BEYOND CAUSAL PAIRS: A PROBABILISTIC APPROACH TO CAUSAL STRUCTURE LEARNING FROM CAUSE-EFFECT PAIR RELATIONSHIPS USING GRAPH NEURAL NETWORK

by

Md Rezaur Rashid

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing & Information Systems

Charlotte

2024

Approved by:

Dr. Gabriel Terejanu

Dr. Razvan Bunescu

Dr. Siddharth Krishnan

Dr. Minwoo Lee

Dr. Shannon Reid

ABSTRACT

MD REZAUR RASHID. Beyond Causal Pairs: A Probabilistic Approach to Causal Structure Learning From Cause-Effect Pair Relationships Using Graph Neural Network. (Under the direction of DR. GABRIEL TEREJANU)

Machine learning has risen to the forefront of scientific research due to its unparalleled predictive capabilities. As a result, researchers have become increasingly interested in uncovering the underlying causal structures that govern the relationships between variables in a system. These causal structures, often represented as directed acyclic graphs (DAGs), provide insights into how changes in one variable may directly or indirectly affect other variables, enabling a deeper understanding of the complex interactions within the system. While it is essential to constrain a model by minimizing spurious correlations and conducting "What-If" analyses, learning causal relationships from observational data, known as causal discovery, remains an active and challenging research area. This is due to factors like finite sampling, unobserved confounding factors, and measurement errors. Current approaches, including constraint-based and score-based methods, often struggle with high computational complexity because of the combinatorial nature of estimating DAGs. Inspired by the workshop on the Causality Challenge 'Cause-Effect Pair' at the Neural Information Processing Systems in 2013, this dissertation adopts a novel approach, generating a probability distribution over all possible graphs based on cause-effect pair features proposed in response to the workshop challenge.

The primary goal of this study is to develop new methods that leverage this probabilistic information and assess their performance. Furthermore, this work introduces a novel causal feature selection (CFS) algorithm using this approach and the establishment of a new evaluation criterion for CFS. To further enhance experimental performance, this dissertation proposes the use of a Graph Neural Networks (GNNs)–based probabilistic predictive framework for causal discovery.

Conventional causal discovery algorithms face significant challenges in dealing with large-scale observational datasets and capturing global structural information. The GNN-based approach addresses these limitations, enabling the learning of complex causal structures directly from data augmented with statistical and information-theoretic measures. The proposed framework represents a significant leap forward in causal discovery, offering improved accuracy and scalability in both synthetic and real-world datasets, as well as introducing a novel synergy between probabilistic learning and causal graph analysis.

In addition to the methodological advancements, this dissertation includes an application of counterfactual analysis to study affective polarization on social media. By comparing scenarios with and without specific influencer-led conversations on platforms like Twitter, I analyze the impact of these conversations on public sentiment. This application highlights the practical implications of the proposed causal modeling techniques, demonstrating their utility in understanding real-world issues and contributing to the broader field of social media analysis.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to the following individuals who have continuously supported me throughout my academic journey, culminating in the completion of this dissertation proposal:

First and foremost, I am deeply grateful to my wife, whose unconditional love, patience, and support have been a constant source of inspiration and comfort throughout my journey. She has been my steadfast pillar of strength. I also extend my heartfelt appreciation to my mother and sisters, whose unwavering love, encouragement, and belief in me have fueled my motivation and resilience.

I am grateful to my advisor, Dr. Gabriel Terejanu, for his continuous guidance, patience, and support throughout my research. His insights and feedback are invaluable touch-stones that are shaping this dissertation.

I am also grateful to my dissertation committee members, Dr. Siddharth Krishnan, Dr. Razvan Bunescu, Dr. Minwoo Lee, Dr. Shannon Reid, and Dr. Anthony Fodor, for their time, expertise, and valuable feedback. Their suggestions and insights have improved as well as will continue to improve the quality of this dissertation.

Furthermore, I thank my lab mates, Jawad Chowdhury and Ouldouz Ghorbani for their support and camaraderie throughout my research. Their input and assistance are invaluable in helping me towards the completion of this dissertation.

Lastly, I would like to thank the Army Research Office for their generous support under Grant Number W911NF-22-1-0035. I am also grateful to Lowe's Innovation Fund for their financial support. The funding from both has been instrumental in enabling the research and development that contributed to this dissertation.

To all of the above, I extend my deepest gratitude and appreciation. Thank you for your belief and support.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiii

CHAPTER 1: INTRODUCTION 1

1.1. Background 2

1.2. Proposed Approach and Contribution 4

1.3. Potential Significance and Outline of Dissertation 5

CHAPTER 2: FROM CAUSAL PAIRS TO CAUSAL GRAPHS 7

2.1. Introduction 7

2.2. Methodology 9

2.2.1. Developing causal-pair models 11

2.2.2. Enforcing DAGness 11

2.3. Experiments 14

2.3.1. Prior Work Used in Numerical Results 14

2.3.2. Data Sets 16

2.3.3. Metrics 17

2.3.4. Simulation 19

2.3.5. Real-World Data 21

2.4. Summary 23

CHAPTER 3: CAUSAL FEATURE SELECTION USING DIRECTED ACYCLIC GRAPHS 25

3.1. Introduction 25

3.2. Methodology 27

3.2.1. Causal Feature Selection (CFS) 29

3.2.2. CFS Evaluation Methodology 31

3.3. Prior Work used in Numerical Results 34

3.3.1. Maximum Likelihood Directed Acyclic Graph (MLDAG) 34

3.3.2. Linear Non-Gaussian Acyclic Model (LiNGAM) 35

3.3.3. Cause-Effect Pair Model 36

3.3.4. Maximum Relevance - Minimum Redundancy (mRMR) 36

3.3.5. Information Gain 37

3.3.6. Lasso Regularization 37

3.4. Experiments 38

3.4.1. Datasets 38

3.4.2. Evaluation Metrics 40

3.4.3. Setup and Simulation 41

3.4.4. Results 42

3.5. Summary 45

CHAPTER 4: A GRAPH NEURAL NETWORK-BASED PROBABILISTIC FRAMEWORK FOR CAUSAL DISCOVERY 47

4.1. Introduction 47

4.2. Related Work 49

4.3. Methodology 51

4.3.1. Feature Engineering and Graph Construction 52

4.3.2. Developing the Graph Neural Network (GNN) Model 54

4.3.3. Probabilistic Inference 56

4.4. Experiments 58

4.4.1. Datasets 58

4.4.2. Metrics 59

4.4.3. Results 59

4.5. Summary 66

CHAPTER 5: CAUSAL MODELING OF SOCIAL MEDIA POLARIZATION: QUANTIFYING INFLUENCER EFFECTS ON AFFECTIVE POLARIZATION 67

5.1. Introduction 67

5.2. Related Work 68

5.3. Methodology 70

5.3.1. Data 71

5.3.2. Quantifying Polarization with E/I Index 72

5.3.3. User Stance Labeling using Graph Neural Networks (GNNs) 75

5.3.4. Approach to Quantifying Influencer Impact on Affective Polarization 76

5.4. Results 77

5.4.1. Effect of an Influential-led Conversation 78

5.4.2. Insight of Influential Users 81

5.4.3. Insight of Conversations Within Stance-Group 82

5.5. Summary 83

CHAPTER 6: CONCLUSIONS 85

6.1. Future Work 86

REFERENCES 88

APPENDIX A: Implementation of Methods Used in Numerical Results 98

APPENDIX B: Sachs Protein Network Dataset 99

LIST OF TABLES

TABLE 2.1: Results of edge probability model applied on 80 Erdos-Renyi (ER) graph structures 19

TABLE 2.2: Results of edge probability model applied on 80 Scale-Free (SF) graph structures 19

TABLE 2.3: Comparison of my probabilistic methods applied on protein network dataset (with scaling) 21

TABLE 2.4: Comparison of my probabilistic methods applied on protein network dataset (without scaling) 22

TABLE 3.1: A description of the real-world data sets 40

TABLE 3.2: Comparison of my causal feature selection method with other methods applied to synthetic data 43

TABLE 3.3: Comparison of my causal feature selection method with other methods applied to real-world data 43

TABLE 3.4: Prediction performance of different methods using several machine learning models applied to classification data sets 44

TABLE 3.5: Prediction performance of different methods using several machine learning models applied to regression data sets 44

TABLE 4.1: Results of GNN-based edge probability framework applied on 80 Scale-Free and 80 Erdos-Renyi (ER) graph structures 60

TABLE 4.2: Results of GNN-based edge probability framework applied on Microsoft CSuite datasets 62

TABLE 4.3: Comparison of my probabilistic methods applied on protein network dataset (with and without scaling) 65

TABLE 5.1: Dataset details 71

TABLE 5.2: Top 5 influential users for each dataset 74

TABLE 5.3: Climate Change Dataset: The effect of a single conversation by influential users on daily polarization score. 81

TABLE 5.4: Gun Control: The effect of a single conversation by influential users on daily polarization score. 81

TABLE 5.5: Gun Control Dataset: Impact of conversation initialized by influential authors (Twitter users) on polarization scores 82

TABLE 5.6: Climate Change Dataset: Impact of conversation initialized by influential authors (Twitter users) on polarization scores 82

TABLE 5.7: Details of conversations' effects based on stances for both gun control and climate change datasets 83

LIST OF FIGURES

FIGURE 1.1: Causality map 2

FIGURE 3.1: Example of a causal network. 28

FIGURE 3.2: Knee/elbow detection using the Kneedle algorithm 31

FIGURE 3.3: Causal feature selection (CFS) method 33

FIGURE 4.1: Illustration of GNN-based edge probability framework 52

FIGURE 4.2: Illustration of results for GNN-based edge probability framework applied on 80 Scale-Free and 80 Erdos-Renyi (ER) graph structures 61

FIGURE 4.3: Performance comparison between GNN-based method and CausalPairs method on smaller CSuite datasets 64

FIGURE 5.1: Interaction networks with and without a specific conversation. 71

FIGURE 5.2: Bar chart distributions of conversation characteristics: (a) Gun Control and (b) Climate Change topics, displaying the frequency of conversations by number of tweets (left) and number of users (right). 73

FIGURE 5.3: Frequency distributions of daily conversation counts over the specified time frame, stratified by topic: (a) Gun Control and (b) Climate Change. 79

FIGURE 5.4: Temporal dynamics of polarization scores in response to events 80

LIST OF ABBREVIATIONS

ACE	Average Causal Effect
ATE	Average Treatment Effect
AUC	Area Under the Curve
BERTweet	Bidirectional Encoder Representations from Transformers for Twitter
CFS	Causal Feature Selection
CNN	Convolutional Neural Network
CSuite	Causal Suite
DAG	Directed Acyclic Graph
E/I Index	External/Internal Index
ER	Erdos-Renyi
FCQ	F Statistic to target over Correlation between features
FDR	False Discovery Rate
FPR	False Positive Rate
GAT	Graph Attention Networks
GBDT	Gradient Boosting Decision Tree
GCN	Graph Convolutional Networks
GES	Greedy Equivalence Search
GNN	Graph Neural Network

HSIC	Hilbert Schmidt Independence Criterion
i.i.d.	independent and identically distributed
ICA	Independent Component Analysis
IGCI	Information Geometric Causal Inference
IPCC	Intergovernmental Panel on Climate Change
KNN	K-Nearest Neighbours
LiNGAM	Linear Non-Gaussian Acyclic Model
LUCAS	Lung Cancer Simple Set
MB	Markov Blanket
MLDAG	Maximum Likelihood Directed Acyclic Graph
MLG	Maximum Likelihood Digraph
mRMR	Maximum Relevance - Minimum Redundancy
MSDAG	Maximum Spanning Directed Acyclic Graph
MST	Minimum Spanning Tree
NLP	Natural Language Processing
NOTEARS	No Tearing DAG Learning Algorithm
PC	Peter-Clark
PCA	Principle Component Analysis
PDAG	Probability Directed Acyclic Graph

PG	Probability Digraph
RBO	Rank Biased Overlap
RDD	Regression Discontinuity Design
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
RoBERTa	Robustly optimized BERT approach
SEM	Structural Equation Modeling
SF	Scale-Free
SHD	Structural Hamming Distance
TCE	Total Causal Effect
TPR	True Positive Rate
VADER	Valence Aware Dictionary and sEntiment Reasoner

CHAPTER 1: INTRODUCTION

The concept of causality, which involves causal relationships between variables, is fundamental in multiple fields such as medicine, economics, and social sciences. It pertains to the relationship between cause and effect, where one variable impacts the outcome of another variable [1–3]. Understanding these relationships is crucial for making informed decisions and accurate predictions. Causality is a vast field of study that consists of various subfields and branches. It is typically divided into two main domains: causal inference and causal discovery, as depicted in Figure 1.1.

Causal inference is concerned with comprehending the effects of actions taken. It provides tools that enable the isolation and calculation of the impact of a change within a system, even if the change did not occur in practice. Causal inference can address various types of queries, including identifying whether taking a specific medication leads to an improvement in an individual's health, determining how much advertising spend is needed to achieve specific revenue targets, and assessing the effect of classroom size on educational attainment.

