experiments

- **Dataset**: real-world contact hypergraph
  - node=person, hyperedge=physical contact
- **Simulation**: treatment=wear face mask, outcome=COVID-19 infection

Method	Linear		Quadratic	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
LR	22.80 ±0.64	21.41± 0.74	414.17 ±3.94	192.80 ±2.97
CEVAE	19.36 ±0.80	8.63 ±0.78	315.01 ±2.53	188.47 ±4.27
CFR	25.23 ±0.01	18.28 ±0.02	392.56 ±4.33	189.75 ±4.80
Netdeconf	11.11 ±0.01	9.22 ±0.03	241.02 ±2.32	147.29 ±1.04
GNN-HSIC	9.38 ±0.44	6.91 ±0.38	114.28 ±3.62	81.21 ±2.53
GCN-HSIC	8.27 ±0.41	6.60 ±0.48	109.57 ±3.85	77.75 ±3.93
HyperSCI	**5.13 ±0.56**	**4.46 ±0.61**	**81.08 ±0.37**	**74.41 ±0.42**

Outcome simulation settings

No graph

Projected graph

Our method

The smaller, The better

HyperSCI outperforms all the baselines

Agenda

1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. **Causality-aided Machine Learning**
5. Applications and Future Directions
6. Conclusions

Explanation in ML

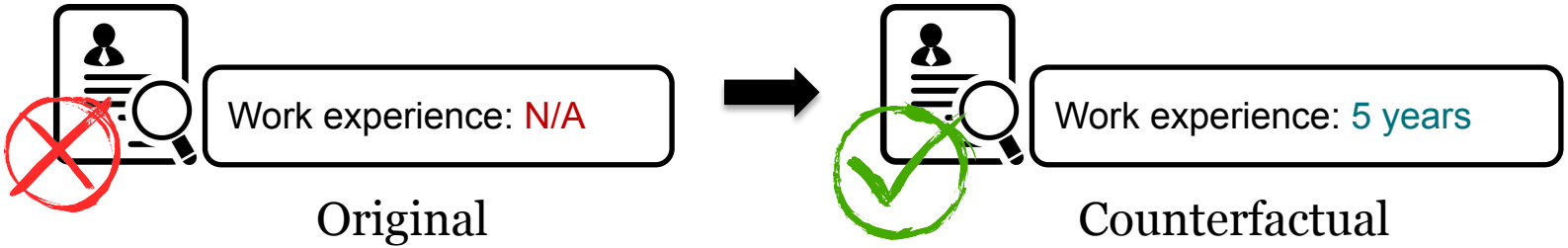
- Human-understandable explanations for machine learning are advantageous in several way

- **Feature importance**: An explanation can be beneficial to the applicant, e.g., helps an applicant understand which of their attributes were important in decision.
- **Fairness**: it can help an applicant challenge a decision if they feel unfair
- **Action**: an explanation provides the applicant with feedback that they can act upon for better outcome in the future.
- **Model improvement**: Explanations can help the model developers find bugs and improve the model.

Counterfactual Explanation

- **Counterfactual explanation (CFE)** [1] is a causality-related explanation strategy, which promotes model explainability by answering the key question

How to make small perturbations for a specific instance to get a desired prediction from the model?



- CFE can be particularly useful when it is not possible to directly observe or manipulate the factors of interest

[1] Verma, Sahil, et al. "Counterfactual explanations for machine learning: A review." arXiv preprint (2020).

Desiderata of Counterfactual Explanation

- **Validity**: generated counterfactuals have desired labels [1]

$$\arg\min_{x'} d(x, x') \text{ subject to } f(x') = y'$$

y' ← Desired label

Original objective

$$\arg\min_{x'} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x')$$

Differentiable, unconstrained objective

$d(x, x')$ ↑ Distance (e.g., L1/L2)

- **Actionability**: which features are mutable
  - E.g., “race”, “country of region” are immutable
  - Some papers ir $\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x')$

$\mathcal{A}$ ↑ Set of actionable features

- **Sparsity**: a counterfactual ideally should change smaller number of features in order to be most effective

[1] Wachter S, et al. Counterfactual explanations without opening the black box: Automated decisions and the GDPR[J]. Harv. JL & Tech., 2017, 31: 841.

Desiderata of Counterfactual Explanation

- **Data Manifold closeness**: a generated counterfactual is realistic in the sense that it is near the training data
  - include a penalty for adhering to the data manifold defined by the training set

$$l(x'; X) \leftarrow \text{training set}$$

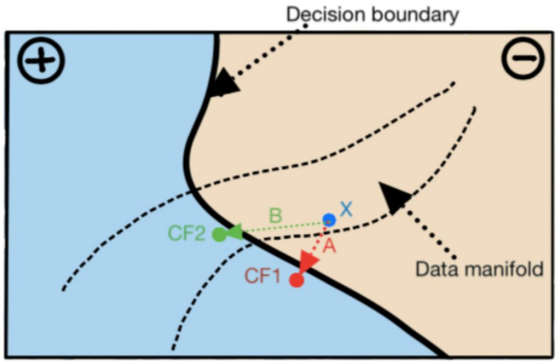
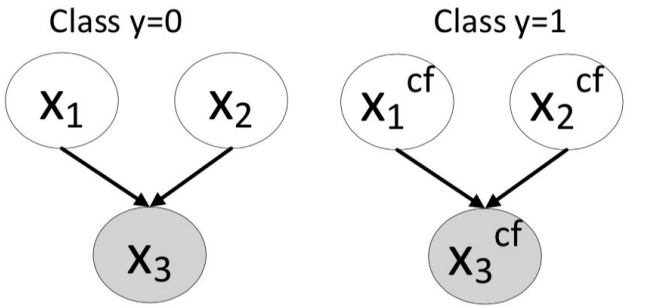


Figure [1]: Two counterfactuals (shown in red and green) are valid for the original datapoint (in blue). The red path is the shortest, whereas the green path adheres closely to the manifold of the training data.

[1] Verma, Sahil, et al. "Counterfactual explanations for machine learning: A review." arXiv preprint (2020).

Desiderata of Counterfactual Explanation

- **Causality**: changing one feature in the real world affects other features
  - E.g., getting a new educational degree necessitates increases the age



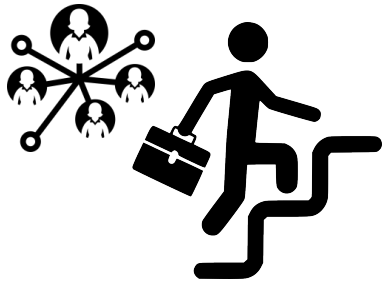
[1] Mahajan D, Tan C, Sharma A. Preserving causal constraints in counterfactual explanations for machine learning classifiers[J]. NIPS Workshop, 2019.

