Graph Neural Networks for Probabilistic Causal Discovery

Rezaur Rashid[1] **Gabriel Terejanu**[1]

[1]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

Abstract

Conventional causal discovery algorithms face significant challenges in dealing with large-scale observational datasets and in capturing global structural information. To address these limitations, we introduce a novel graph neural network (GNN)–based probabilistic framework for causal structure learning that generates a probability distribution over possible causal graphs. By encoding the node and edge attributes into a unified graph representation, our framework enables the GNN to learn the complex causal structure directly from the data augmented with statistical and information-theoretic measures, which exploit the local and global data properties. Our approach outperforms benchmark methods, both traditional and recent non-GNN-based, in terms of accuracy and scalability on synthetic and real-world datasets.

1 INTRODUCTION

Causal inference from observational data is a fundamental task in many disciplines and forms the backbone of many practical decision-making procedures as well as theoretical developments. Classical causal discovery algorithms test hypotheses of conditional independences to learn causal structure Spirtes et al. [2001]. Score-based causal discovery algorithms optimize fit scores over various graph structures Chickering [2002]. While effective in many situations, these approaches suffer from exponential run-times and combinatorial explosions in statistic complexity as the data sets grow Heckerman et al. [1995]. Recent machine learning advances optimize continuous functions to impose acyclicity and achieve improved computational performance NOTEARS algorithm Zheng et al. [2018]. These approaches typically identify a single best causal graph rather than a probability distribution over multiple possible graphs, which can limit its ability to account for uncertainty in the causal discovery process.

Graph neural networks (GNNs) have revolutionized learning on data with underlying graph structures. From social network analysis to molecular property prediction (e.g., modeling interactions of atoms in a chemical molecule) Kipf and Welling [2016], Velickovic et al. [2017], Graph Convolutional Networks (GCN) and other sophisticated variants such as Graph Attention Networks (GAT), have successfully exploited node and edge features to learn deep and hierarchical representations Zhou et al. [2020], Waikhom and Patgiri [2023]. Despite their success in areas such as network analysis and bioinformatics Hamilton et al. [2017], Lacerda et al. [2012], these methods have yet to be fully integrated into causal discovery frameworks. Such developments strongly motivate and justify the idea of utilizing GNNs for causal learning tasks Yu et al. [2019], Brouillard et al. [2020], Peters et al. [2017].

This research proposes a novel GNN-based probabilistic framework for causal discovery that addresses the limitations of our prior work using causal pairs by capturing global information in the graph structure Rashid et al. [2022]. This work makes several key contributions: (1) it enhances causal structure learning by refining the probability distribution of all possible digraphs; (2) it provides a comprehensive understanding of causal discovery by learning a spectrum of causal graphs instead of producing a single deterministic graph; and (3) it outperforms conventional non-GNN-based methods in terms of accuracy and scalability.

2 METHODOLOGY

Given an observational dataset $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i$ represents an attribute with multiple observations, the goal is to learn possible directed acyclic graph (DAG) representations that capture the causal relationships among these attributes. This involves predicting the direction of edges (forward, reverse, or no edges) between nodes or data attributes.

2.1 DATA REPRESENTATION

We first construct a fully connected graph $\mathcal{G} = (V, E)$, where V is the set of all attributes in the observational dataset, and E is the set of edges between nodes (attributes) such that every node is connected with every other node which leads to $N(N-1)/2$ edges in the graph for a dataset with N attributes. We then extract statistical and information-theoretic measures, such as mutual information, entropy, and conditional independence test, on the attributes in the observational dataset to represent each node with 11 features and each edge with 115 features between node pairs in the graph. Additionally, we use the probability distribution over the edge direction from our causal-pairs model as an additional edge feature. In total, we have 118 features for each edge in the graph.

2.2 DEVELOPPING GNN MODEL

We use a graph neural network model, trained on synthetic datasets with underlying causal graphs, as an edge classifier with supervised learning to infer the probability distribution over the edge direction. We use the 'SAGEConv' layer from the GraphSAGE framework as our GNN model which performs the message-passing operation and updates the node and edge features iteratively. The model is then able to capture local and global dependencies in the graph structure and augment the accuracy of causal relations between nodes. Finally, after multiple rounds of message passing, we can obtain the final node and edge features that represent each node and each edge in the graph, and use them to predict the probabilities of edge directions (forward, reverse, or no edge) between any pair of nodes (attributes).

2.3 PROBABILISTIC INFERENCE

Samples from the probability distribution generated by the GNN model are digraphs, with no guarantees of acyclicity. We enforce acyclicity by approximating the most likely probability distribution of directed acyclic graphs (DAGs) using the maximum spanning DAG approach Rashid et al. [2022]. Finally, we employ the same method described in our prior work to estimate a sample digraph (PG), a maximum likelihood digraph (MLG), a sample DAG (PDAG), and a maximum likelihood DAG (MLDAG) from these two new probability distributions over all graph structures.