Conversely, causal discovery is a discipline within causal inference that endeavors to infer the causal topology of a system from observational data, striving to discern the causal connections and directional influences that shape the data. Unlike statistical correlations, causal discovery methods reveal relationships that represent the fundamental basis for understanding a system and are invariant to change. Therefore, the primary focus of this dissertation is to explore in-depth causal structure learning, which is a key aspect of causal discovery through the use of causal graphs.

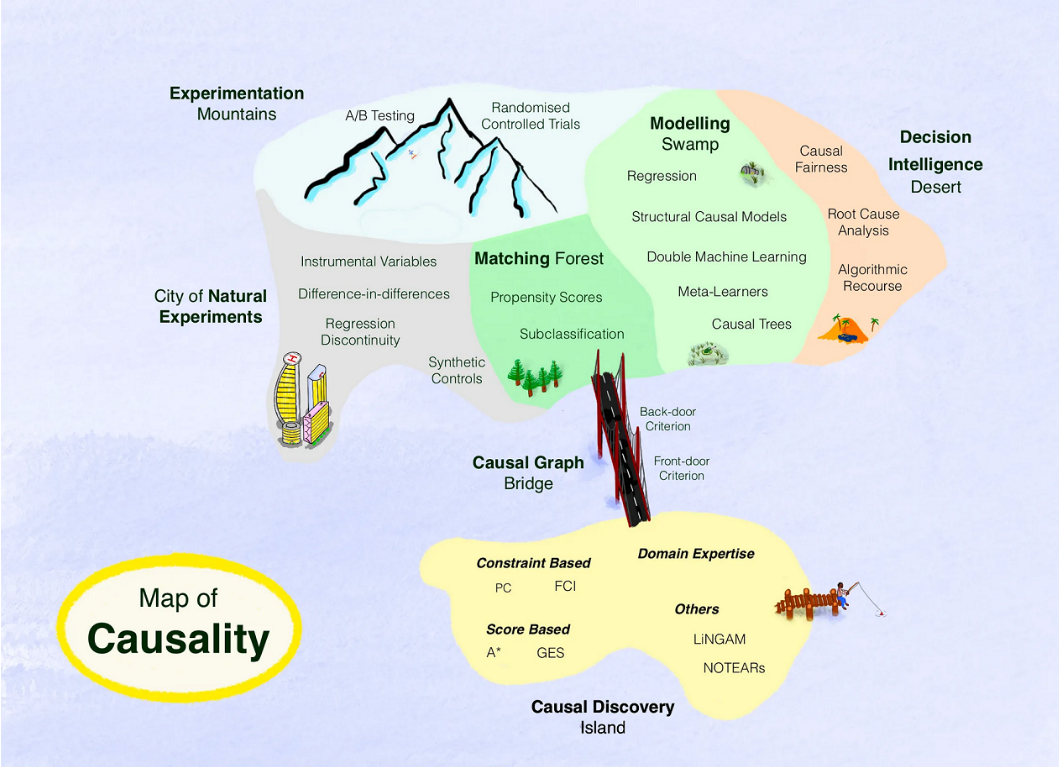


Figure 1.1: The world of causality. Image used from Ferrand (2023) [4].

1.1 Background

Causal Graph Discovery

Causal graph discovery identifies causal relationships among variables in a complex system using directed acyclic graphs (DAGs) to represent these relationships. This technique relies on observational data and statistical methods rather than experimental manipulation [5–9]. For instance, in investigating the causal relationship between physical exercise and mental health, researchers analyze data on various factors like exercise frequency and mental health status to uncover causal links.

One approach to causal graph estimation involves cause-effect pairs, which test whether one variable causes another, especially when experimental manipulation is impractical [10–12]. Despite its potential, causal graph discovery faces challenges like high-dimensional data, computational complexity, and confounding variables [5,13]. Hybrid methods that combine constraint-based and score-based approaches attempt to mitigate these issues but still rely on local heuristics without a standard way of selecting score functions and search strategies [14,15].

Causal Feature Selection

Causal feature selection addresses some challenges of causal graph discovery by identifying a subset of features with a causal effect on the outcome variable, reducing data dimensionality, and eliminating confounding variables [16–18]. This technique employs a causal score to evaluate and select features that demonstrate the most robust causal associations with the outcome variable, enabling the identification of key drivers and causal factors. On the contrary, traditional feature selection methods often overlook causal relationships, leading to suboptimal results [19]. Effective evaluation of causal feature selection requires a reliable criterion that accurately measures prediction accuracy.

Graph Neural Networks for Causal Discovery

Graph Neural Networks (GNNs), encompassing architectures such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), have demonstrated exceptional proficiency in learning from data represented as graphs, effectively capturing complex relational patterns and structural information [20–23]. Despite their success, GNNs' application in causal discovery is limited due to challenges like enforcing the acyclic nature of causal graphs, as traditional GNNs handle general graph structures without this constraint [24,25]. DAG-GNN addresses this by using a variational autoencoder to learn DAGs, but it focuses on deterministic structures and misses causal relationship uncertainties [26]. Another challenge is integrating local and global information in the causal graph [3,26,27]. While GNNs capture local interactions through node-to-node message passing, they often struggle with global structural information. This dissertation develops a probabilistic framework incorporating both node and edge features to address these challenges.

Application: Affective Polarization on Social Media

Affective polarization, characterized by increasing emotional divide and animosity between opposing political groups, is amplified by social media platforms like Twitter [28–31]. Influencers on these platforms significantly impact public sentiment and contribute to polarization through their large followings and frequent interactions [32]. This dissertation employs counterfactual analysis to compare scenarios with and without influencer-led conversations, providing insights into influencers' impact on social media dynamics. This analysis highlights the practical implications of advanced analytical techniques in studying complex social phenomena and their utility in real-world contexts.

1.2 Proposed Approach and Contribution

The challenges in causal discovery and causal feature selection have motivated us to study several research questions to address the causal graph learning problem. The contributions are multi-faceted: a novel probabilistic approach is proposed to uncover causal structures using cause-effect pairs; a new method for causal feature selection is introduced, leveraging causal graphs and causal metrics to enhance model accuracy and reliability; a Graph Neural Network (GNN)-based framework is developed to incorporate global information and overcome the limitations of causal-pairs methods; and counterfactual analysis is applied to social media data to provide insights into affective polarization. The objectives of this study are to investigate and provide insights into the following research questions:

1. How can predictions of cause-effect pairs be utilized to efficiently and accurately discover the causal graph of a set of variables?

2. Given a high-dimensional observational dataset, how can a subset of variables that are causally related to the target variable be selected?

3. What is an appropriate evaluation framework for causal feature selection algorithms?

4. How can Graph Neural Networks (GNNs) be leveraged to enhance causal discovery by incorporating global information beyond local causal pairs?

5. What is the impact of influencer-led conversations on public sentiment, and how can counterfactual analysis provide insights into affective polarization on social media?

The first two chapters, addressing research questions 1-3, have already been published as two papers in peer-reviewed conferences, ICMLA 2022 [33] and DSAA 2023 [34]. Furthermore, preliminary findings related to question 4 have been published in an extended abstract in the 2024 UAI Causality workshop [35], and a final manuscript is also ready for submission. Furthermore, a manuscript addressing question 5 has been submitted to another peer-reviewed conference for review.

1.3 Potential Significance and Outline of Dissertation

The proposed solutions have significant potential in various fields, including healthcare, economics, and social sciences. They can identify causal relationships for decision-making and improve machine learning algorithms by providing a causal framework, leading to more accurate and interpretable models. The causal feature selection method aids in dimensionality reduction and identifies influential features, offering computationally faster solutions. The GNN-based framework for causal discovery extends the applicability and scalability of causal inference methods, benefiting fields dealing with complex and high-dimensional data, such as genomics, environmental science, and social network analysis. The counterfactual analysis of affective polarization on social media provides insights into the impact of influencer-led conversations on public sentiment, highlighting the practical implications of advanced analytical techniques.

Chapter 2 details the proposed approach to causal graph estimation using cause-effect pairs, covering datasets, algorithms, and evaluation metrics. This approach generates a probability distribution over all possible graphs based on cause-effect pair features.

Chapter 3 describes the methodology for causal feature selection, combining causal graph discovery and causal metrics to identify features with a causal effect on the outcome variable. The performance is evaluated on synthetic and real-world datasets compared with other benchmark methods.

Chapter 4 introduces a GNN-based probabilistic predictive framework for causal discovery, refining the probability distribution from the causal-pairs approach. This framework is evaluated on various synthetic, benchmark, and real-world datasets.

Chapter 5 explores the application of the concept of counterfactual analysis to study affective polarization on social media. By comparing scenarios with and without specific influencer-led conversations on Twitter, the study highlights the practical implications of advanced analytical techniques.

Finally, Chapter 6 presents the main conclusions and contributions to machine learning and causal discovery. The dissertation discusses potential future work, including enhancing the proposed methods and exploring their applications in diverse fields.

CHAPTER 2: FROM CAUSAL PAIRS TO CAUSAL GRAPHS

2.1 Introduction

Machine learning methods, deep learning in particular, have achieved unparalleled predictive performance in the past two decades. Nevertheless, these correlation-based models exhibit significant limitations when applied to out-of-distribution data and prescriptive analytics, which is grounded in causal inference. Learning the underlying causal structure is an important task to both constrain a model by reducing spurious correlations [2] and perform What-If analysis [1]. Learning causal relationships from observational data, also known as causal discovery, remains an active and challenging research topic [1, 3, 5].

Several causal discovery methods have been proposed in the literature. Constraint-based approaches learn the causal skeleton using conditional independence test using the joint probability distribution of the data and identify edge directions up to their Markov equivalence class [5–9].

Score-based approaches learn the causal graph $\mathcal{G}$ by optimizing a score function generally computed with respect to observational data [13, 36, 37]. Unfortunately, these methods suffer from super-exponential computational complexity in the number of nodes. Tsamardinos et al. [14] propose a hybrid method where they use a constraint-based approach to reduce the search space in score-based methods. However, this method relies on local heuristics and lacks a standard way of choosing score functions and search strategies [15].

A promising direction, NOTEARS [38], formulates a smooth characterization of acyclicity that can be incorporated into a continuous optimization and solved using well-known numerical methods. NOTEARS was later extended to parametric nonlinear models and nonparametric models [39]. GOLEM [40] also adopts a continuous optimization framework, however, it makes use of a linear DAG learning model and doesn't capture non-linear relationships. A different approach looks at identifying cause-effect pairs using the statistical techniques from observational data [41, 42]. Singh et al. [11] use deep convolutional neural network (CNN) models to determine the directions of pairwise causal edges from observational data. Hassanzadeh et al. [12] formulate the pairwise causal discovery techniques as binary causal problems where they try to answer if there exist any causal relations between two variables in the context of Natural Language Processing (NLP). Nevertheless, they have not studied how the predicted edge directions can be used to provide a solution to causal graph identification.

Motivated by applications in biological networks, Medvedovsky et al. [10] propose an approximation algorithm to orient a graph by maximizing the number of pairs that admit a directed path from known pairs of sources and targets. Nevertheless, given the peculiarities of the application based on prior knowledge constrains, this approach falls short from identifying the entire causal graph. Therefore, discovering the causal graph, including pairwise causal relations from observational data remains a challenging task due to various factors such as finite sampling and measurement errors. In general, identifying cause-effect relationships requires controlled experimentation which is expensive and/or technically and ethically impossible to perform [43].

I propose a probabilistic approach to discover causal structures using the cause-effect pairs features proposed in response to the 'Cause-Effect Pair' at the NIPS 2013 Workshop on Causality challenge. This chapter introduces the following novel contributions: (1) generate a probability distribution over all the edges of a digraph using various statistical and information-theoretic features that describe the relationships between any two variables in the dataset; (2) generate the most likely probability distribution of directed acyclic graphs (DAG) using the maximum spanning DAG; (3) generate an approximate solution to the causal graph problem by estimating the digraph and DAG using maximum likelihood estimate with the probability distributions in (1) and (2) respectively; (4) overall, my proposed methods are comparable with traditional ones (PC, GES), while benefiting from polynomial time complexity as compared to super-exponential time complexity; and (5) finally, by comparing with state-of-the-art methods such as NOTEARS-MLP, I show that future improvements are possible by further leveraging global graph information.

Section 2.2 introduces the problem formulation and details of my causal discovery approach based on causal-effect pairs. The empirical evaluation of my methods is described in Section 2.3. Lastly, in Section 2.4 I present my findings in brief and opportunities for future improvements.

2.2 Methodology

Given n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_d] \in \mathbb{R}^{n \times d}$, the goal of causal discovery is to estimate the underlying causal relations encoded by the directed acyclic graph (DAG), $\mathcal{G}_{\text{DAG}} = (V, E)$. V comprises of nodes corresponding to the observed random variables X_i for $i = 1 \dots d$ and the edges in E correspond to the causal relations encoded by $\mathcal{G}_{\text{DAG}}$. Namely, the existence of the edge $i \to j$ corresponds to a direct causal relationship between X_i (the cause) and X_j (the effect).