Evaluation of Counterfactual Explanation

- Commonly used datasets
  - Image - MNIST
  - Tabular - Adult income, German credit, Compas recidivism, etc.
  - Graphs – motifs, molecular graphs
- Metrics
  - **Validity**: the ratio of the counterfactuals that have the desired class label to the total number of counterfactuals.
  - **Proximity**: the distance of a counterfactual from the input datapoint.
  - **Sparsity**: the number of modified features
  - **Diversity**: diversity is encouraged by maximizing the distance between the multiple counterfactuals
  - **Causal constraint**: whether the counterfactuals satisfy the causal relation between features

Counterfactual Explanation on Graph

- Counterfactual explanations for graphs: the minimal perturbation to the input (graph) data such that the prediction changes
- Motivation & applications:
  - Drug discovery: CFE can help identify the minimal change one should make to a molecule with a desired property
  - Career plan: optimize professional network for better career outcome



Counterfactual Explanation for Node Classification

- CF-GNNExplainer [1]
  - based on GNN models
  - focus on node classification task
  - focus on perturbing graph structure

$$\mathcal{L} = \mathcal{L}_{pred}(v, \bar{v} \mid f, g) + \beta \mathcal{L}_{dist}(v, \bar{v}),$$

Prediction model

Original data for a node

Counterfactual data for the node

CFE generator

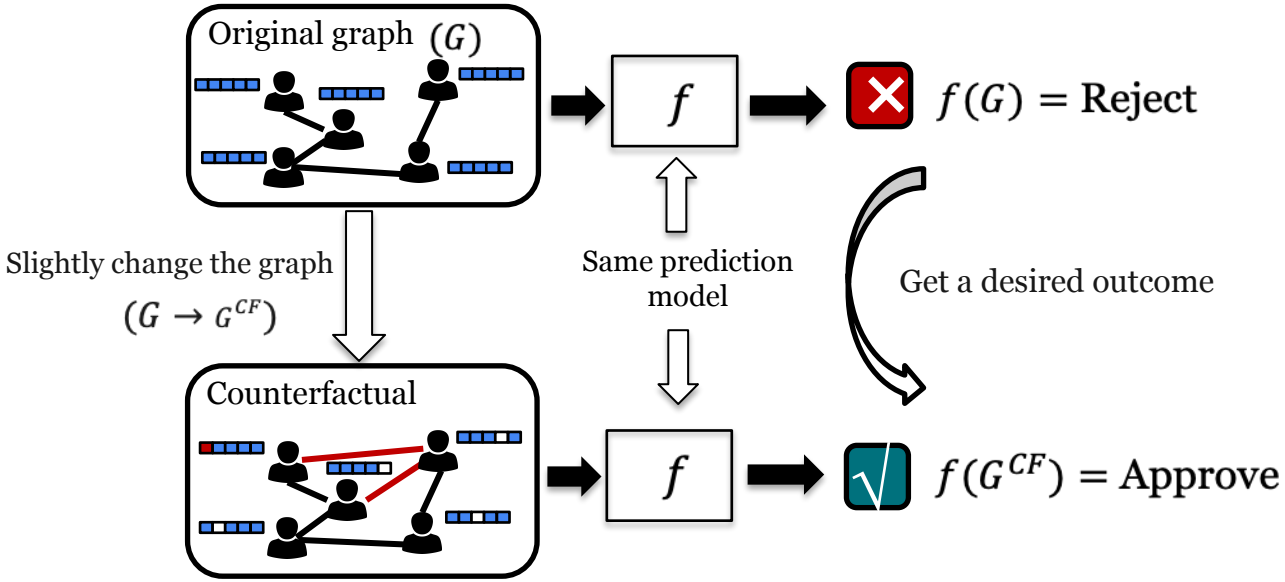
Element-wise difference

- Main idea of the method
  - iteratively remove edges (learn a perturbation matrix) from the original adjacency matrix based on matrix sparsification techniques
  - track of the perturbations that lead to a change in prediction
  - return the perturbation with the smallest change w.r.t. the number of edges

[1] Lucic A, ter Hoeve M, Tolomei G, et al. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks, AISTATS 2021.

Counterfactual Explanation for Graph Classification

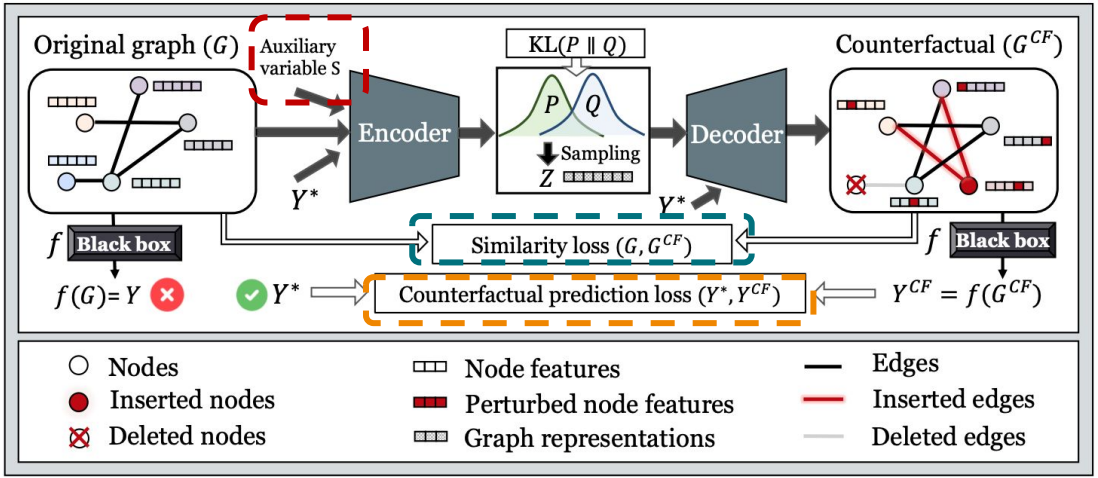
- Example: a graph ML model f trained for grant application decision-making



CLEAR: Graph Counterfactual Explanation

- CLEAR [1]: Model-agnostic, focus on graph classification task
- Use a graph-VAE based framework to enable optimization and generalization on graph data, promote the causality of CFEs with an auxiliary variable *S*

$$\mathcal{L} = \mathbb{E}_Q[d(G, G^{CF}) + \alpha \, l(f(G^{CF}), Y^*)] + \mathrm{KL}(Q(Z|G, S, Y^*) \| P(Z|G, S, Y^*))$$



[1] Ma, Jing, et al. "CLEAR: Generative Counterfactual Explanations on Graphs." NeurIPS 2022.

