Fast Proxy Experiment Design for Causal Effect Identification

Sepehr Elahi [1] Sina Akbari [1] Jalal Etesami [2] Negar Kiyavash [1] Patrick Thiran [1]

[1]EPFL, Switzerland [2]TUM, Germany

EPFL — NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

- Reformulated the minimum-cost intervention for causal effect identification problem using SAT and ILP frameworks.
- Developed algorithms that solve MCID up to six orders of magnitude faster than existing methods.
- Proposed a polynomial-time heuristic for MCID using adjustment sets.

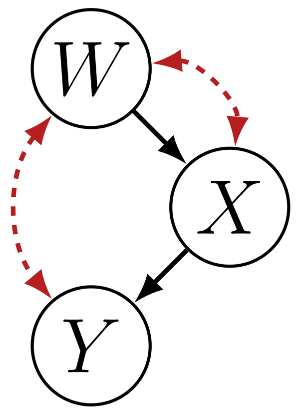
Motivation

Goal: Estimate the causal effect of a treatment on an outcome.

Challenges:

Impossible with obs. data (confounding) → Need for making interventions → Can't intervene on any var. (costly)

Example:



→ : Direct causal link

⇠⇢ : Confounded link

W: Tax on cigarettes (proxy)

X: Smoking (treatment)

Y: Lung cancer (outcome)

Cannot intervene on directly on X, but can intervene on W (proxy) to estimate the causal effect of X on Y.

Solution: Identify set of **M**inimum-**C**ost **I**nterventions to **ID**entify causal effect of X on Y: MCID problem

MCID Problem

Given: An acyclic directed mixed graph (ADMG) $\mathcal{G} = \langle V, \overrightarrow{E}, \overleftrightarrow{E} \rangle$, and $X, Y \in V$.

Goal: Find min-cost interventions to id. causal effect of X on Y:

$$\boldsymbol{\mathcal{I}}^* \in \operatorname*{argmin}_{\boldsymbol{\mathcal{I}} \in 2^{2^V}} C(\boldsymbol{\mathcal{I}}), \quad \text{s.t.}$$

$$\exists \text{ functional } f(\cdot) : \mathbb{P}_X(Y) = f(\{\mathbb{P}_{\mathcal{I}}\}_{\mathcal{I} \in \boldsymbol{\mathcal{I}}}).$$

MCID Problem is NP-Hard

The MCID problem is at least as hard as the weighted minimum hitting set problem, which is NP-hard to solve and approximate.

Solving the MCID Problem: Exact Algorithms

Previous approach: Required an exponential number of calls to an exponential-time algorithm (Akbari et al., 2022).

Our approach: Reformulate MCID problem as a **weighted Partial MAX-SAT** problem: **WPMAX-SAT**.

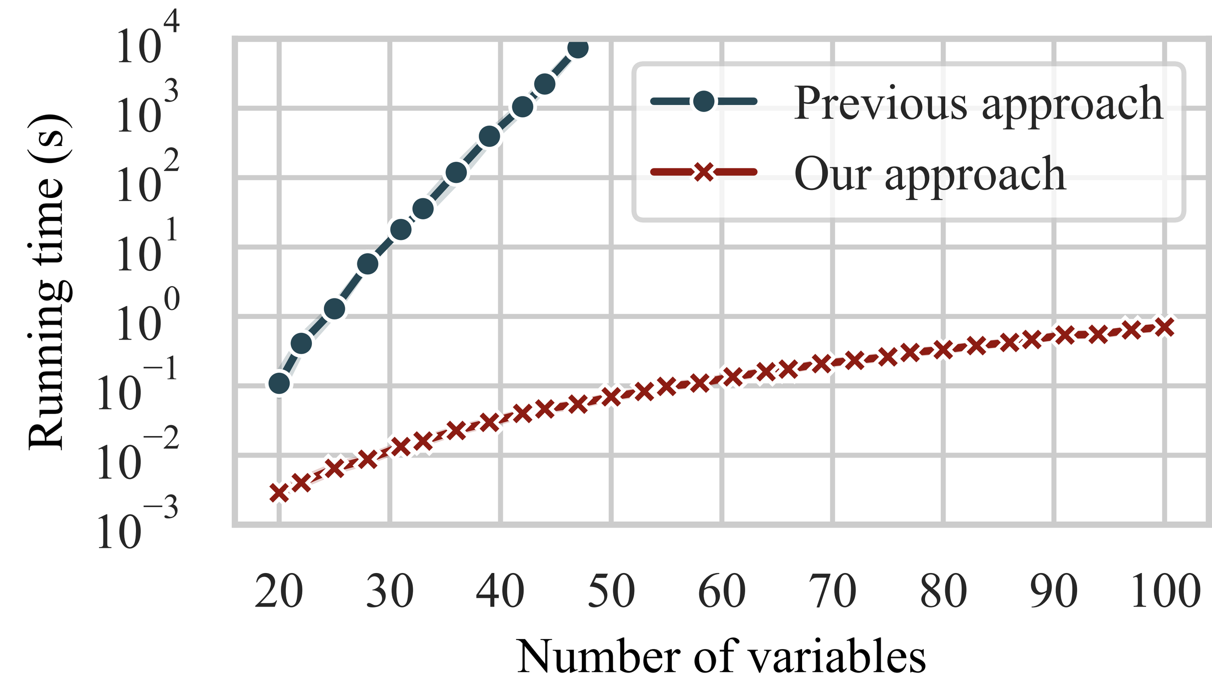
3-SAT Formula for MCID Problem

The 3-SAT formula constructed as in Section 3.1 given $\mathcal{G}$, X, and Y has a satisfying solution $\{x^*_{i,j}\}$ where $x^*_{i,0} = 0$ if $i \in \mathcal{I}$, if and only if $\mathcal{I}$ is a feasible solution to the MCID problem.