The approach is to leverage the work on cause-effect pairs which uses a classifier model to predict the probability distribution $p(y_{ij}|f)$ of causal relation between two variables X_i and X_j given the observational dataset $[\mathbf{x}_i, \mathbf{x}_j] \in \mathbb{R}^{n \times 2}$.

$$p(y_{ij}|f) \quad = \quad f([\mathbf{x}_i, \mathbf{x}_j]), \text{ for } i < j \tag{2.1}$$

Here, I assume that $f(\cdot)$ is a pre-trained machine learning model and $y_{ij} \in [-1, 0, 1]$.

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

After calculating the probability distributions of causal relations between all the pairs in the dataset, a naive approach to construct the probability distribution of a digraph $\mathcal{G}$ is to assume that the causal-pairs are independent. In Section 2.2.2, I show that my proposed approach of enforcing DAGness does correlate these causal-pairs and provides us with additional global information to constrain the graph probability distribution and allows us to move beyond the initial edge independence.

$$p(\mathcal{G}|f) = \prod_{i<j} p(y_{ij}|f) \tag{2.2}$$

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

$$\mathcal{G}_{\mathrm{ML}} = \arg\max_{\mathcal{G}} p(\mathcal{G}|f) \tag{2.3}$$

Note that the samples from the probability distribution, Eq. 2.2, and the maximum likelihood estimate, Eq. 2.3, are digraphs with no guarantees that they are acyclic. In the following, I propose to generate the most likely probability distribution of directed acyclic graphs (DAG) using the maximum spanning DAG approach [44] as well as estimate a representative DAG using the maximum likelihood estimate.

2.2.1 Developing causal-pair models

The model $f(\cdot)$ in Eq. 4.1 can be trained using synthetic datasets or real datasets with known causal relations. Given a set of labeled datasets $\{([\mathbf{x}_i, \mathbf{x}_j], y_{ij})_k\}$, in this study I take the approach of engineering features from this dataset using various statistical and information-theoretic measures such as: minimum or maximum value of a variable; number of unique samples of a variable; entropy, mutual information, uniform divergence; slope-based information geometric causal inference (IGCI), Hilbert Schmidt independence criterion (HSIC); Pearson R coefficient; Spearman's rank coefficient; moments and mixed moments such as skewness and kurtosis. Therefore, the machine learning model $f(\cdot)$ can be trained on the new engineered dataset with features previously introduced. Note that my proposed methodology is agnostic to the features deployed and it works with any causal-pairs model. Also, the computational runtime of calculating any of these features such as HSIC, Pearson R coefficient has no effect asymptotically on the computational complexity - it just increases the constant of the polynomial runtime. I also note that additional improvements might be brought by developing deep neural network architectures capable to extract informative representations for predicting the target $y_{ij} \in [-1, 0, 1]$ directly from the sample dataset, but it is left as future work.

2.2.2 Enforcing DAGness

In this section, I propose to derive a probability distribution that guarantees that the sample graphs are DAGs. This takes the form of the probability distribution in Eq. 2.4 which unlike Eq. 2.2 contains the DAGness condition.

$$p(\mathcal{G}|f, \mathrm{DAG}) = \sum_{\pi} p(\mathcal{G}|f, \mathrm{DAG}, \pi)p(\pi|f) \tag{2.4}$$

Due to computational intractability, I have chosen to build this conditional distribution not using the Bayes rule and utilizing Eq. 2.2 as prior, but rather as the law of total probability where I integrate out the topological ordering π of the vertices. I note that for a topological ordering where node i comes before node j, it implies that a directed edge can only happen from i to j. I also note that the possibility of no causal relation between i and j is not excluded in this context. Both causal, $i \rightarrow j$, and noncausal, $i \not\rightarrow j$ and $j \not\rightarrow i$, are possible.

To generate a representative DAG one can use the maximum likelihood estimate. This however is intractable, and I do assume that there is a topological sorting of vertices that also covers the maximum likelihood DAG.

$$\begin{aligned} \mathcal{G}_{\text{DAG}} &= \arg\max_{\mathcal{G}} p(\mathcal{G}|f, \text{DAG}) && (2.5) \\ &\approx \arg\max_{\mathcal{G}} p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) && (2.6) \end{aligned}$$

I propose to approximate the topological ordering, π_{ML}, by the topological sorting of the Maximum Spanning DAG (MSDAG) [44] of the induced weighted graph by the probability of causal relations.

$$\begin{aligned} \pi_{\text{ML}} &= \arg\max_{\pi} p(\pi|f) && (2.7) \\ &\approx \text{toposort}(\text{MSDAG}(\mathcal{G}_A)) && (2.8) \end{aligned}$$

I build the following weighted adjacency matrix $A \in \mathcal{R}^{d \times d}$, which contains the probability of all directed edges as weights.

$$\begin{aligned} A[i,j] &= p(y_{ij} = 1|f) \\ A[j,i] &= p(y_{ij} = -1|f) \end{aligned} \qquad (2.9)$$

Let $\mathcal{G}_A$ be a weighted graph induced by the adjacency matrix A. The goal is to find the topological sorting of the MSDAG of $\mathcal{G}_A$. The motivation is to accommodate as many directed edges with large probabilities as possible. I use the approach introduced by [45] to approximate the MSDAG by first constructing the maximum spanning tree and greedily adding edges in the descending order of the weights as long as no cycles are formed. Note that a topological sorting derived this way still accommodates the possibility of no edges to account for their probability in the maximum likelihood DAG, Eq. 2.6.

$$p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) = \prod_{\pi_{\text{ML}}^{-1}[i] < \pi_{\text{ML}}^{-1}[j]} p(y_{i\to j}|f) \tag{2.10}$$

Given the maximum likelihood of topological ordering, one can easily calculate the probability distribution in Eq. 2.6 by constraining the direction of the edge based on the node ordering, see Eq. 2.10. In this context, I are left with only two possibilities when node i appears before j in π_{ML}. Either there is an edge from i to j or there is no edge between them.

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

By constraining the direction of edges I need to re-normalizing the edge probabilities as follows.

$$p(y_{i\to j} = 1|f) = \frac{p(y_{ij} = 1|f)}{p(y_{ij} = 1|f) + p(y_{ij} = 0|f)} \tag{2.11}$$

Enforcing DAGness in Eq. 2.10 and consequently Eq. 2.6 provides us with additional global information to constrain the graph probability distribution and allows us to move beyond the initial edge independence in Eq. 2.2 and Eq. 2.3 respectively, which was derived from just pair-wise (local) information.

2.3 Experiments

I show the empirical results of my approaches applied to both synthetic and real-world datasets. I have used the following labels for my approaches: PG given by Eq. 2.2, MLG given by Eq. 2.3, PDAG given by Eq. 2.10, and MLDAG given by Eq. 2.6.

2.3.1 Prior Work Used in Numerical Results

I compare my methods' performance with two traditional approaches: the PC algorithm [5] and the GES algorithm [13] and with a state-of-the-art approach, NOTEARS-MLP [39].

PC Algorithm [5]. The PC (Peter and Clark) algorithm is based on the concept of conditional independence. It takes a dataset consisting of variables and their corresponding values as input such as the data matrix, $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_d] \in \mathbb{R}^{n \times d}$ given n i.i.d. observations. The algorithm then analyzes the statistical relationships between variables to identify causal connections. The PC algorithm generates a causal graph (DAG), that visually encapsulates the inferred causal connections and directional dependencies between variables, providing a graphical representation of the underlying causal structure.

GES Algorithm [13]. The GES (Greedy Equivalence Search) algorithm tests conditional independence and incorporates a search strategy to explore different causal structures. It aims to find the most likely causal graph that fits the observed data matrix, $\mathbf{X}$. The algorithm uses a score-based approach to evaluate and compare different causal graph structures. These score-based metrics penalize complex models and favor simpler explanations that fit the data well. The output of the GES algorithm is also a causal graph or DAG that captures causal connections.

It is important to note that both the PC and GES algorithms are capable of outputting a Partially Directed Acyclic Graph (PDAG) alongside a causal graph or DAG. In these outputs, directed edges indicate causal relationships, while undirected edges signify conditional independence relationships among the variables.

NOTEARS [38]. The NOTEARS algorithm offers a promising approach for inferring the causal structure of variables from observational data. It is specifically designed with the assumption that the causal relationships within the data are acyclic, ensuring the absence of cycles in the causal graph. To capture both linear and nonlinear causal dependencies, NOTEARS incorporates a nonlinear transformation of the data. Furthermore, the algorithm promotes sparsity in the estimated causal graph by imposing a constraint that encourages a sparse representation of significant causal connections. By employing an optimization algorithm, NOTEARS estimates the causal structure by optimizing an objective function that strikes a balance between data fit, sparsity, and acyclicity. These two properties, sparsity, and acyclicity, are essential characteristics of causal graphs that the objective function aims to capture. Therefore, given the observational data matrix, **X**, the objective function used by NOTEARS is defined as follows:

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times d}} F(\mathbf{W}) = \frac{1}{2n} \| \mathbf{X} - \mathbf{X}\mathbf{W} \|_F^2 + \lambda \| \mathbf{W} \|_1 \tag{2.12}$$

where,

- **W** is the causal graph adjacency matrix
- $\| \cdot \|_F^2$ denotes the squared Frobenius norm
- λ is a regularization parameter that controls the sparsity of the graph by using smaller edge weights
- $\| \cdot \|_1$ denotes the l_1 norm, which serves as a penalty term

The objective function is composed of two primary terms: the first evaluates the model's ability to replicate the observed data, driving the learning of a graph structure that accurately represents the conditional dependencies; the second term introduces a penalty for complex graph structures, promoting sparsity by discouraging large edge weights and unnecessary edges.

In this study, I utilize the NOTEARS-MLP [39] variant of the algorithm where the input is the same observational data matrix, $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_d] \in \mathbb{R}^{n \times d}$ that I use for my method to derive causal pairs features and the output is a estimated causal structure represented by a DAG or a weighted adjacency matrix. However, Reisach et. al. [46] in their paper highlight that continuous score-based approaches i.e. NOTEARS-MLP [39] in particular suffer highly from data scaling which was addressed from a theoretical perspective. Therefore, I standardize the features for both the synthetic and real-world datasets by removing the mean and scaling to unit variance.

2.3.2 Data Sets

Cause-Effect Pair Train Data. To train my model, I have used the Cause-effect pairs dataset[1] from the NIPS 2013 Workshop on Causality. The train data is a set of labeled datasets $\{([\mathbf{x}_i, \mathbf{x}_j], y_{ij})_k\}$ where $[\mathbf{x}_i, \mathbf{x}_j] \in \mathbb{R}^{n \times 2}$ and $k = 4050$ with known causal relationships. Namely, this is a set of variable pairs (variable X_i and variable X_j) with known ground truth such that label, $y_{ij} \in [-1, 0, 1]$ where -1: causal relation exists from X_j to X_i, 0: no causal relationship exists between X_i and X_j, and 1: causal relation exists from X_i to X_j. I trained the model $f(\cdot)$ in Eq. 4.1 on this set of labeled causal pairs using the engineering features to calculate the probability distribution of causal relations between all the pairs in the testing datasets. In addition, these known ground truths are derived from expert domains such as chemistry, ecology, engineering, medicine, physics, sociology, etc.,

[1] Cause-Effect Pair Dataset: https://www.kaggle.com/c/cause-effect-pairs/data

and these pairs are intermixed with controls such as pairs of independent variables and pairs of dependent variables but not causally related.

Synthetic Test Data. To evaluate the performance of my methods on causal graph estimation, I have generated synthetic data for testing. Synthetic graphs provide a benchmark for evaluating the performance of causal discovery algorithms. In addition, synthetic graphs offer control over the complexity of the causal relationships where I can design the graphs with varying degrees of complexity, including the number of nodes, edge density, and type of causal relationships (e.g., direct, indirect).

In this study, I have considered 16 types of different graph combinations having similar criteria: number of nodes, $d = [10, 20]$, number of edges, $e = [1d, 4d]$, number of data samples per node, $n = [200, 1000]$, and graph models from Erdos-Renyi(ER) and Scale-Free (SF). I have generated non-linear data samples for the graph nodes similar to data generation utilities available in the NOTEARS-MLP implementation. In addition, for each of these 16 graph types, I have generated 10 random graph structures with ground truths to test my methods. The outputs are then summarized over these 10 graph structures to report my results for all 16 graph combinations.

2.3.3 Metrics

I consider three performance metrics to evaluate the causal graphs: True Positive Rate (TPR), False Positive Rate (FPR) and Structural Hamming Distance (SHD). A lower SHD and FPR indicate a better performance whereas a higher TPR is better. However, since both PC and GES may generate outputs with undirected edges, I treated an undirected edge as a true edge with a probability of 0.5 and a false edge with the same probability. SHD, TPR, and FPR were implemented by their definition for PC, GES, NOTEARS-MLP, and my two maximum likelihood estimates (MLG given by Eq. 2.3 and MLDAG given by Eq. 2.6).