Fairness in ML

- Discrimination widely exists in the training data and algorithms of ML, leading to biases in ML predictions
- It is important to be aware of these potential sources of discrimination in machine learning and take steps to address them

October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

By Roberto Iriondo

Fairness in ML

- Fairness in ML: a model should treat all individuals or groups equally without discrimination based on sensitive attributes like race, gender, religion, …

- Fairness in ML is progressing at an astounding rate in recent years!

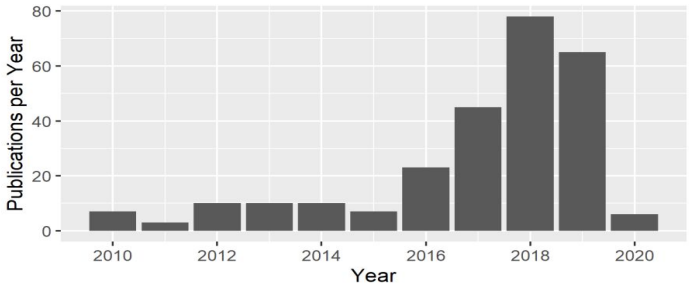


Figure [1]: Number of Papers related to Fairness in ML research

[1] Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey."

Fairness Notions

The prediction should be statistically independent with the sensitive attribute!

The prediction should not be influenced by the sensitive attribute from a causal perspective!

The predictions for similar individuals should be similar!

How to define and improve fairness in ML?

Fairness Notions

Group Fairness

Causality-based Fairness

Individual Fairness

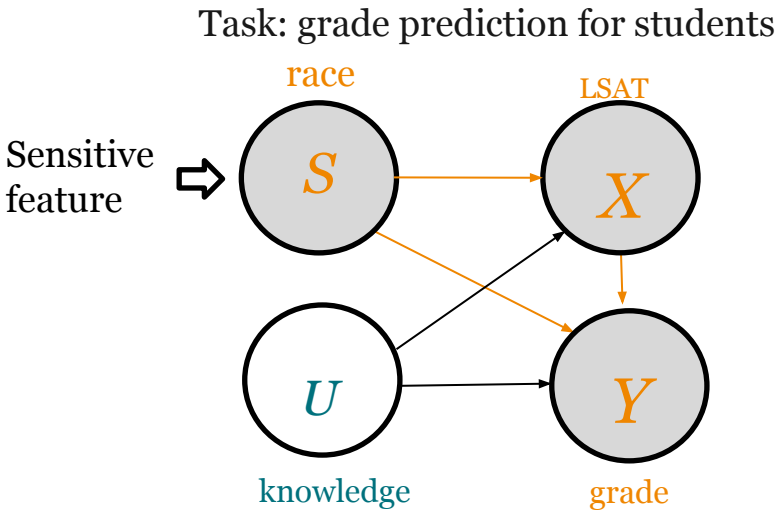
Fairness can be defined in many different ways!

Fairness Notions

- Fairness Through Unawareness (FTU): sensitive features are not explicitly used in the decision-making process.
  - Limitation: non-sensitive features may also be biased
- Equality of Opportunity: the positive rate are enforced to be the same between demographic subgroups conditional on the positive ground truth class labels $P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1)$
  - Limitation: the class label *Y* may also be biased

Causality-based Fairness

- Compared with other fairness notions, causality-based fairness can explicitly model how discriminations would happen, and how to identify, track, and eliminate them



- Discrimination is spreading through the causal pathways
- Descendants of the sensitive feature in the causal graph can also be biased

Counterfactual Fairness

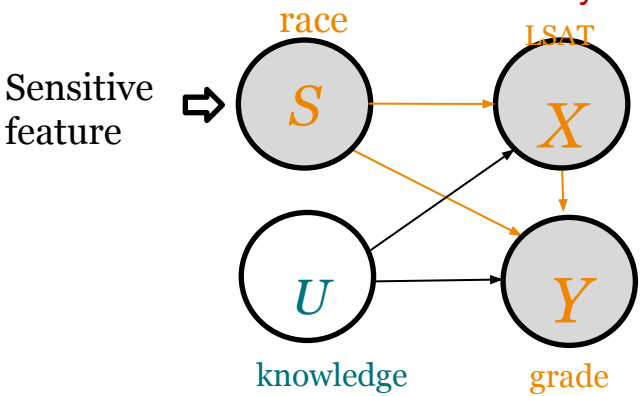
- Prediction $\hat{Y}$ is **counterfactually fair** [1] if under any features $X = x$ and sensitive attribute $S = s$:

$$P(\hat{Y}_{S\leftarrow s} = y | X = x, S = s) = P(\hat{Y}_{S\leftarrow s'} = y | X = x, S = s)$$

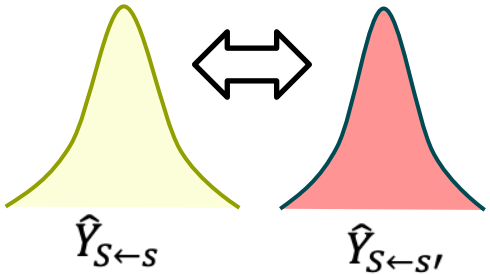
The value of the prediction if S had been set to s (s')
Notice: other features may change correspondingly.

Features

Sensitive attribute



A fair predictor should give the **same** prediction for an individual even if this individual had a **different race/gender/…**



[1] Kusner, Matt J., et al. "Counterfactual fairness." NIPS 2017.