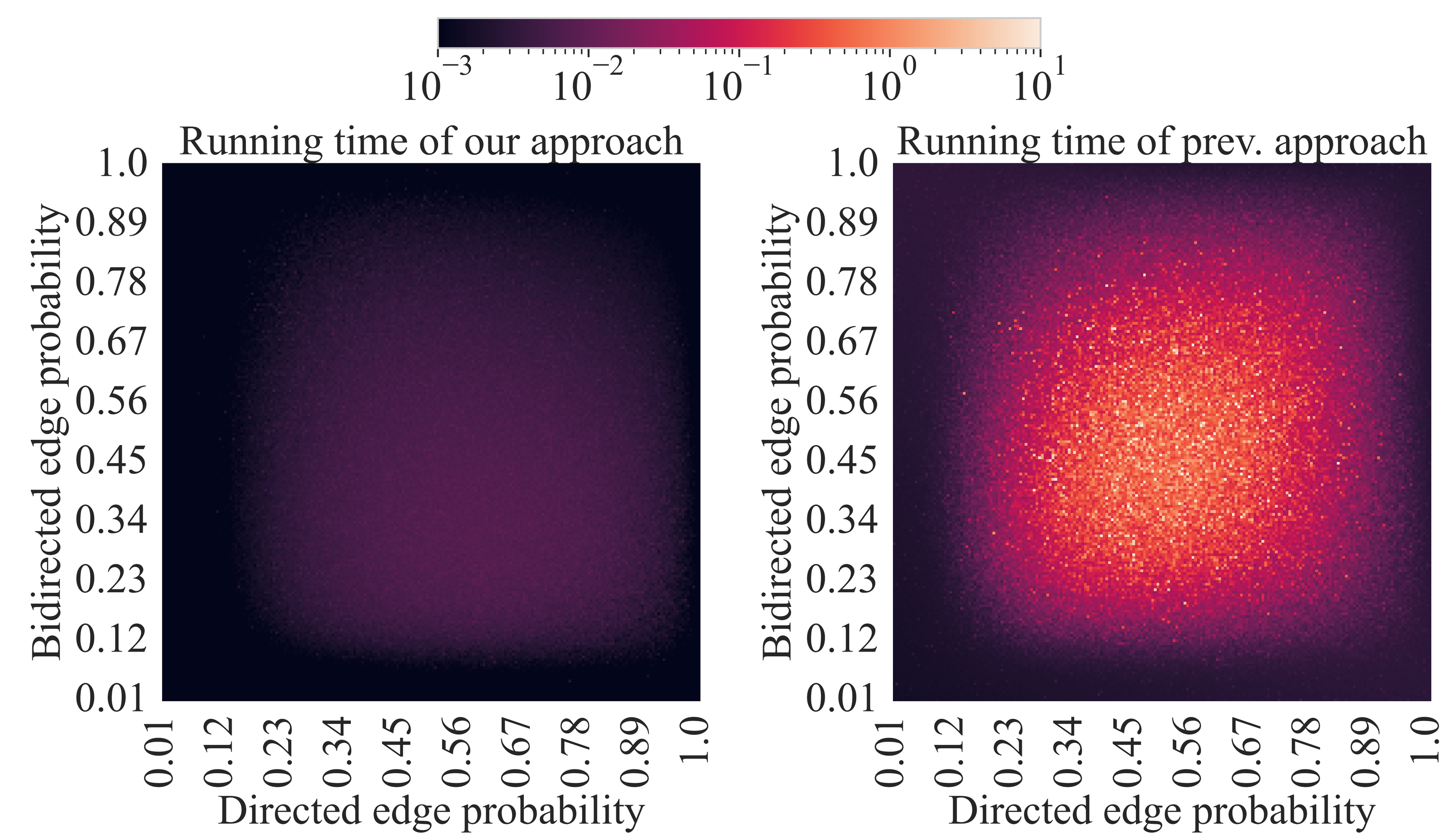
WP-MAXSAT Algorithm for MCID Problem: Find the optimal solution to the 3-SAT formula, $\{x^*_{i,j}\}$ that minimizes $\sum_{i=1}^m (1 - x^*_{i,0}) C(v_i)$.

WPMAX-SAT Results: Running Time (⇓)

Experiment #1: 30,000 ADMGs with varying (bi)/directed sparsities with $\{20, \ldots, 100\}$ variables.



Experiment #2: 40,000 ADMGs with varying (bi)/directed sparsities with 20 variables.