As for my probabilistic approaches (PG given by Eq. 2.2 and PDAG given by Eq. 2.10), given a true graph $\mathcal{G}_{true}$ and an adjacency matrix $\mathcal{A}$ with edge probabilities, I calculate SHD using Eq. 2.13, TPR using Eq. 2.15 and FPR using Eq. 2.17.

$$\begin{aligned} SHD = & \sum_{\substack{i<j \\ (i,j)\in E(\mathcal{G}_{true}) \\ (j,i)\notin E(\mathcal{G}_{true})}} (1-\mathcal{A}[i,j]) + \sum_{\substack{i<j \\ (i,j)\notin E(\mathcal{G}_{true}) \\ (j,i)\in E(\mathcal{G}_{true})}} (1-\mathcal{A}[j,i]) \\ & + \sum_{\substack{i<j \\ (i,j)\notin E(\mathcal{G}_{true}) \\ (j,i)\notin E(\mathcal{G}_{true})}} (\mathcal{A}[i,j]+\mathcal{A}[j,i]) \end{aligned} \tag{2.13}$$

$$TP = \sum_{\substack{i<j \\ (i,j)\in E(\mathcal{G}_{true}) \\ (j,i)\notin E(\mathcal{G}_{true})}} \mathcal{A}[i,j] + \sum_{\substack{i<j \\ (i,j)\notin E(\mathcal{G}_{true}) \\ (j,i)\in E(\mathcal{G}_{true})}} \mathcal{A}[j,i] \tag{2.14}$$

$$TPR = \frac{TP}{max(|E(\mathcal{G}_{true})|,1)} \tag{2.15}$$

$$\begin{aligned} FP = & \sum_{\substack{i<j \\ (i,j)\in E(\mathcal{G}_{true}) \\ (j,i)\notin E(\mathcal{G}_{true})}} \mathcal{A}[j,i] + \sum_{\substack{i<j \\ (i,j)\notin E(\mathcal{G}_{true}) \\ (j,i)\in E(\mathcal{G}_{true})}} \mathcal{A}[i,j] \\ & + \sum_{\substack{i<j \\ (i,j)\notin E(\mathcal{G}_{true}) \\ (j,i)\notin E(\mathcal{G}_{true})}} (\mathcal{A}[i,j]+\mathcal{A}[j,i]) \end{aligned} \tag{2.16}$$

$$FPR = \frac{FP}{max((M-|E(\mathcal{G}_{true})|),1)} \tag{2.17}$$

Here, $M = \frac{d(d-1)}{2}$ is the number of possible edges of the graph $\mathcal{G}_{true}$ and d is the total number of nodes in the graph.

Note that since I am calculating these metrics over 160 different graph structures of various sizes in my test data, I report a normalized SHD over the number of graph nodes (SHD/d) as my SHD measure. TPR and FPR are normalized by definition.

Table 2.1: Edge probability model trained on cause-effect pairs data provided at the NIPS 2013 Workshop on Causality. The means and standard errors of the performance metrics are based on the 80 Erdos-Renyi (ER) graph structures in the test data.

Metrics	PG (Eq. 2.2)	MLG (Eq. 2.3)	PDAG (Eq. 2.10)	MLDAG (Eq. 2.6)	PC	GES	NOTEARS-MLP
SHD/d	2.38±0.14	2.32±0.17	2.30±0.15	2.18±0.16	2.40±0.21	1.78±0.13	1.33±0.10
TPR	0.39±0.02	0.15±0.02	0.38±0.02	0.28±0.02	0.17±0.02	0.48±0.02	0.58±0.02
FPR	0.72±0.10	0.07±0.01	0.61±0.09	0.29±0.05	0.22±0.04	0.87±0.15	0.32±0.06

Table 2.2: Edge probability model trained on cause-effect pairs data provided at the NIPS 2013 Workshop on Causality. The means and standard errors of the performance metrics are based on the 80 Scale-Free (SF) graph structures in the test data.

Metrics	PG (Eq. 2.2)	MLG (Eq. 2.3)	PDAG (Eq. 2.10)	MLDAG (Eq. 2.6)	PC	GES	NOTEARS-MLP
SHD/d	2.02±0.12	1.97±0.13	1.96±0.12	1.88±0.13	1.93±0.15	1.43±0.11	1.36±0.11
TPR	0.31±0.01	0.12±0.01	0.30±0.01	0.20±0.01	0.17±0.02	0.51±0.03	0.47±0.02
FPR	0.26±0.02	0.03±0.01	0.21±0.02	0.09±0.01	0.08±0.01	0.26±0.04	0.12±0.02

2.3.4 Simulation

For my implementation, I first extract features from the data-pairs using the feature extraction method of Team-Jarfo [42], the second winner from the NIPS 2013 Workshop on Causality challenge. I have also used the causal pairs model of the third winning team from the above-mentioned competition and I have found that it doesn't have better performance than Team-Jarfo [42] as expected. I couldnât run a performance analysis with the model proposed by the No.1 team due to the lack of availability of their implemented code. I have extracted 130 features from the pairs for all training and testing datasets using the code[2] of Team-Jarfo [42]. This feature set contains some standard statistical features as well as new measures based on variable measures of the conditional distribution. I train a multi-classifier model based on LightGBM [47] which is a Gradient Boosting Decision Tree (GBDT) algorithm developed by Microsoft. LightGBM speeds up the training

[2] Team-Jarfo Code: https://github.com/jarfo/cause-effect

process by using a histogram-based algorithm [48, 49] and combines weak learners into strong ones using an iterative approach [50] to optimize parallel learning. I create the LightGBM classifier from the Python library PyCaret [51] and select the hyper-parameters by tuning the model using the PyCaret 'tune_model()' function optimized over AUC.

I train my classifier model on cause-effect pairs data using the 130 extracted features and make predictions on the synthetic testing data. I predict a probability distribution over the three classes of edge directions (backward edge, no edge, forward edge) for all edges in the testing dataset. Finally, using the methods described in Section 2.2 I calculate the metrics of my causal graph estimation methods from the predicted probability distribution and compare the results with the benchmark approaches.

Table 2.1 and Table 2.2 show the empirical results of my methods applied to 80 different Erdos-Renyi graph structures and 80 different Scale-Free graphs, respectively. From these two tables, we see that estimating causal graphs is more challenging for Erdos-Renyi graph structures than Scale-Free graphs for all the methods. I also note that the performance of NOTEARS-MLP is superior to all other methods in terms of SHD and TPR in particular and it does have implications for further developing my proposed methods as detailed later.

Regardless of the graph structure, I make the following observations. (1) PG performs better than PC in terms of SHD and TPR and is better than GES in terms of FPR. (2) MLG does improve the FPR but at the cost of degrading the TPR. (3) I do see a significant improvement by enforcing DAGness. Namely, PDAG does improve FPR over PG, but not sufficient to be statistically better/similar to PC. (4) However, this does happen when I take the maximum likelihood of the conditional probability. Namely, MLDAG performs at the same FPR level as PC.

From these results, it becomes clear that using global information by constraining the graphs to be DAGs does improve the performance compared to using the pair-wise probabilities naively. However, when I compare it with NOTEARS-MLP, it also becomes clear that this is not sufficient and that the next iteration of methods needs to develop features that intrinsically exploit global information.

Nevertheless, one of the distinguishable advantages my methods have over PC and GES is that they not only perform statistically better/similar but also they have significantly low computational complexity. While both PC and GES have exponential time complexity due to their combinatorial approach, my methods run in polynomial time $\mathcal{O}(d^2)$ in the number of nodes d as they exploit local node pairs information.

2.3.5 Real-World Data

For real data, I consider the dataset published by [52], which is based on the expression level of proteins. This signaling network is largely used in the scientific community as a real application due to the consensus ground truths. It has 11 different protein cells represented as nodes d and the causal relationships were represented as directed edges ($e = 17$) between the nodes. This direction of the edge indicates the direction of influence, where the activity or behavior of one protein (cause) directly influences the activity or behavior of another protein (effect). Added, I aggregated the 9 different data files, resulting in a sample size $n = 7466$ in my experiment. The intensive details about this dataset are presented in the appendix.

Table 2.3: Comparison of my probabilistic methods with GES and NOTEARS-MLP that were applied on protein network dataset using cause-effect pairs as training data.

Metrics	PG (Eq. 2.2)	MLG (Eq. 2.3)	PDAG (Eq. 2.10)	MLDAG (Eq. 2.6)	GES	NOTEARS-MLP
Predicted Edges	36.14	9.82	33.16	18.48	34	42.23
Correct Edges	6.7	3.04	7.42	4.91	5.5	5.83
Reversed Edges	7.77	4.26	6.62	5.41	9.5	7.18

Table 2.4: Comparison of my probabilistic methods with GES and NOTEARS-MLP (results reported from the original manuscript [39]). These methods were applied to a non-standardized protein network dataset.

Metrics	PG (Eq. 2.2)	MLG (Eq. 2.3)	PDAG (Eq. 2.10)	MLDAG (Eq. 2.6)	GES	NOTEARS-MLP
Predicted Edges	38.01	10.41	34.81	20.60	34	13
Correct Edges	6.21	1.52	6.47	4.71	5.5	7
Reversed Edges	8.26	4.04	7.49	6.32	9.5	3

Unlike synthetic data sets, I have considered three different metrics for this protein dataset: the total number of predicted edges, the number of correct edge predictions, and the number of reversed edge predictions. Since this protein network uses a consensus over the number of true edges (17 known edges as ground truth), I do not know the actual true graph for the entire network. Therefore, a metric such as SHD becomes meaningless. Furthermore, since GES algorithms generate graphs that contain bidirectional edges, similarly to the synthetic results, I considered the bidirectional edge as increasing the number of corrected edges with 0.5 and the number of reversed edges with 0.5 as well.

In Table 2.3, I show the performance evaluation of my methods applied to the protein network dataset. I note that both PG and PDAG perform better than GES and NOTEARS-MLP in terms of predicting the number of correct edges. In addition, I find that taking the maximum likelihood estimates, MLG and MLDAG, significantly reduce the number of false edge predictions but their performances in predicting correct edges also degrade. The degradation is less severe between PDAG and MLDAG than PG and MLG.

PG, PDAG, as well as GES have the best performance in terms of the total number of predicted edges than NOTEARS-MLP, but MLDAG has the best total number of predicted edges which is close to the number of true edges in the dataset. I also note that MLG and MLDAG have better performance than NOTEARS-MLP in terms of fewer reversed edges. Finally, PGDAG shows an overall better performance than PG, further demonstrating the impact of enforcing DAGness.

Sensitivity to data scaling. Table 2.4 shows the results of my methods as well as GES and NOTEARS-MLP [39] applied on protein dataset before scaling them. We observe that my methods: PG, PDAG, MLG, MLDAG have almost similar results in both Table 2.3 (after data scaling) and Table 2.4 (before data scaling). GES is not affected by the data scaling. However, as expected, we see that NOTEARS-MLP has very different results in these two tables where it suffers highly from data-scaling, which is consistent with the sensitivity to scaling results in Ref. [46]. It is to mention that in Table 2.4, NOTEARS-MLP results are reported from the original manuscript [39], whereas in Table 2.3, I use the same implementation[??] of NOTEARS-MLP [39] with default parameters and applied on the standardized protein dataset by removing the mean and scaling to unit variance.

2.4 Summary

In this study, I have introduced a novel approach to causal discovery by leveraging the probabilistic information of pairwise causal edges. I have proposed to go beyond the naive approach to generate graph probabilities from causal pair probabilities by enforcing the graph to be acyclic and approximating its solution using the maximum spanning directed acyclic graph approach. Enforcing acyclicity clearly improves the performance on both synthetic and real datasets compared with the naive approach.

I have shown that my methods have statistically better and/or similar performances than some traditional methods. More importantly, this performance comes with just polynomial run-time as compared with the exponential run-time of traditional methods that are combinatorial in nature which presents a promising and feasible approach for approximating the solution to the NP-hard problem of causal graph discovery using a novel probabilistic framework. To further enhance computational efficiency and lower complexity, I want to leverage the concept of a Markov blanket approach as well as the subgraph decomposition in my future work.

Moreover, these promising results prompt us to further look into improving the causal pair feature generation to intrinsically capture global information which I plan to implement using graph neural networks that have been discussed in chapter 4. Based on this causal graph learning, the next chapter addresses the challenge of the high dimensionality of data and I propose an approach called causal feature selection, which involves selecting a subset of features that have a causal effect on the outcome variable.

CHAPTER 3: CAUSAL FEATURE SELECTION USING DIRECTED ACYCLIC GRAPHS

3.1 Introduction

The rise of big data has led to a rapid increase in data collection and database creation across various industries, including healthcare, social media, finance, and retail. Consequently, high-dimensional data has become more publicly available, with widespread usage in numerous applications, presenting new challenges for research communities [53]. Real-world high-dimensional datasets, such as gene expression datasets in bioinformatics, can contain hundreds of thousands of features. This large number of features poses significant challenges for machine learning models, which often struggle to handle such high dimensionality effectively [54].

To address the challenges posed by high-dimensional data, feature extraction and feature selection methods have emerged as essential data pre-processing techniques with demonstrated effectiveness. Feature extraction involves transforming a large set of original (raw) features into a new, lower-dimensional, and meaningful feature set, which retains essential information from the original data while reducing computational requirements. In contrast, feature selection methods directly identify a subset of features from the original dataset that carry relevant information about the target concept for model building [55,56]. While feature extraction generates a new feature space, feature selection preserves the interpretability of the original features by retaining relevant ones and removing irrelevant or redundant ones, ultimately enhancing model interpretability [57].