Counterfactual Fairness

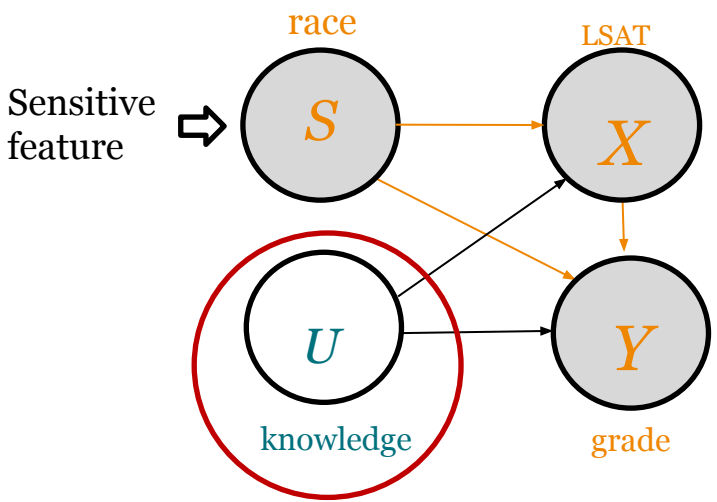
- FairLearning [1]
  - Fit the causal model
  - For each data instance i, compute the unobserved variables with $P(U|X,S)$

$$D' = \{s^i, x^i, y^i, u^i\}$$

  - Train a fair predictor

$$\hat{\theta} \leftarrow \text{argmin}_\theta \sum_{i' \in D'} l\left(y^{i'}, g_\theta(U^{i'}, x^{i'}_{\not\in S})\right)$$

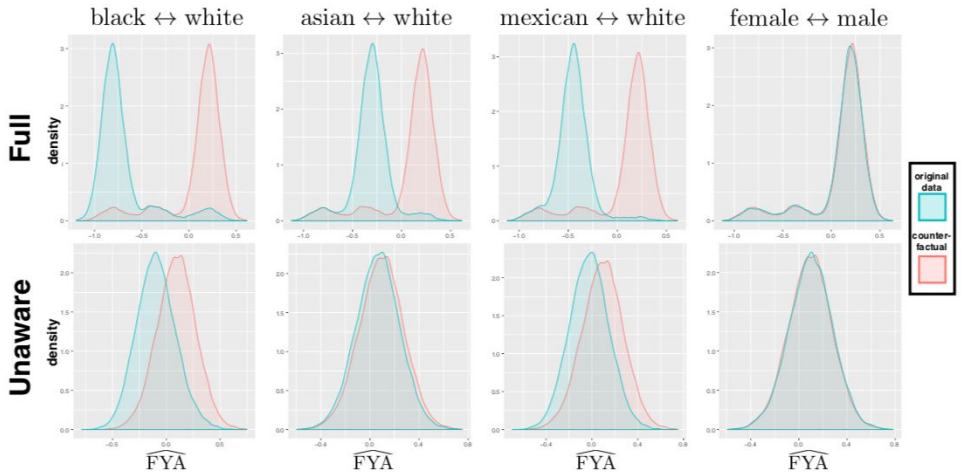
Non-descendants of S



[1] Kusner, Matt J., et al. "Counterfactual fairness." NIPS 2017.

Evaluation of Counterfactual Fairness

- ❑ Fit the parameters of causal model using the observed data
- ❑ Generate samples with *counterfactual* sensitive feature values
- ❑ Compare the predictions for the original and counterfactual data

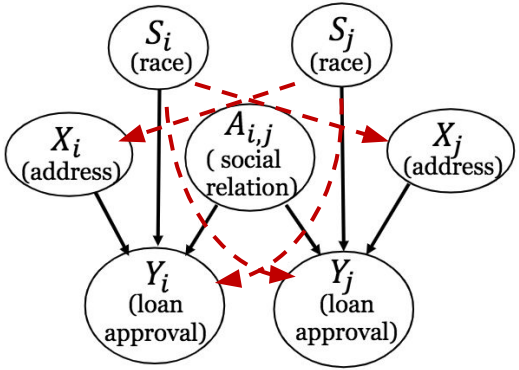


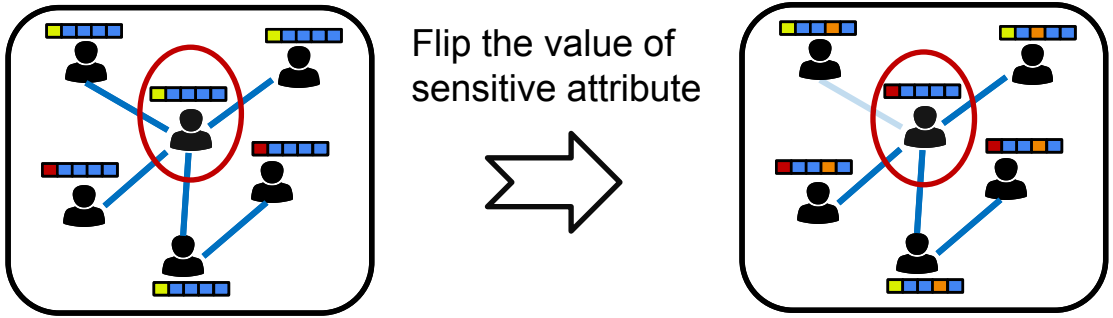
If a predictor is counterfactually fair => the distributions of these two predictions would be similar

[1] Kusner, Matt J., et al. “Counterfactual fairness.” NIPS 2017.

Counterfactual Fairness on Graphs

- In graphs, the sensitive attributes of each node's neighbors may causally affect the prediction w.r.t. this node (red dashed edges);





Counterfactual Fairness on Graphs

- **Graph counterfactual fairness** [1]: An encoder $Z_i = (\Phi(X, A))_i$ satisfies graph counterfactual fairness if for any node i:

$$P((Z_i)_{S \leftarrow s'} | X = \mathbf{X}, A = \mathbf{A}) = P((Z_i)_{S \leftarrow s''} | X = \mathbf{X}, A = \mathbf{A}),$$

Node representation for node i after intervention on S with value s'

Sensitive features

Node features

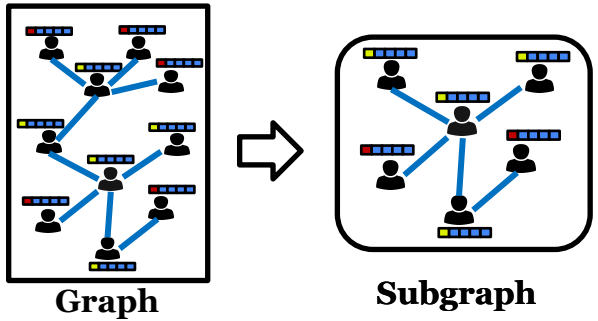
Graph structure

- Example: the prediction for one's loan application being approved should be the same regardless this applicant's and his/her friends' (connected in a social network) race information.

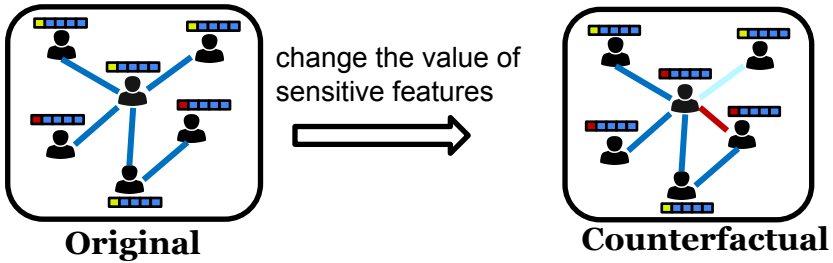
[1] Ma, Jing, et al. "Learning fair node representations with graph counterfactual fairness." *WSDM* 2022.