3 RESULT

We evaluate the causal discovery performance of our GNN-based framework on synthetic datasets and real-world datasets and compare it with traditional and recent non-GNN-based methods. The synthetic datasets generation process and evaluation metrics are structural Hamming distance (SHD), true positive rate (TPR), and false positive rate (FPR) Rashid et al. [2022]. In Table 1, we show that for Scale-Free (SF) graph structure, our GNN-based framework significantly improves upon SHD compared to traditional and causal-pairs methods, and compares favorably to LiNGAM, DAG-GNN and NOTEARS-MLP, which implies better recovery of causal structures. Moreover, from the TPR and FPR results, the GNN-based framework outperforms all other methods in identifying true causals while avoiding false positives. Additional results are provided in the supplementary material section.

Table 1: Comparison of edge probability model trained on GNN framework. The means and standard errors of the performance metrics are based on the 80 Scale-Free (SF) graph structures in the test data.

Method ↓ / Metrics →	**SHD/d**	**TPR**	**FPR**
GNN-PG	1.88±0.08	0.51±0.02	0.30±0.01
GNN-MLG	1.85±0.13	0.20±0.02	0.01±0.00
GNN-PDAG	1.55±0.07	0.56±0.02	0.19±0.01
GNN-MLDAG	1.40±0.11	0.48±0.03	0.08±0.01
CausalPairs-PG	2.02±0.12	0.31±0.01	0.26±0.02
CausalPairs-MLG	1.97±0.13	0.12±0.01	0.03±0.01
CausalPairs-PDAG	1.96±0.12	0.30±0.01	0.21±0.02
CausalPairs-MLDAG	1.88±0.13	0.20±0.01	0.09±0.01
PC	1.93±0.15	0.17±0.02	0.08±0.01
GES	1.43±0.11	0.51±0.03	0.26±0.04
LiNGAM	1.68±0.11	0.35±0.02	0.34±0.04
DAG-GNN	1.75±0.12	0.24±0.02	0.02±0.00
NOTEARS-MLP	1.36±0.11	0.47±0.02	0.12±0.02

4 CONCLUSION

In this work, we present a probabilistic causal discovery framework utilizing graph neural networks (GNN). Our results on both synthetic and real-world datasets demonstrate that GNN methods outperform causal-pair methods by leveraging multi-attribute information captured by the GNN. Future work will focus on incorporating acyclicity constraints directly into the GNN framework, potentially improving the robustness and accuracy of causal discovery. Additionally, exploring more sophisticated GNN architectures could further enhance the performance of our approach.

ACKNOWLEDGEMENT

The research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-22-1-0035. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33: 21865–21877, 2020.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197–243, 1995.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Rezaur Rashid, Jawad Chowdhury, and Gabriel Terejanu. From causal pairs to causal graphs. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 802–807. IEEE, 2022.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Natalie Schluter. On maximum spanning dag algorithms for semantic dag parsing. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pages 61–65, 2014.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

Lilapati Waikhom and Ripon Patgiri. A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artificial Intelligence Review*, 56(7):6295–6364, 2023.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. Pmlr, 2020.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

Graph Neural Networks for Probabilistic Causal Discovery (Supplementary Material)

Rezaur Rashid[1] **Gabriel Terejanu**[1]

[1]Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

A EQUATIONS USED IN THE TABLES

Assuming we have n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x1} \dots \mathbf{x_d}] \in R^{n \times d}$, causal discovery attempts to estimate the underlying causal relations encoded by the directed acyclic graph (DAG), $\mathcal{G}_{\text{DAG}} = (V, E)$. V contains of nodes associated with the observed random variables X_i for $i = 1 \dots d$ and the edges in E associate the causal relations encoded by $\mathcal{G}_{\text{DAG}}$. In other words, the presence of the edge $i \to j$ corresponds to a direct causal relation between X_i (cause) and X_j (effect).

Our approach uses a graph neural network model to predict the probability $p(e_{ij}|f)$ of an edge e_{ij} between nodes X_i and X_j given their feature representations.

$$p(e_{ij}|\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}) \quad = \quad f([\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}]), \text{ for } i < j \tag{1}$$

Here,

- $\mathbf{h}_i$ and $\mathbf{h}_j$ represent the feature vectors of nodes X_i and X_j after the GNN's message passing and aggregation operations.
- $\mathbf{e}_{ij}$ represents the feature vector of the edge e_{ij} between nodes X_i and X_j.
- $[\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}]$ denotes the concatenation of the feature vectors of nodes X_i and X_j and the edge features $\mathbf{e}_{ij}$.
- The function f represents the GNN classifier that outputs the probability $p(e_{ij}|\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij})$ of there being an edge $e_{ij} \in [-1, 0, 1]$.