Machine learning models applied to real-world high-dimensional data often suffer from significant information loss, which can degrade their learning performance due to the presence of numerous irrelevant or redundant features. Feature selection methods have been widely employed as a potential solution to this issue, as they reduce storage and memory requirements, increase computational efficiency, and enhance the performance of machine learning models. However, most feature selection techniques rely on correlations or associations between features and the target variable, without providing any direct or causal relationships [19]. Consequently, causal feature selection algorithms have gained increased attention in recent years, aiming to identify a subset of features with a causal effect on the target variable [16]. These algorithms employ causal inference techniques to discern the Markov blanket (MB) of class attributes or a subset thereof, distinguishing between causal and non-causal relationships to improve the interpretability and performance of machine learning models [17, 18, 58].

Yu et al. [59] recently introduced a novel feature selection algorithm, framing it as a local causal structure learning problem. This algorithm is formulated as a multi-label feature selection approach that learns the underlying causal mechanisms of data and selects causally informative features shared by common class labels. In a separate study, Yu et al. [60] presented a multi-source feature selection algorithm leveraging the concept of causal invariance in causal inference. Similarly, Peters et al. [61] proposed a method to identify all 'direct causes' of a target variable of interest by exploiting the invariance of a prediction under a causal model. Paul (2017) [62] suggested a matching technique for identifying meaningful features using causal inference in document classification, despite the fact that the concept of causality may not apply as naturally to document classification as it does to other tasks.

Many existing algorithms, both causal and non-causal, evaluate their performance based on correlation-based metrics such as accuracy, AUC, mean-squared loss, etc. [63–66]. However, causal metrics are necessary for evaluating causal feature selection methods, as they offer a rigorous and principled approach to assessing the effectiveness of these algorithms. Recently, Panda et al. [67] proposed an instance-wise causal feature selection method for interpreting black-box models. In their work, they utilized a variant of the average causal effect (ACE) as a causal evaluation metric, although their formulation was specifically tailored for image data.

In this chapter, I investigate the application of causal inference methodologies for reducing redundant features and selecting optimal relevant features in observational data. My approach involves constructing a directed acyclic graph (DAG) to represent the causal relationships among variables, with edges indicating direct causal effects. This chapter: (1) presents a novel formulation of a causal feature selection (CFS) algorithm that leverages causal structure learning techniques; (2) introduces a new evaluation criterion for CFS using causal metrics; and (3) provides quantitative comparisons on several synthetic and real-world datasets, demonstrating that the truncated subsets of features selected by the CFS algorithm yield comparable or improved performance relative to baseline methods while utilizing fewer causal features.

In Section 3.2 of this chapter, I present the methodology that I use to address the problem at hand. I leverage the existing prior work outlined in Section 3.3 to obtain numerical results. The experimental evaluation of my approach is presented in Section 3.4, where I provide empirical evidence to support my approach. Finally, I summarize my findings and provide insights for future enhancements in Section 3.5.

3.2 Methodology

Causality refers to a relationship between two events, where one event causes the other event to occur. In statistical terms, a dependent variable's value is determined by the value of an independent variable [68]. This relationship can be observed in the causal network depicted in Figure 3.1, where features X_1, X_2, X_3 have a direct causal effect on feature Y, while features X_4, X_5, X_6 have an indirect causal effect. However, features X_7, X_8, and X_9 are only correlated with Y, and not causally related. It is important to note that while causal relationships involve correlation, not all correlations imply causality. Therefore, a causal feature selection method should ideally select only a subset of features from X_1 to X_6 based on their relevance and importance, while ignoring correlated features X_7 to X_9.

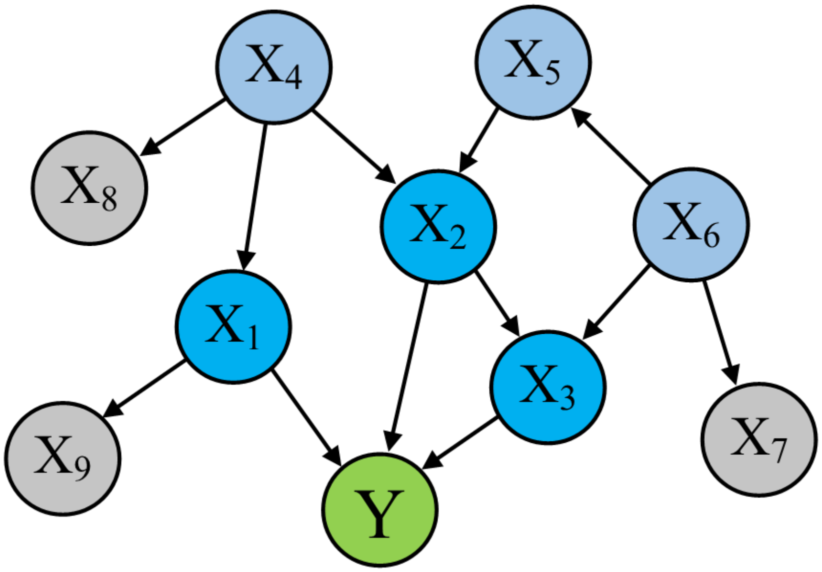


Figure 3.1: Example of a causal network.

The causal effect refers to the distinction between an actual outcome and the alternative outcome that would have arisen if a specific treatment or intervention had not been applied. In the context of causal feature selection evaluation, it is vital to appraise the causal effect of selected features, rather than merely focusing on their predictive power. This is crucial because a feature may have a strong predictive association with the outcome variable, even though it may not have a causal connection. To address this, causal metrics such as average causal effect (ACE), total causal effect (TCE), and average treatment effect (ATE) can be utilized to determine the causal influence of chosen features on the outcome in question. In this study, I employ TCE, which encompasses both the direct and indirect consequences of the treatment on the outcome. Figure 3.1 illustrates the causal relationships between features X_1 to X_9 and the outcome variable Y, where features X_1 to X_3 have a direct causal effect on Y while features X_4 to X_6 have an indirect causal effect. Features X_7 to X_9 have no causal effect on Y as there is no path from these features to Y. It is important to note that the causal effect of a grandparent feature on a variable is often mediated through its parent(s), which can result in a dampening of the causal effect of the grandparent on the variable compared to the direct causal effect of the parent. TCE measures the difference between a hypothetical scenario where all individuals receive the treatment and an alternative scenario where none do. Therefore, in evaluating the causal influence of selected features, it is important to consider their causal effect rather than just their predictive power.

3.2.1 Causal Feature Selection (CFS)

This section introduces a method for causal feature selection (CFS) based on causal graph discovery and the total causal effect (TCE) metric. The TCE of features can be computed by identifying all possible paths between the treatment variable and the outcome (target) variable and aggregating the direct and indirect effects along each path, utilizing randomized controlled trials, natural experiments, or observational studies with adjustments for confounding variables such as DAGs and the do-calculus, Structural equation modeling (SEM), and propensity score matching, etc. This methodology provides a ranking of features that have a higher causal impact on the target variable, indicating the importance of each feature.

A similar ranking of feature importance can also be obtained using a cause-effect pairwise model based on the probability of direction from each feature to the target variable. However, using such a model to get the ranking may be biased if there are confounding factors that affect both the features and the target variable. On the other hand, the causal feature selection using TCE isolates the effect of the features on the target variable while controlling for other factors that may affect the outcome.

However, selecting a subset of features from the TCE ranking requires domain knowledge, which may not always be available. To address this issue, the 'Kneedle' algorithm is proposed, which uses the mathematical definition of curvature for continuous functions to identify the optimal cut-off point for selecting a subset of features from the TCE ranking. The algorithm is explained in more detail in the following paragraph.

It is worth noting that a similar procedure can be applied to non-causal feature selection methods to obtain a ranking of feature importance based on their score criteria. By applying the 'Kneedle' algorithm, a subset of features can be selected from the ranking. For instance, for the correlation-based feature selection method Maximum Relevance - Minimum Redundancy (mRMR), the FCQ score of each feature can be used to obtain a similar ranking.

Kneedle Algorithm. The 'Kneedle' algorithm was proposed by Satopää et al. [69], which detects the optimal cut-off point in a curve by identifying the knee or elbow point, where the curve changes from a steep slope to a flatter one. It is a computationally efficient method with a complexity of O(n log n), where n is the number of points in the curve. The algorithm works by computing the change in slope at each point along the curve, and then looking for the point where the change in slope is maximal, indicating the knee point. The algorithm then uses a statistical

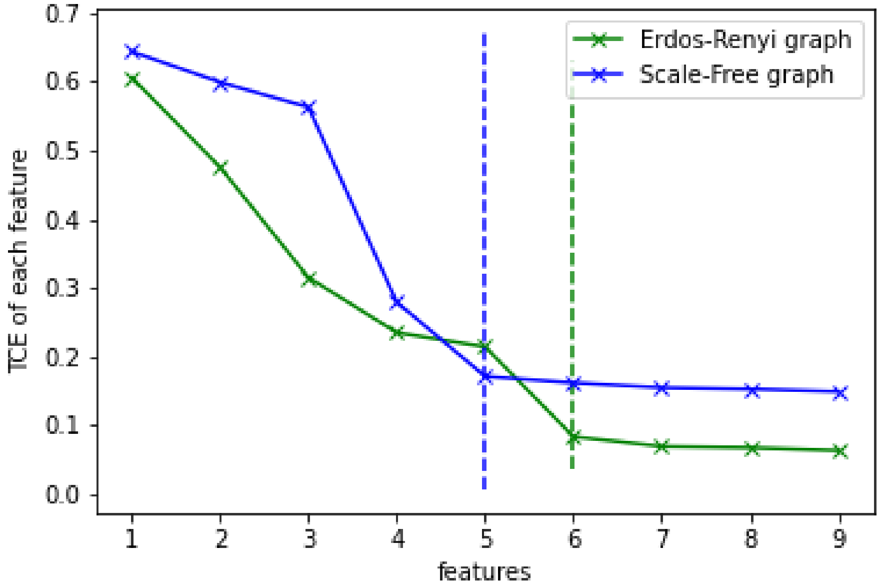


Figure 3.2: An example of the Kneedle algorithm for knee/elbow detection applied to two of the 10-node graph datasets. Given the TCE-based sorted ranking for the features, the algorithm selects the first 6 features for the Erdos-Renyi graph and the first 5 features for the Scale-Free graph as important causal features.

test to determine if the maximal change in slope is statistically significant, and if so, returns the knee point. Otherwise, it continues searching for the next maximal change in slope. Figure 3.2 shows an example of the 'Kneedle' algorithm detecting the cut-off point.

3.2.2 CFS Evaluation Methodology

In the literature discussed in Section 3.1, many causal feature selection (CFS) methods are evaluated without using causal metrics as their criterion [60, 64–66]. To address this issue, I propose a new evaluation criterion for CFS methods based on their causal effect, leveraging the Rank Biased Overlap (RBO) similarity measure algorithm [70]. This new approach offers a different way to assess the performance of causal graph discovery methods compared to traditional evaluation metrics, such as Structural Hamming Distance (SHD), True Positive Rate (TPR), and False Discovery Rate (FDR).

My proposed causal feature selection method uses observational data to identify the subset of causal features through a causal graph discovery method and the total causal effect (TCE) metric, as described in Section 3.2.1. I then identify a baseline ranking of feature importance by applying the same procedure to the ground truth graph. Next, I apply the RBO measure to compare the subset of features given by the CFS method with the baseline subset of features given by the true graph in terms of their ranking positions. The RBO measure produces a similarity score between 0 and 1, with a score of 1 indicating identical subsets and a score of 0 indicating no common features between the two subsets. Figure 3.3 illustrates the steps of my causal feature selection method using the new evaluation criterion.