GEAR: Graph Counterfactual Fairness

1. Subgraph generation: split the input big graph into small subgraphs for each centroid node for better efficiency



2. Counterfactual (CF) augmentation: generate CFs for each subgraph with perturbation on sensitive features of different nodes



3. Fair representation learning: learn fair representations which elicit the same predicted label across different CFs w.r.t. the same node

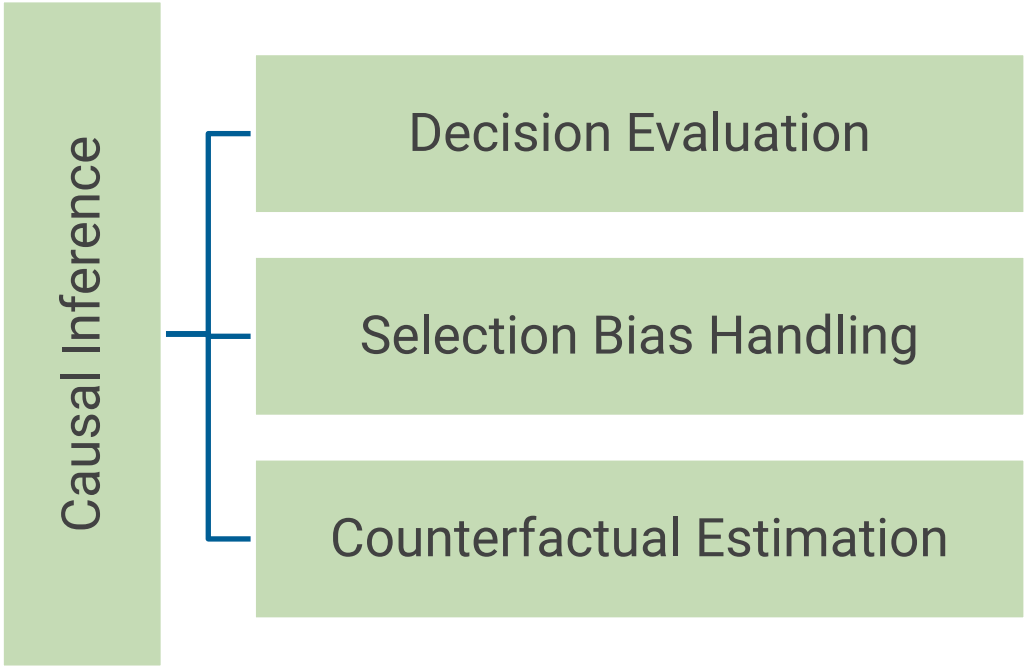
$$\mathcal{L}_f = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}((1-\lambda_s)d(\mathbf{z}_i,\bar{\mathbf{z}}_i)+\lambda_s d(\mathbf{z}_i,\underline{\mathbf{z}}_i))$$

Fairness loss: Encourage the node representations learned from the original graph and CFs to be the same

[1] Ma, Jing, et al. "Learning fair node representations with graph counterfactual fairness." *WSDM* 2022.

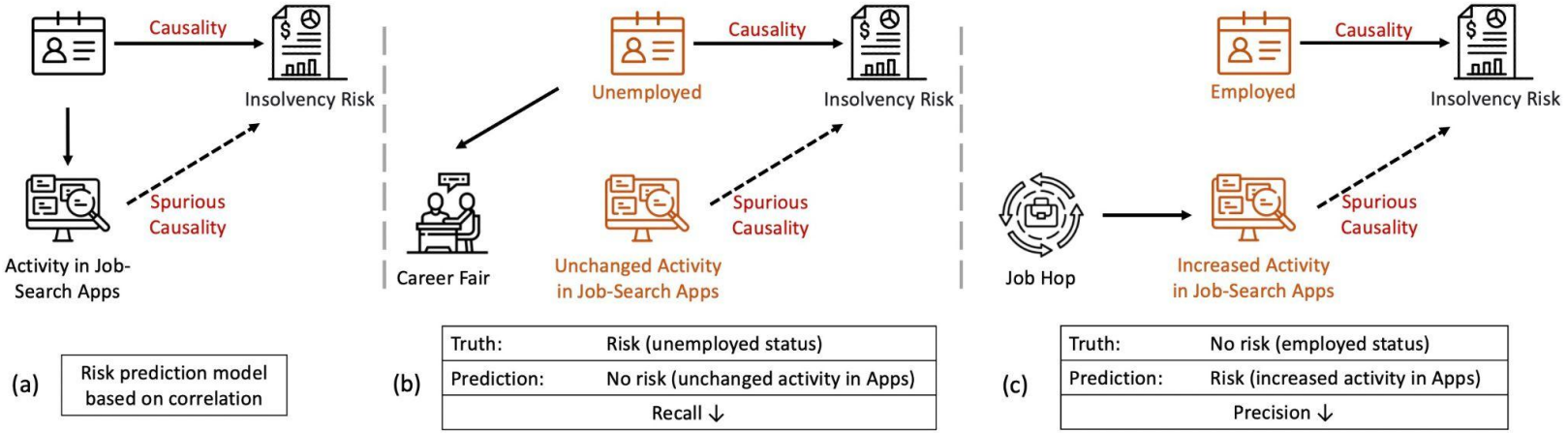
1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causality-aided Machine Learning
5. **Applications and Future Directions**
6. Conclusions

Applications



Decision Evaluation: Risk Prediction

- Always unable to produce trustworthy results on risk prediction tasks:
  - A lack of interpretability
  - No insight into cause relationships
  - Low precision and recall