Solving the MCID Problem: Heuristics

Generalized adjustment criterion: intervention set $\mathcal{I}$ and set Z s.t.

$$\mathbb{P}_X(Y) = \mathbb{E}_{\mathbb{P}_\mathcal{I}}[\mathbb{P}_\mathcal{I}(Y \mid X, Z)].$$

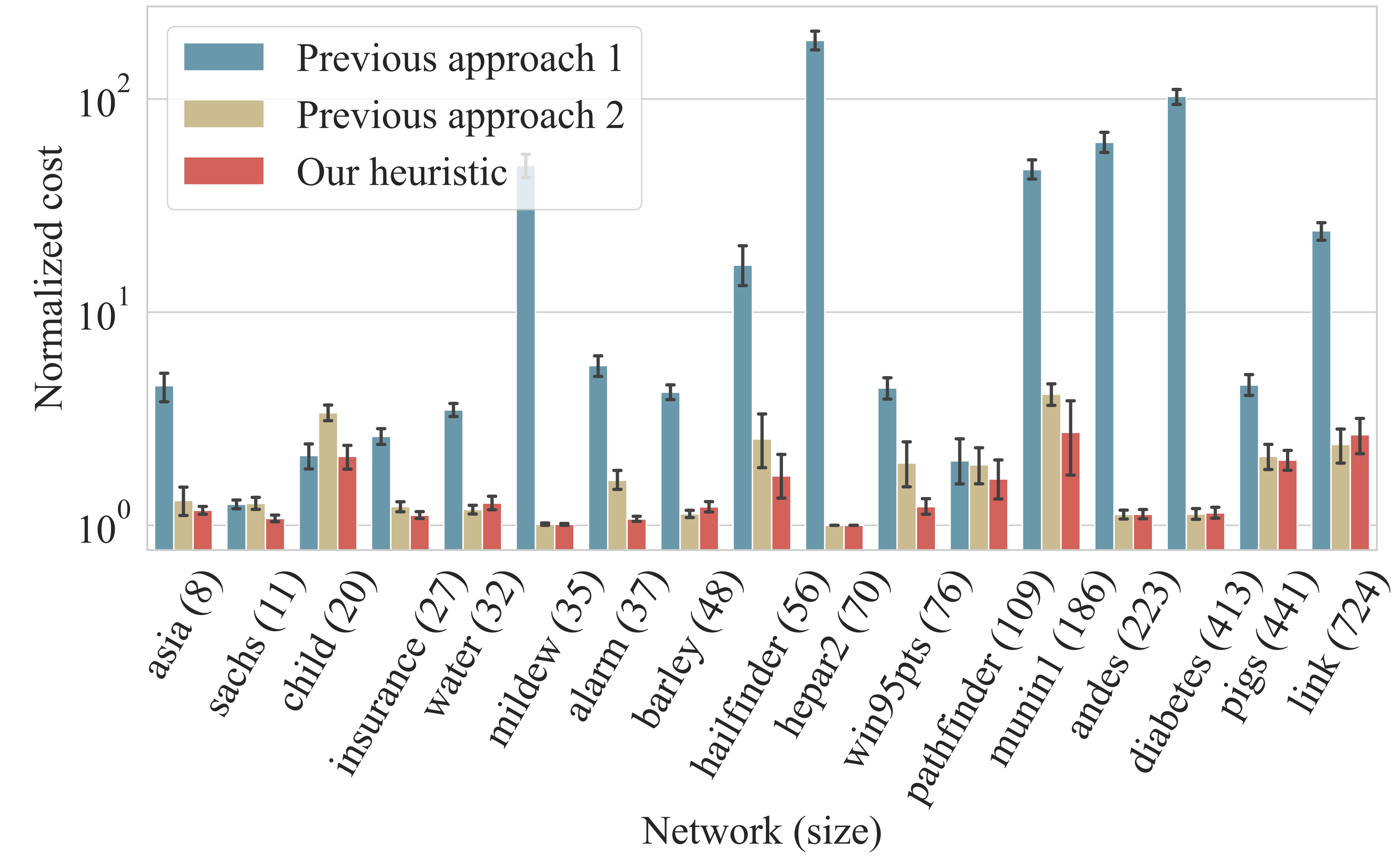
Surrogate problem for MCID: $\mathcal{I}^* \in \operatorname{argmin}_{\mathcal{I} \in 2^V} C(\mathcal{I})$

s.t. $\mathbb{P}_X(Y) = \mathbb{E}_{\mathbb{P}_\mathcal{I}}[\mathbb{P}_\mathcal{I}(Y \mid X, Z)]$ *for some* Z.

- Reduces to minimum cut $\Longrightarrow$ Solvable in polynomial time.
- Solves a special case of MCID $\Longrightarrow$ efficient heuristic for MCID.

Heuristic Results: Cost (⇓)

Experiment #3: Real-world DAGs from the Bayesian Network Repository.



Future Work

- Do log-factor approximation algorithms for MCID exist?
- Incorporating the sample complexity of the estimators corresponding to each intervention set.

More Information

[sepehr.elahi,sina.akbari]@epfl.ch