Rank Biased Overlap. Rank Biased Overlap (RBO) is an intersection-based similarity measure that compares two ranked lists of items. The measure is proposed by Webber et al. in their paper on similarity measures [70]. Similar to Jaccard similarity, RBO counts the proportion of overlapped items as the depth of the ranking increases. However, RBO differs by introducing weights for each rank position. The weights are derived from a convergent series, which means that items with (dis)agreement appearing at the top of the two lists will weigh more than those at the bottom. Given two infinite ranked lists L and M, the RBO score can be calculated using the equation 3.1:

$$RBO(L, M, p) = (1 - p)\sum(p^{d-1})A_d \tag{3.1}$$

Here, d is the depth of the ranking examined, ranging from 1 to ∞. A_d is the agreement between L and M, determined by the ratio of the size of the overlap up to depth d, represented by $(|L_{:\,d} \cap M_{:\,d}|)/d$. The tunable parameter p ranges

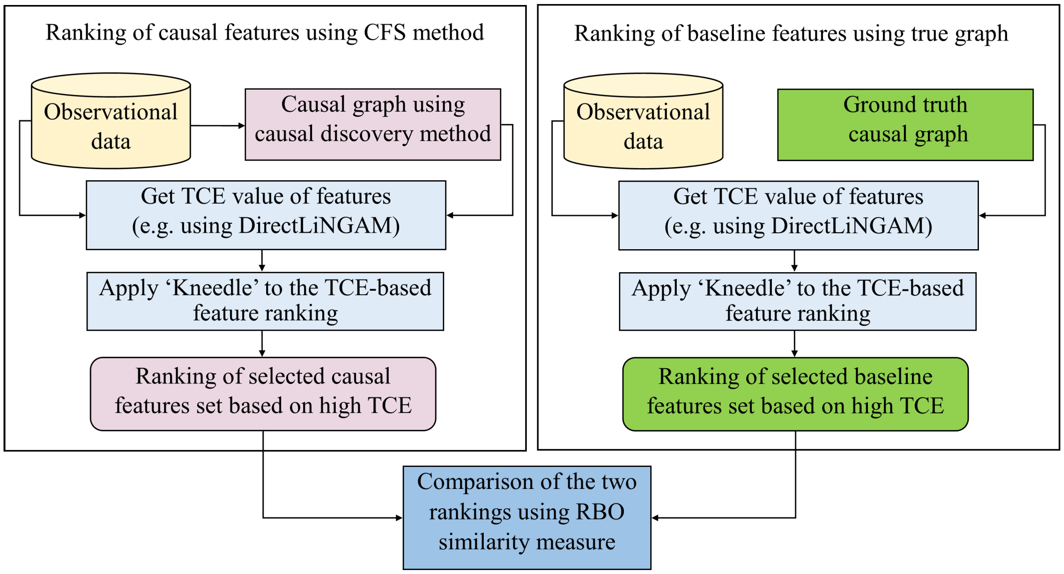

Ranking of causal features using CFS method
Observational data
Causal graph using causal discovery method
Get TCE value of features (e.g. using DirectLiNGAM)
Apply 'Kneedle' to the TCE-based feature ranking
Ranking of selected causal features set based on high TCE
Ranking of baseline features using true graph
Observational data
Ground truth causal graph
Get TCE value of features (e.g. using DirectLiNGAM)
Apply 'Kneedle' to the TCE-based feature ranking
Ranking of selected baseline features set based on high TCE
Comparison of the two rankings using RBO similarity measure

Figure 3.3: Causal feature selection method and the new evaluation criteria using RBO score. First, on the left side box, I estimate a causal graph using a suitable method and select the ranking of the causal features subset. Next, on the right side box, I extract another ranking of causal features subset using the true causal graph as a baseline. Finally, I compare the two rankings of selected features using the Rank-Biased Overlap (RBO) similarity measure to evaluate the performance of my method.

from $(0, 1)$ and contributes towards the final RBO score by determining the top d ranks' contribution.

It is notable that the Rank Biased Overlap (RBO) measure addresses three common issues associated with correlation-based similarity measures, such as Kendall Tau. Firstly, unlike Kendall Tau, RBO does not require the two ranking lists to be conjoint. Secondly, RBO is a weighted measure that assigns greater weight to items with (dis)agreement appearing at the top of the two lists than those appearing at the bottom. Finally, the contribution of a single discordant pair does not decrease as the depth of the ranking increases. A more comprehensive discussion of these advantages can be found in the paper by Webber et al. [70].

3.3 Prior Work used in Numerical Results

In this section, I will review prior work on feature selection methods that I will compare with my proposed causal feature selection (CFS) method. Specifically, I will examine correlation and mutual information based methods such as Maximum Relevance - Minimum Redundancy (mRMR) [71], Information Gain (IG) [72], and Lasso regularization [73] as well as causality-based methods such as Linear Non-Gaussian Acyclic Models (LINGAM) [74], Causal-pairs model [42], and Maximum Likelihood DAG (MLDAG) [33] method. I will explain each method and discuss how they will be used to obtain numerical results for my proposed CFS method.

3.3.1 Maximum Likelihood Directed Acyclic Graph (MLDAG)

MLDAG is a causal graph learning approach introduced by Rashid et al. [33]. The authors describe causal discovery as the process of estimating the underlying causal relations encoded by the directed acyclic graph (DAG), $\mathcal{G}\text{DAG} = (V, E)$, given n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_d] \in \mathbb{R}^{n\times d}$.

The method predicts the probability distribution $p(y_{ij}|f)$ of the causal relation between two variables X_i (cause) and X_j (effect) given the observational dataset $[\mathbf{x}_i, \mathbf{x}_j] \in \mathbb{R}^{n\times 2}$, where $y_{ij} \in [-1, 0, 1]$ is the directions of edges between variable pairs X_i and X_j (1 = causal edge from X_i to X_j, -1 = causal edge from X_j to X_i and 0 = no causal relation between pairs), and $f(\cdot)$ is a pre-trained machine learning model.

$$p(y_{ij}|f) \quad = \quad f([\mathbf{x}_i, \mathbf{x}_j]), \text{ for } i < j \tag{3.2}$$

$$p(\mathcal{G}|f) = \prod_{i<j} p(y_{ij}|f) \tag{3.3}$$

The author constructs a probability distribution of all possible digraphs, $\mathcal{G}$, using equation 3.3 assuming that the variable pairs X_i and X_j are independent, where $\mathcal{G}$ is not always a DAG. To enforce DAGness, the author constrains the conditional distribution in equation 3.3 by integrating out the topological ordering π of the vertices, resulting in equation 3.4. This ensures the probability distribution of digraphs to be DAGs. Using the maximum likelihood estimate from equation 3.5, a representative DAG can be extracted, which I intend to use in my study to estimate a causal graph.

$$p(\mathcal{G}|f, \mathrm{DAG}) = \sum_{\pi} p(\mathcal{G}|f, \mathrm{DAG}, \pi) p(\pi|f) \tag{3.4}$$

$$\mathcal{G}_{\mathrm{DAG}} = \arg\max_{\mathcal{G}} p(\mathcal{G}|f, \mathrm{DAG}) \tag{3.5}$$

3.3.2 Linear Non-Gaussian Acyclic Model (LiNGAM)

Linear Non-Gaussian Acyclic Models (LiNGAM) is a statistical technique for estimating the structural equation models that assume the linear and non-Gaussian relationship between variables and identify the underlying causal structure of observed data. It was introduced by Shimizu et al. [74]. LiNGAM applies statistical tools to estimate the parameters of the model, which separates the observed variables into their causal and non-causal components. This separation helps LiNGAM identify the structure of the causal graph. To do this, LiNGAM assumes that there are no unmeasured (latent) confounding variables that influence both the causal variables and the outcome variables. Note that, this is a hard assumption to make and in many real-world scenarios this will not hold. Although I will use LiNGAM estimation methods in this study since this assumption simplifies the modeling process, I plan to go beyond this limitation in the future by considering alternative modeling approaches that can account for the confounding latent variable and their influences such as Structural Equation Modeling (SEM) or other causal inference techniques.

3.3.3 Cause-Effect Pair Model

The cause-effect pair model is designed to identify the most relevant features that have a strong cause-effect relationship with a specific effect or target variable. The model I focus on is the one proposed by Fonollosa (2019) [42] in response to the 2013 NIPS Workshop on Causality Challenge. According to the author, a new set of features is generated for every feature-target pair, which includes standard statistical features combined with information-theoretic measures and conditional distribution variability measures. The model then predicts the probability of causal direction from each feature to the target using a machine learning model for classification. The higher the probability, the more relevant the feature is to the target. The cause-effect pair model has the potential to be an effective tool for identifying causal relationships in high-dimensional datasets, which can have important applications in various fields, including healthcare, finance, and social sciences.

3.3.4 Maximum Relevance - Minimum Redundancy (mRMR)

The mRMR algorithm is a mutual information-based feature selection method that aims to select the most informative features while minimizing redundancy between them. This algorithm was proposed by Zhao et al. [71], who claim that selecting just a few features from tens of thousands can achieve maximum accuracy for the task at hand. In this study, I will work with the FCQ variant of the mRMR algorithm, which stands for F Statistic to target / Correlation between features.

The relevance of a feature f at the i-th iteration is calculated as the F-statistic between the feature and the target variable, while the average Pearson correlation between the feature and all the features selected in earlier iterations is used to calculate redundancy. This is expressed by the following equation:

$$\mathrm{FCQ}(f) = \frac{\mathrm{F}(f, y)}{\frac{1}{k}\sum_{j=1}^{k} \mathrm{corr}(f, f_j)} \tag{3.6}$$

where y is the target variable, f_j represents the j-th feature already selected in previous iterations, k is the number of features selected so far, and $\mathrm{F}(f, y)$ and $\mathrm{corr}(f, f_j)$ denote the F-statistic between feature f and target variable y and the Pearson correlation between feature f and feature f_j, respectively. By using this criterion, mRMR selects the feature with the highest FCQ value and iteratively adds features that provide maximum information gain while minimizing redundancy with the previously selected features.

3.3.5 Information Gain

Information Gain (IG) or InfoGain is a feature selection technique that quantifies the amount of information provided by a feature in predicting the target variable [72]. It measures the reduction in entropy achieved when the dataset is split based on the feature. By calculating the entropy of the original dataset and the weighted average entropies of the feature's values, Information Gain captures the usefulness of a feature in terms of the predictability it offers. Higher Information Gain values indicate more important features. This technique enables the identification and removal of irrelevant or redundant features, resulting in a more concise and informative feature set for subsequent analysis or modeling tasks.

3.3.6 Lasso Regularization

Lasso regularization [73] is an embedded method utilized to incorporate the l_1-norm of the coefficient of a linear model as a penalty term. The objective function for Lasso regularization is defined as follows:

$$\frac{1}{2N}\sum_{i=1}^{N}\left(y_{true}^{i}-y_{pred}^{i}\right)^{2}+\alpha\sum_{j=1}^{n}|a_j| \tag{3.7}$$

In this equation, a_j represents the coefficient of the j-th feature. The last term represents the l_1 penalty, while α is a hyperparameter that adjusts the strength of the penalty term. The cost function aims to optimize by reducing the absolute values of the coefficients. Higher coefficients contribute to a larger value of the cost function.

Consequently, the objective is to minimize the cost function by shrinking the coefficients toward zero. In cases where two features exhibit linear correlation, their joint presence leads to an increase in the cost function value. Therefore, Lasso regression actively works to shrink the coefficient of the less significant feature to zero, effectively selecting the most relevant features. This property enables feature selection, where features associated with non-zero coefficients are retained, while those with zero coefficients are discarded.

3.4 Experiments

In this section, I will begin by introducing the datasets employed in my experiment, followed by the presentation of the novel evaluation methodologies, where I empirically assess and discuss the efficacy of my methods in comparison to several existing feature selection algorithms and approaches discussed in Section 3.3.

3.4.1 Datasets

Synthetic Data. The benefit of using synthetic data in causal estimation lies in the ability to control and know the true relationships between features when compared to real data. To this end, I generated small graphs that encompassed 16 different combinations of graph characteristics with similar criteria, including the number of nodes, $d = [10, 20]$, the number of edges, $e = [1d, 4d]$, the number of data samples per node, $n = [200, 1000]$, and graph models from Erdos-Renyi (ER) and Scale-Free (SF). Non-linear data samples for the graph nodes were generated, similar to the data generation utilities provided in the NOTEARS-MLP [39] implementation. Furthermore, I generated 10 random graph structures with ground truths for each of these 16 graph types to test my methods.

Similarly, large graphs were generated using comparable criteria, including the number of nodes, $d = [100, 500]$, and the number of data samples per node, $n = [2000, 10000]$.

Real-world Data. I utilized a combination of real-world and semi-synthetic datasets to present the new evaluation criteria. The Protein Expression Network Dataset, as presented in Sachs et al. [52], and the widely employed Lung Cancer Simple Set (LUCAS) data set presented in the Causality Challenge (2008) [75], were the primary real-world datasets used. The same evaluation criteria as the synthetic datasets, as presented in Table 3.3, were employed to evaluate the performance of the CFS methods on these datasets.

Additionally, I employed four datasets from the UCI Machine Learning repository [76], including Communities and Crime, Tom's Hardware (a dataset in Buzz social media), Heart Disease, and Parkinson's Disease, to assess the predictive performance of my method and compare it with other methods. The details of the datasets are presented in Table 3.1.

Table 3.1: A description of the real-world data sets

Dataset	#Dimensions	#Instances
Sachs	11	7466
LUCAS	12	2000
Communities and Crime	123	1994
Tom's Hardware	96	28179
Heart Disease	14	303
Parkinsons Disease	23	197

3.4.2 Evaluation Metrics

In this study, I evaluate the performance of my proposed causal feature selection method using both causal and correlation-based metrics. To quantify the performance, I introduce a new evaluation approach for causal feature selection.

Following the methodology prescribed in section 3.2, I first evaluate the performance of the causal feature selection methods using the Total Causal Effect (TCE) of features on the target variable, which is extracted using the causal graph, as a ranking criterion for the features. I then apply a modified version of the Kneedle algorithm [69] to identify the subset of selected features for each method based on the estimated TCE ranking. I compare the subsets of selected features given by the Kneedle algorithm for each method with the subsets of features extracted using the true graph ranking (as baseline) using the Rank Biased Overlap (RBO) similarity measure. In addition to the causal metrics, I use the scores from equation 3.6 for mRMR, the mutual-information score for InfoGain, and the magnitude of feature coefficients for Lasso regularization as ranking criteria to select features subset using the Kneedle algorithm.