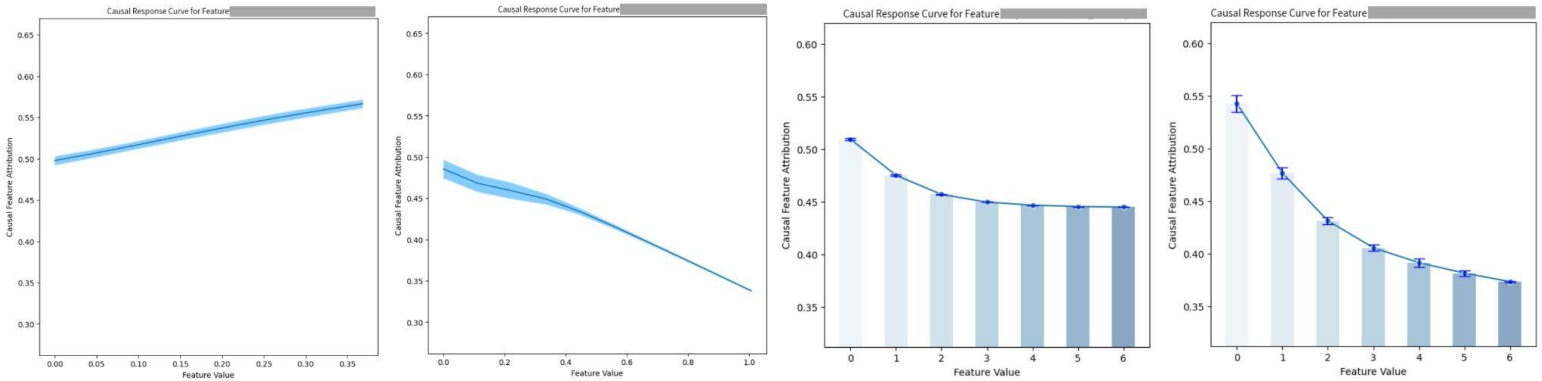
Decision Evaluation: Risk Prediction

- Without figuring out the true causal features, it is challenging to produce trustworthy predictions with **high recall and precision** in the risk prediction task
- Task-Driven Causal Feature Distillation model (TDCFD)
  - Incorporate the Potential Outcome Framework (POF) to model explanation.
  - Distill the task-driven causal feature attributions from the original feature values
    - Represent how much contribution each feature makes to this specific risk prediction task
  - Train on distilled causal feature attributions

	Synthetic data			Real corporate risk data		
Method	Accuracy	Precision	Recall	Accuracy	Precision	Recall
LR	0.92	0.64	0.54	0.83	0.21	0.16
SVM	0.94	0.68	0.65	0.87	0.40	0.27
KNN	0.91	0.55	0.60	0.91	0.62	0.47
RF	0.95	0.72	0.78	0.90	0.60	0.43
XGBoost	0.94	0.67	0.83	0.91	0.61	0.63
DNN	0.95	0.73	0.80	0.93	0.70	0.66
Transformer	0.96	0.77	0.85	0.93	0.71	0.71
TDCFD	**0.97**	**0.82**	**0.90**	**0.96**	**0.86**	**0.80**

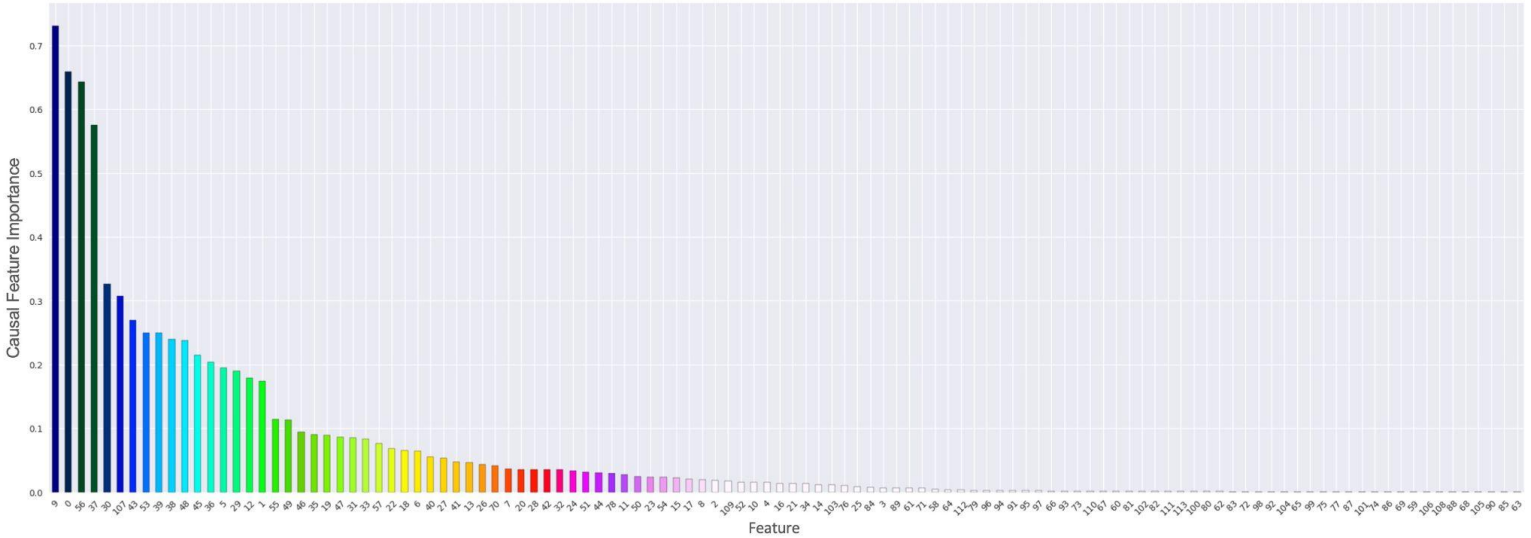
Decision Evaluation: Risk Prediction Explanations

- The causal response curve provides the expectation of outcome across a specific feature
- Observe the causal effect of feature changes on the outcome result
  - Directly applied to business decision-making, precise medication guidance, or other scenarios requiring actionable justification
  - Users do not have to repeatedly change their models to find the best combination of feature inputs. Can directly find out the optimal point for each feature