$$e_{ij} = \begin{cases} -1: & j \to i, \text{ causal relation exists from } X_j \text{ to } X_i \\ 0: & i \not\to j \text{ and } j \not\to i, \text{ no direct causal relation between } X_i \text{ and } X_j \\ 1: & i \to j \text{ causal relation exists from } X_i \text{ to } X_j \end{cases}$$

After computing the probability distributions of causal relations between all the pairs in the dataset, a naive way to construct the probability distribution of a digraph $\mathcal{G}$ is to assume that the causal-pairs are independent.

$$p(\mathcal{G}|f) = \prod_{i<j} p(e_{ij}|f) \tag{2}$$

Having this rich probabilistic information on all the causal relationships in the dataset, one may choose to generate the maximum likelihood digraph.

$$\mathcal{G}_{\text{ML}} = \arg\max_{\mathcal{G}} p(\mathcal{G}|f) \tag{3}$$

Note that the samples from the probability distribution, Eq. 2, and the maximum likelihood estimate, Eq. 3, are digraphs with no guarantees that they are acyclic. We then generate the most likely probability distribution of directed acyclic graphs

(DAG) leveraging the maximum spanning DAG approach Schluter [2014] as well as estimate a representative DAG using the maximum likelihood estimate using the following equations. We have used the following labels for our approaches: PG given by Eq. 2, MLG given by Eq. 3, PDAG given by Eq. 4, and MLDAG given by Eq. 5; see also Rashid et al. [2022] for an in-detailed derivation for the equations.

$$p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) = \prod_{\pi_{\text{ML}}^{-1}[i] < \pi_{\text{ML}}^{-1}[j]} p(e_{i \to j}|f) \tag{4}$$

$$\mathcal{G}_{\text{DAG}} \approx \arg\max_{\mathcal{G}} p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) \tag{5}$$

B ADDITIONAL SIMULATION RESULTS

Table 2 lists additional simulation results for Erdos-Renyi (ER) graph structures and Table 3 show results for real-world protein network dataset Sachs et al. [2005].

Table 2: Comparison of edge probability model trained on GNN framework. The means and standard errors of the performance metrics are based on the 80 Erdos-Renyi (ER) graph structures in the test data.

Method ↓ / Metrics →	**SHD/d**	**TPR**	**FPR**
GNN-PG	2.08±0.11	0.52±0.02	0.52±0.06
GNN-MLG	2.17±0.17	0.25±0.02	0.01±0.00
GNN-PDAG	1.75±0.11	0.61±0.03	0.28±0.03
GNN-MLDAG	1.66±0.15	0.54±0.03	0.13±0.02
CausalPairs-PG	2.38±0.14	0.39±0.02	0.72±0.10
CausalPairs-MLG	2.32±0.17	0.15±0.02	0.07±0.01
CausalPairs-PDAG	2.30±0.15	0.38±0.02	0.61±0.09
CausalPairs-MLDAG	2.18±0.16	0.28±0.02	0.29±0.05
PC	2.40±0.21	0.17±0.02	0.22±0.04
GES	1.78±0.13	0.48±0.02	0.87±0.15
LiNGAM	1.97±0.13	0.43±0.02	1.04±0.17
DAG-GNN	2.10±0.17	0.27±0.02	0.06±0.00
NOTEARS-MLP	1.33±0.10	0.58±0.02	0.32±0.06

Table 3: Comparison of our GNN-based probabilistic methods with GES and NOTEARS-MLP that were applied to both standardized and non-standardized protein network datasets. DAG-GNN and NOTEARS-MLP results for non-standardized data are reported from the original manuscriptsYu et al. [2019], Zheng et al. [2020]

Dataset type →	**Standardized**			**Non-standardized**		
Method ↓ / Metrics →	**Predicted Edges**	**Correct Edges**	**Reversed Edges**	**Predicted Edges**	**Correct Edges**	**Reversed Edges**
GNN-PG	19.68	6.60	6.98	19.40	5.86	7.79
GNN-MLG	12.07	5.13	5.64	13.81	5.48	6.86
GNN-PDAG	17.09	6.96	5.81	16.74	4.14	8.62
GNN-MLDAG	14.12	6.96	5.81	12.54	4.71	7.77
CausalPairs-PG	36.14	6.70	7.77	38.01	6.21	8.26
CausalPairs-MLG	9.82	3.04	4.26	10.41	1.52	4.04
CausalPairs-PDAG	33.16	7.42	6.62	34.81	6.47	7.49
CausalPairs-MLDAG	18.48	4.91	5.41	20.60	4.71	6.32
GES	34.00	5.50	9.50	34.00	5.50	9.50
LiNGAM	36.00	4.00	11.00	36.00	4.00	11.00
DAG-GNN	6.00	1.00	5.00	18.00	8.00	3.0
NOTEARS	42.33	5.83	7.18	13.00	7.00	3.00