To compare the performance of my causal feature selection method with correlation based methods, I evaluate the predictive performance of different models. I report the number of features selected by each method along with the performance metrics of the predictive models. I use the root mean squared error (RMSE) for the regression data sets (Communities and Crime, Tom's Hardware) and the area under the curve (AUC) score for the classification data sets (Heart Disease, Parkinson's Disease).

3.4.3 Setup and Simulation

To quantify the TCE of features for each method, I utilized the DirectLiNGAM[1] Python tool and used the true graph as prior knowledge to establish the baseline ranking based on the TCE value. Likewise, I used MLDAG causal graph estimation approach proposed by Rashid et al. [33] as prior knowledge to derive the TCE-based feature ranking for the MLDAG method. Subsequently, I utilized DirectLiNGAM [1] to learn a causal graph from the data, which provided a TCE-based feature ranking for the LiNGAM method. In addition, I implemented a modified version [2] of the cause-effect pair method proposed by Fonollosa (2019) [42] (hereafter referred to as 'Causal Pairs' method) and mRMR-based feature selection algorithms. I implement the Kneedle algorithm [69] in Python to identify the knee/elbow point and select a subset of features from the TCE-sorted rankings given by each method. Lastly, I implement the variant[3] of Rank Biased Overlap (RBO) algorithm to compare the subsets of selected features.

To compare the predictive performances of the selected features by each method on the selected data sets I have used the scikit-learn implementation of Linear regression, Logistic regression, K-Nearest Neighbours (KNN), and Random Forest models for regression and classification data sets respectively.

[1] DirectLiNGAM: https://lingam.readthedocs.io/en/latest/tutorial/lingam.html

[2] Jarfo Cause-effect Pair Model: https://github.com/jarfo/cause-effect

[3] RBO implementation: https://github.com/dlukes/rbo

3.4.4 Results

Table 3.2 presents the outcomes of my novel evaluation criteria applied to synthetic data for various causal feature selection methods. The table provides information on the number of features selected by each method and the RBO (Rank Biased Overlap) score of the selected feature subset, compared to the subset of features selected based on the true graph ranking as the baseline. Our observations indicate that the causal probability-based method, 'Causal Pairs', outperforms the correlation-based methods. Moreover, causal graph discovery methods demonstrate superior performance compared to causal probabilities. Specifically, MLDAG and LiNGAM outperform Causal Pairs and other methods across all graph structures. While MLDAG exhibits superior performance for small graph structures, LiNGAM performs better for larger graphs. Notably, the correlation-based method mRMR achieves comparable performance to the causality-based method for small graphs. These findings suggest that MLDAG and LiNGAM estimate distinct causal structures from the data, resulting in different feature rankings based on their Total Causal Effect (TCE) outcomes. Therefore, combining both methods to create a hybrid causal feature selection approach can yield a robust set of selected features that outperforms correlation-based methods.

It is worth noting that the mRMR feature selection ranking score curve is unusual for these datasets. The modified f-score calculated in each feature selection for the ranking is independent of the others. Therefore, a feature with a lower rank can have a higher f-score than the previously selected higher-rank feature, which results in a steep rise-fall of peaks in the middle of the curve. To address this, I used the same number of features selected by the baseline true graph for the mRMR method in Table 3.2.

Table 3.2: Comparison of my causal feature selection method with other methods applied to synthetic data in terms of the number of features selected and the RBO similarity measure against the ranking of selected features by the true graph. The means and standard errors of the performance metrics are based on the 40 different graph structures for each category in the test data.

Graph Type ↓	Graph Size →	10 Nodes		20 Nodes		100 Nodes		500 Nodes	
	Method ↓	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score
Erdos-Renyi	Causal Pairs	6.05	0.327±0.03	13.88	0.248±0.02	49.00	0.205±0.06	312.25	0.150±0.05
	LiNGAM	3.30	0.419±0.04	5.63	0.409±0.03	12.50	**0.239±0.06**	21.25	**0.277±0.08**
	MLDAG	4.30	**0.514±0.04**	8.35	**0.454±0.03**	33.25	0.138±0.04	107.25	0.159±0.07
	mRMR	4.25	0.417±0.04	6.93	0.370±0.03	24.75	0.169±0.05	46.25	0.189±0.08
	InfoGain	3.70	0.382±0.03	7.23	0.311±0.04	21.25	0.161±0.07	90.25	0.072±0.02
	Lasso	4.00	0.420±0.03	7.13	0.397±0.03	9.25	0.208±0.07	12.00	0.218±0.09
Scale-Free	Causal Pairs	5.98	0.369±0.03	13.35	0.284±0.03	62.00	0.308±0.09	244.75	0.193±0.08
	LiNGAM	3.65	0.521±0.04	5.73	**0.523±0.03**	12.50	**0.343±0.05**	50.25	**0.206±0.06**
	MLDAG	3.50	**0.550±0.04**	5.70	0.516±0.03	24.25	0.269±0.09	61.75	0.171±0.08
	mRMR	3.95	0.453±0.03	7.13	0.399±0.03	14.50	0.253±0.03	30.75	0.142±0.04
	InfoGain	4.48	0.368±0.04	7.60	0.342±0.02	27.00	0.127±0.05	97.50	0.085±0.06
	Lasso	4.03	0.462±0.03	7.05	0.446±0.03	6.50	0.263±0.04	8.00	0.170±0.05

Table 3.3: Comparison of my causal feature selection method with other methods in terms of the number of features selected and the RBO similarity measure against the ranking of selected features by the true graph for Sachs and LUCAS data sets. Note, the Sachs data set does not have any defined target, therefore, probable 3 different targets from the protein signaling network proposed in the manuscript [52] have been experimented

Graph Dataset	Total	MLDAG		Causal Pairs		LiNGAM		mRMR		InfoGain		Lasso	
	#feat.	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score	#feat.	RBO score
Sachs (P38)	10	5	0.341	9	0.344	4	0.392	5	0.368	3	0.161	2	**0.407**
Sachs (ERK)	10	3	0.357	5	0.577	8	**0.727**	2	0.614	3	0.583	2	0.614
Sachs (AKT)	10	3	0.161	5	0.358	2	**0.581**	2	**0.581**	3	0.322	2	0.133
LUCAS	11	2	0.416	8	**0.674**	5	0.502	5	0.633	5	0.652	6	0.462

Table 3.4: Prediction performance of different methods using several machine learning models applied to classification data sets. The mean and standard errors of testing AUC shown here are based on 10 different test runs for each data set.

Dataset	HEART Disease				Parkinson's Disease			
Methods	#features	Logistic Reg.	KNN	Random Forest	#features	Logistic Reg.	KNN	Random Forest
Original	13	0.822±0.01	**0.818±0.01**	0.834±0.01	22	0.769±0.02	**0.828±0.03**	0.711±0.03
Causal Pairs	7	0.808±0.01	0.782±0.01	0.835±0.01	2	**0.785±0.03**	0.740±0.03	**0.819±0.02**
MLDAG	6	0.814±0.03	0.812±0.01	0.826±0.01	4	0.776±0.03	0.721±0.03	0.812±0.02
Lingam	2	0.776±0.01	0.725±0.02	0.749±0.03	2	0.766±0.03	0.774±0.02	0.775±0.03
mRMR	7	**0.827±0.01**	0.810±0.01	0.838±0.01	4	0.763±0.03	0.812±0.03	0.811±0.02
InfoGain	3	0.814±0.01	0.771±0.01	0.810±0.01	8	0.777±0.03	0.803±0.03	0.772±0.03
Lasso	3	0.819±0.01	0.801±0.02	**0.845±0.01**	5	**0.785±0.03**	0.787±0.03	0.756±0.03

Table 3.5: Prediction performance of different methods using several machine learning models applied to regression data sets. The mean and standard errors of testing RMSE shown here are based on 10 different test runs for each data set.

Dataset	Communities and Crime				Tom's Hardware			
Methods	#features	Linear Reg.	KNN	Random Forest	#features	Linear Reg.	KNN	Random Forest
Original	122	0.599±0.01	0.652±0.01	0.596±0.00	96	**0.196±0.00**	0.267±0.01	0.216±0.01
Causal Pairs	15	0.607±0.00	0.633±0.01	0.604±0.01	48	0.199±0.00	0.223±0.01	**0.215±0.01**
MLDAG	98	0.604±0.01	0.663±0.01	0.599±0.01	54	0.224±0.00	0.245±0.01	0.233±0.00
Lingam	30	0.608±0.01	0.652±0.01	0.603±0.01	12	0.204±0.00	0.177±0.01	**0.215±0.01**
mRMR	99	0.602±0.01	0.656±0.00	**0.595±0.00**	55	0.197±0.00	0.232±0.01	**0.215±0.01**
InfoGain	28	0.597±0.01	**0.632±0.01**	0.599±0.01	9	0.206±0.00	**0.155±0.01**	0.216±0.01
Lasso	26	**0.579±0.01**	**0.632±0.01**	0.596±0.01	17	0.199±0.00	0.195±0.01	**0.215±0.01**

In Table 3.3, I present the results obtained from applying various causal feature selection methods to the Sachs and LUCAS datasets. The findings reveal that LiNGAM and mRMR exhibited superior performance on the Sachs dataset, while Causal Pairs outperformed other methods on the LUCAS dataset, utilizing a larger set of features. These results demonstrate the enhanced efficacy of incorporating causal inference techniques compared to traditional correlation-based approaches for feature selection. Notably, we observed that mRMR can achieve comparable or even superior performance to causality-based methods in the context of causal network datasets. This compelling finding motivates us to explore the potential benefits of combining mRMR with other causal discovery methods.

To assess the predictive capabilities of the causal feature selection methods in comparison to other techniques, I employed various machine learning predictive models on UCI datasets, utilizing the subset of features selected by each method. The results of the classification and regression tasks are presented in Table 3.4 and Table 3.5, respectively. Interestingly, we observed that causality-based methods exhibit similar performance to correlation-based approaches in terms of predictive accuracy. Moreover, incorporating causality generally leads to a reduction in the number of features while maintaining comparable predictive performance. Additionally, I found that correlation-based methods, such as mRMR and Lasso regularization, perform equally well or better than other techniques in the regression task. These findings open up possibilities for integrating such correlation-based methods into causal discovery frameworks.

3.5 Summary

This study proposed the causal feature selection (CFS) method as an approach to select informative and relevant features from observational data. The use of causal graphs provided a unique advantage over traditional correlation-based metrics. I introduced new CFS evaluation criteria using causal metrics, such as the total causal effect and the Kneedle algorithm. The experimental results on synthetic and real-world datasets demonstrated that the proposed CFS method is statistically better or on par with other baseline and traditional approaches.

In future work, I plan to explore other causal graph estimation methods to calculate the total causal effect and use additional causal metrics to improve the efficiency of my method which I will explain in brief in the next chapter. Moreover, I aim to apply my method to different domain datasets such as vision and text data to evaluate its effectiveness in various applications. Overall, this study contributes to the development of causal feature selection methods and highlights the importance of causal inference in feature selection for machine learning tasks.

CHAPTER 4: A GRAPH NEURAL NETWORK-BASED PROBABILISTIC FRAMEWORK FOR CAUSAL DISCOVERY

4.1 Introduction

Causal inference from observational data is a fundamental task in many disciplines [3, 52, 77–79] 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 [80]. Score-based causal discovery algorithms optimize fit scores over various graph structures [81]. While effective in many situations, these approaches suffer from exponential run-times and combinatorial explosions in statistic complexity as the data sets grow [82]. Recent advancements in machine learning, such as the NOTEARS algorithm, employ continuous optimization to enforce acyclicity, enhancing computational efficiency [83]. 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.

The emergence of graph neural networks (GNNs) has revolutionized the field of supervised learning on graph-structured data, enabling powerful representations and insights from complex networks and relationships. From social network analysis to molecular property prediction (e.g., modeling interactions of atoms in a chemical molecule) [20, 21], 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 [22, 23]. Despite their success in areas such as network analysis and bioinformatics [24, 25], 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 [3, 26, 27]. DAG-GNN, for instance, focuses on deterministic structure learning, while my methods use a probabilistic framework to better capture the inherent uncertainties in causal relationships [26].

This paper proposes a novel GNN-based probabilistic framework for causal discovery that goes beyond the existing causal pairs methods, including the work by Rashid, Chowdhury, and Terejanu [33], by capturing global information in the graph structure.

This work makes several key contributions:

- This research enhances causal structure learning by refining the probability distribution of all possible digraphs.
- It provides a comprehensive understanding of causal discovery by learning a spectrum of causal graphs instead of producing a single deterministic graph.
- It outperforms conventional non-GNN-based methods in terms of accuracy and scalability.

The proposed framework's capabilities are validated through a comprehensive evaluation process, involving the analysis of both synthetic data derived from non-linear SEMs, benchmark datasets, and real-world datasets, demonstrating its effectiveness in diverse scenarios. My approach surpasses benchmark methods, including traditional techniques (PC [80], GES [81]) and recent non-GNN-based methods (LiNGAM [74], NOTEARS-MLP [83]) and GNN-based method: DAG-GNN [26], in terms of accuracy and scalability on synthetic datasets, while also performing favorably compared to DAG-GNN and NOTEARS-MLP, and outperforming LiNGAM and GES for real-word dataset.