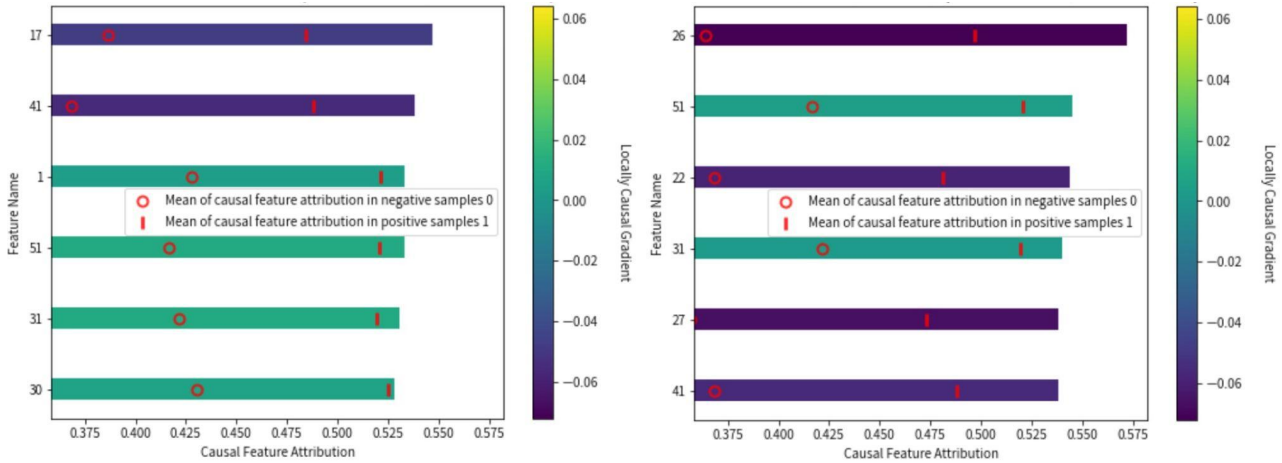
Decision Evaluation: Risk Prediction Explanations

- The causal feature importance clearly illustrates the importance of features
  - Based on real causal relationships rather than spurious correlations
  - Help people understand the **underlying data (traditional XAI based on underlying model cannot work)**



Decision Evaluation: Risk Prediction Explanations

- Individual prediction result explanations summarize the top-K important features
  - Explain why the model can produce this outcome
- Mean of causal feature attribution in positive and negative samples can be treated as the threshold of their contributions to the outcome



Decision Evaluation: Online Advertising

Will the ad attract user clicks?

Will a campaign increase sales?

Randomized experiments such as A/B testing?

Time-consuming and Expensive

Estimating the ad effect from observational data!

Decision Evaluation: Online Advertising

- Online Advertising as Causal Inference:

  Estimating the ad effect from observational data

Observational data → Logged feedback records under **current** advertising system's policy

Treatment W

Ads

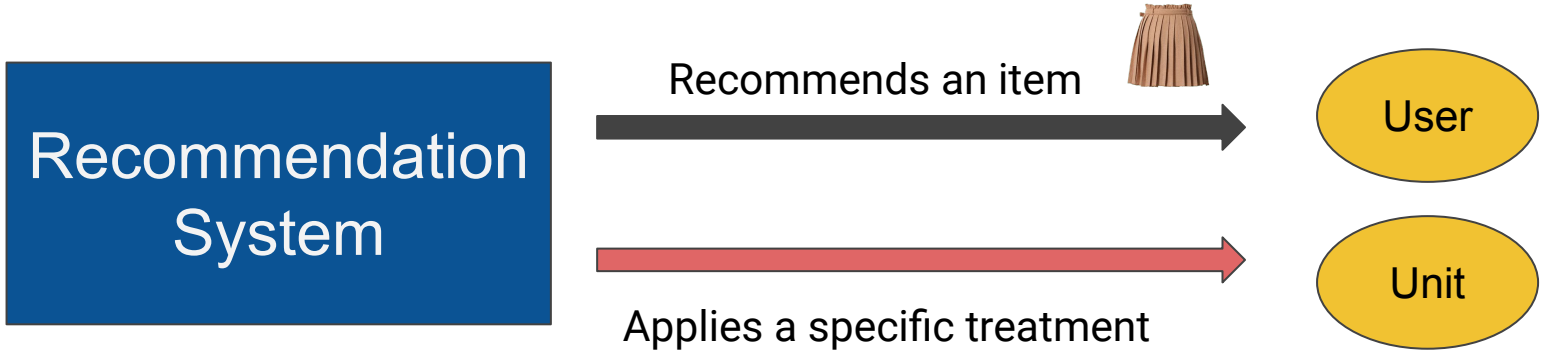
Outcome Y

Click

Variable X

Ad content

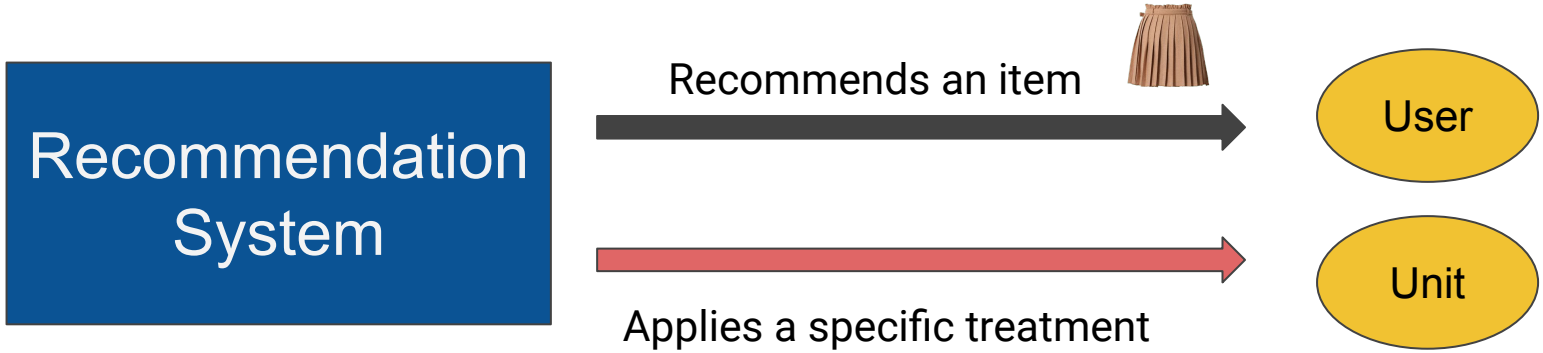
Selection Bias Handling: Recommendation

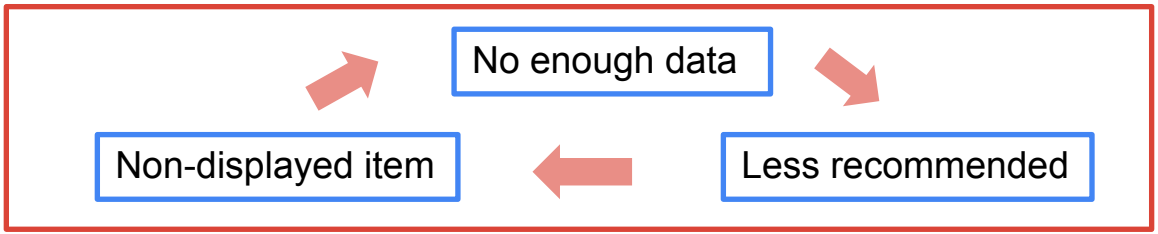


Selection bias:

- Users tend to rate the items that they like:
  - The horror movie ratings are mostly made by horror movie fans and less by romantics movie fans.
- The records in the datasets are **not representative** of the whole population.

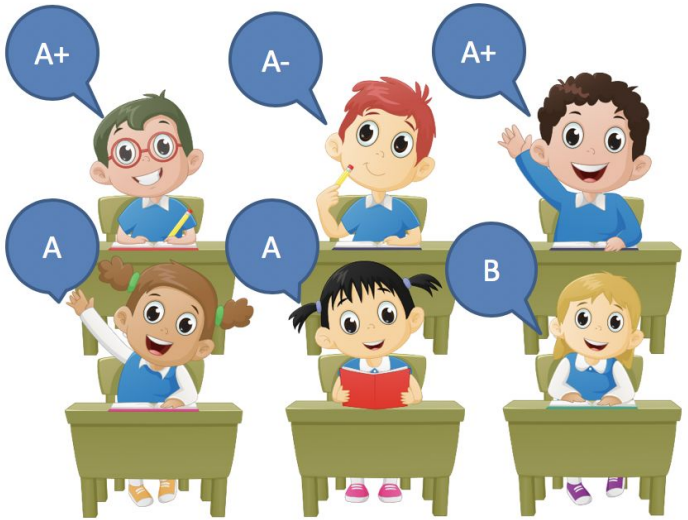
Selection Bias Handling: Recommendation





Counterfactual Estimation: Education

What would happen if the teacher adopted another teaching method?



Teachers can find the best teaching method for each individual!

Potential Direction: Perspective of Treatment

- **Similar treatments**:

  Finding neighbors that have similar treatments.

- **Multiple treatments**:

  Each treatment has different levels.

- **Continuous treatments:**

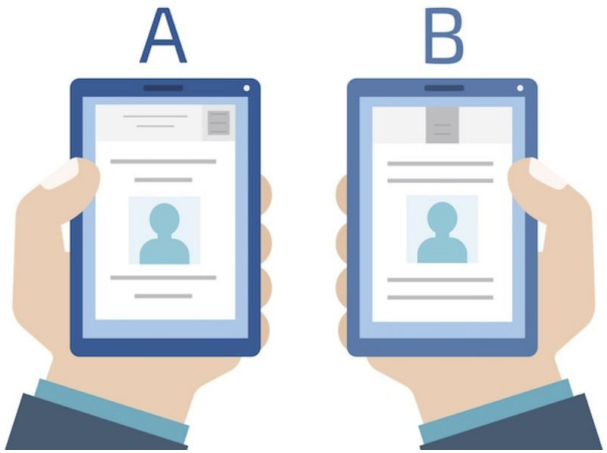
  Treatment can take values from a continuous range.

- **Causal interaction**:

  Identifying the effect of combinations of treatments.

Potential Direction: Evaluation

For real-world applications, how can we evaluate the performance of different causal inference methods?



Potential Direction: Data Fusion

Experimental Data

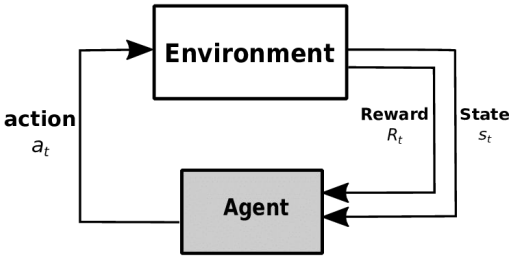
A small set of records under the **randomized experiments**

Observational Data

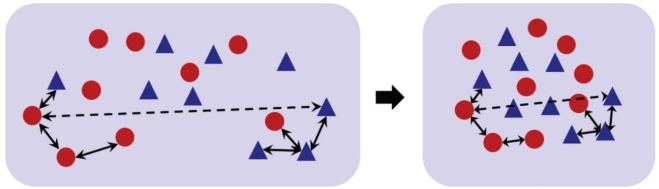
A large set of logged feedback records under **current system**

- ❑ Unobserved confounders
- ❑ Non-displayed items
- ❑ Improve existing methods
- ❑ ... ...

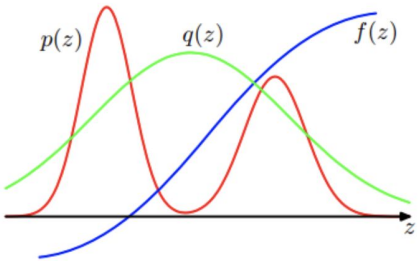
Potential Direction: Connecting Other Areas



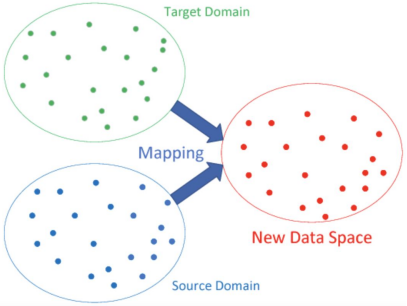
Reinforcement Learning



Representation Learning based Causal Inference



Statistical Sampling



Transfer Learning

Agenda

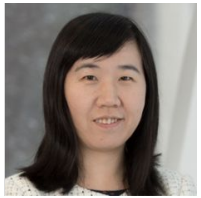
1. Background on Causal Inference
2. Representation Learning based Methods
3. Graph Neural Networks based Methods
4. Causality-aided Machine Learning
5. Applications and Future Directions
6. **Conclusions**

Resources

- Causal Effect Estimation: Recent Advances, Challenges, and Opportunities
  - https://arxiv.org/abs/2302.00848
- Learning Causality with Graphs, AI Magazine
  - https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12070
- Continual Treatment Effect Estimation: Challenges and Opportunities
  - https://arxiv.org/pdf/2301.01026.pdf
- Causal Inference in Recommender Systems: A Survey of Strategies for Bias Mitigation, Explanation, and Generalization
  - https://arxiv.org/pdf/2301.00910.pdf
- A Survey on Causal Inference
  - https://arxiv.org/abs/2002.02770
- A Survey of Learning Causality with Data: Problems and Methods, ACM Computing Surveys, 2020
  - https://arxiv.org/pdf/1809.09337.pdf

Acknowledgements

Aidong Zhang
Professor
University of Virginia



Jing Gao
Associate Professor
Purdue University

Liuyi Yao
Research Scientist
Alibaba

Yaliang Li
Research Scientist
Alibaba

Wei Chu
Researcher
Ant Group

Ruopeng Li
Researcher
Ant Group

Nikos Vlassis
Principal Scientist
Adobe

Stephen Rathbun
Professor
University of Georgia

Thank you!