Section 4.2 reviews the related work, followed by the problem formulation and a detailed explanation of my causal discovery approach using GNNs in Section 4.3. The empirical evaluation of my methods is presented in Section 4.4. Finally, Section 4.5 summarizes the findings and discusses potential future improvements.

4.2 Related Work

Structure learning from observational data typically follows either constraint-based or score-based methodologies. Constraint-based approaches, like the PC algorithm [80], start by employing conditional independence tests to map out the underlying causal graph's skeleton. Alternatively, score-based strategies, such as those implemented by GES [81], involve assigning scores to potential causal graphs according to specific scoring functions [82, 84], and then systematically exploring the graph space to identify the structure that optimizes the score [14,85]. However, the challenge of pinpointing the optimal causal graph is NP-hard, largely due to the combinatorial nature of ensuring acyclicity in the graph [86, 87]. Although these methods provide theoretical performance guarantees under certain conditions, their practical application often falls short, particularly when faced with the complexities of real-world data.

Another approach focuses on identifying cause-effect pairs using statistical techniques from observational data. Fonollosa's work on the JARFO model [42] is a notable effort in this direction, employing a conditional independence-based approach to infer causal relationships from pairs of variables. Despite the promise of these pairwise methods, they often fail to leverage global structural information, limiting their effectiveness in constructing comprehensive causal graphs.

Recent advancements, such as the NOTEARS algorithm [88], incorporate continuous optimization techniques to ensure the acyclicity of the learned graph without requiring combinatorial constraint checks, representing a significant improvement in computational efficiency and scalability. However, experiments indicate that this method is highly sensitive to data scaling [89].

On the other hand, geometric deep learning, specifically GNNs, combined with causality has revolutionized learning paradigms in domains dealing with graph-structured data [20, 21, 24]. Despite the success of GNNs in various domains, their application in causal discovery remains limited. A few pioneering works have begun exploring this avenue, each with its own perspective [11, 90–92]. Li et al. [93] propose a probabilistic approach for whole DAG learning using permutation equivariant models. This method demonstrates how supervised learning can be applied to structure discovery in graphs. DAG-GNN [26] uses a variational autoencoder parameterized by GNNs to learn directed acyclic graphs (DAGs), focusing on deterministic structure learning and primarily utilizing node features. My methods, in contrast, emphasize a probabilistic framework, incorporating both node and edge features. Interestingly, my algorithm can complement DAG-GNN by providing a probabilistic distribution over possible DAGs, potentially refining its causal structure learning. Another study presents a gradient-based method for causal structure learning with a graph autoencoder framework, accommodating nonlinear structural equation models and vector-valued variables, and outperforming existing methods on synthetic datasets [94]. Furthermore, the Gem framework provides model-agnostic, interpretable explanations for GNNs by formulating the explanation task as a causal learning problem, achieving superior explanation accuracy and computational efficiency compared to state-of-the-art alternatives [95].

However, these attempts have often not fully exploited the capabilities of GNNs, particularly in learning complex causal structures dynamically from data. This is primarily because many of these studies either focus on deterministic approaches or do not incorporate the rich information available from graph structures (e.g. edge features, probabilistic nature of causal relationships), leading to a lack of comprehensive modeling of causal dependencies. My proposed work seeks to bridge this gap by developing a GNN-based probabilistic framework specifically tailored for causal graph learning, which utilizes both the intrinsic graph-based nature of causal relationships and the powerful representational learning capabilities of GNNs.

4.3 Methodology

Assuming I have n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x_1} \dots \mathbf{x_d}] \in \mathbb{R}^{n \times d}$, causal discovery attempts to estimate the underlying causal relations encoded by the di-graph, $\mathcal{G} = (V, E)$. 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 $\mathcal{G}$. In other words, the presence of the edge $i \rightarrow j$ corresponds to a direct causal relation between X_i (cause) and X_j (effect).

This 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.

$$p(e_{ij}|\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}) \quad = \quad f([\mathbf{h}_i, \mathbf{h}_j, \mathbf{e}_{ij}]), \text{ for } i < j \tag{4.1}$$

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 $e_{ij} \in [-1, 0, 1]$.

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

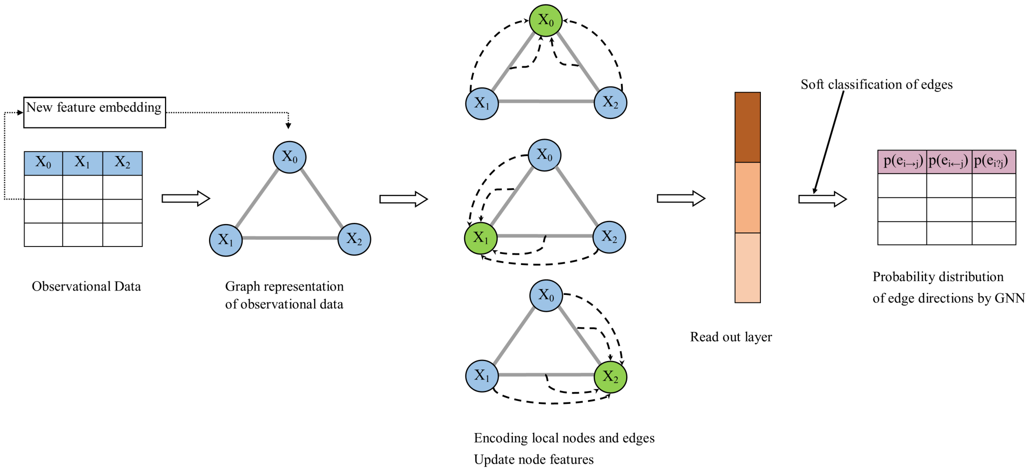


Figure 4.1: Graph representation of observational data and predicting Edge directions. Each node in the graph is initialized with node features and each edge between node pairs is initialized with aggregated edge features of extracted new edge features from attribute pairs and probabilities of edge directions from CausalPairs model [33].

4.3.1 Feature Engineering and Graph Construction

I 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 $d(d-1)/2$ edges in the graph for a dataset with d attributes.

Initially, a comprehensive graph $\mathcal{G} = (V, E)$ is constructed, wherein V denotes the set of all attributes present in the observational dataset, and E represents the set of edges connecting nodes (attributes) in a manner that ensures complete interconnectivity. Consequently, the graph comprises $d(d-1)/2$ edges for a dataset comprising d attributes, thereby establishing a fully connected structure.

The construction of the graph $\mathcal{G} = (V, E)$ begins with a complete graph, where V is the set of attributes from the observational dataset, and E consists of all possible edges between nodes, yielding a total of $d(d-1)/2$ edges for a dataset with d attributes, thereby ensuring that every attribute is connected to every other attribute. I 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.

For node features, I include measures such as normalized entropy, skewness, kurtosis, and log of the number of unique samples, providing a comprehensive representation of each attributeâs distribution. For edge features, I use metrics like mutual information, conditional entropy, poly-fit error, and normalized error probability, which capture the relationships between pairs of attributes. Additionally, features like conditional distribution entropy variance and Pearson correlation coefficient are used to further enrich the edge feature set. I also incorporate the probability distribution over the edge direction using the causal-pairs model [33] as an additional edge feature aggregated with the extracted 115 edge features. In total, I have 118 features for each edge in the graph.

A simplified illustration is shown in Figure 4.1. The intuition behind this approach is that by creating a comprehensive feature set that includes both node and edge features, I can capture a rich representation of the underlying dependencies and interactions between variables. The fully connected graph ensures that all possible relationships are considered, allowing the model to learn from a wide range of potential causal connections. Furthermore, incorporating the probability distribution from the causal-pairs model adds another layer of probabilistic reasoning, enhancing the model's ability to infer causal directions accurately. This multi-faceted feature representation enables the GNN to leverage both local and global information, leading to more accurate and reliable causal predictions.

4.3.2 Developing the Graph Neural Network (GNN) Model

Graph neural networks (GNNs) are a family of architectures that leverage graph structure, node features, and edge features to learn dense graph representations. GNNs employ a neighborhood aggregation strategy, iteratively updating node representations by aggregating information from neighboring nodes. For example, a basic operator for neighborhood information aggregation is the element-wise mean.

In this study, I utilize a GNN model to predict edge directions by training it on synthetic datasets with underlying causal graphs. The GNN model, serving as an edge classifier, infers the probability distribution over edge directions through supervised learning. I specifically adopt the GraphSAGE framework, which performs the message-passing operation and iteratively updates node features. GraphSAGE samples a fixed number of neighbors for each node rather than using the entire neighborhood, enhancing scalability for large graphs and making neighborhood aggregation computationally tractable since I am using a complete graph as input. Although GraphSAGE is primarily designed to update node features based on neighboring node features, I extend it to incorporate edge features in the message-passing process.

To integrate both node and edge features, I define the message $m_{uv}^{(k)}$ as a combination of the feature vectors of nodes u and v at layer $k-1$, along with the edge feature vector e_{uv}. The updated equations for message passing and node feature updates are as follows:

$$m_{uv}^{(k)} = \text{CONCAT}(h_u^{(k-1)}, h_v^{(k-1)}, e_{uv}) \tag{4.2}$$

$$m_v^{(k+1)} = \frac{1}{|N(v)|} \sum_{u \in N(v)} m_{uv}^{(k)} \tag{4.3}$$

$$h_v^{(k+1)} = \sigma \left(W \cdot \text{CONCAT}(h_v^{(k)}, m_v^{(k+1)}) \right) \tag{4.4}$$

Here,

- For each neighboring node u of node v, I calculate a message $m_{uv}^{(k)}$ by concatenating the feature vectors of node u and node v at layer $k-1$ along with the edge feature vector e_{uv}.

- The messages $m_{uv}^{(k)}$ from all neighbors $u \in N(v)$ are aggregated by summing them and normalizing by the number of neighbors $|N(v)|$. This normalization ensures that contributions from all neighbors are equally weighted.

- The aggregated message $m_v^{(k+1)}$ is concatenated with the current feature vector of node v ($h_v^{(k)}$).

- The concatenated vector is then passed through a linear transformation defined by the learnable weight matrix W, followed by a non-linear activation function σ (e.g., ReLU).

This model captures both local and global dependencies in the graph structure, enhancing the accuracy of inferred causal relations between nodes considering their relationships with neighbors. After multiple rounds of message passing, the final node embeddings represent each node and edge in the graph, allowing for the prediction of edge direction probabilities (forward, reverse, or no edge) between any pair of nodes.

4.3.3 Probabilistic Inference

The probabilities on edges (edge directions) predicted by the GNN model represent a distribution over all possible graphs, rather than a single $p(\mathcal{G}_{DAG})$. This approach captures a comprehensive view of potential causal structures instead of committing to a single deterministic graph.

To estimate a sample digraph (PG), a maximum likelihood digraph (MLG), a sample DAG (PDAG), and a maximum likelihood DAG (MLDAG) from the GNN-derived probability distributions over all possible graph structures, I employ the method described in [33], involving several steps:

Sample Digraph (PG). After calculating the probability distributions of causal relations between node pairs or edge directions, a straightforward approach to construct the probability distribution of a digraph $\mathcal{G}$ is to assume that the directions of edges are independent (Eq. 4.5).

$$p(\mathcal{G}|f) = \prod_{i<j} p(e_{ij}|f) \tag{4.5}$$

Maximum Likelihood Digraph (MLG). By selecting the edge directions with the highest probabilities, I construct a maximum likelihood digraph, representing the most probable causal structure using Eq. 4.6.

$$\mathcal{G}_{\mathrm{ML}} = \arg\max_{\mathcal{G}} p(\mathcal{G}|f) \tag{4.6}$$

Note that the samples from the probability distribution, Eq. 4.5 and Eq. 4.6, are digraphs with no guarantees of acyclicity. To ensure the graphs are acyclic, I incorporate DAG constraints using the maximum spanning DAG approach [44] and topological sorting:

$$p(\mathcal{G}|f, \text{DAG}) = \sum_{\pi} p(\mathcal{G}|f, \text{DAG}, \pi) p(\pi|f) \tag{4.7}$$

Here, π represents the topological ordering of the vertices, ensuring acyclicity. Due to computational intractability, I use the law of total probability, integrating out π. I approximate the topological ordering, π_{ML}, by the topological sorting of the Maximum Spanning DAG (MSDAG):

$$\pi_{\text{ML}} \quad = \arg\max_{\pi} p(\pi|f) \approx \text{toposort}(\text{MSDAG}(\mathcal{G}_A)) \tag{4.8}$$

To find the topological sorting of the MSDAG of $\mathcal{G}_A$, I use the approach introduced by [96], constructing the maximum spanning tree and adding edges in descending order of weights while avoiding cycles using the following equation:

$$p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) = \prod_{\pi_{\text{ML}}^{-1}[i] < \pi_{\text{ML}}^{-1}[j]} p(e_{i \to j}|f) \tag{4.9}$$

Sample DAG (PDAG). I ensure acyclicity by sampling edges based on their probabilities and enforcing the DAG constraint using MST and topological sorting methods as described in Eq. 4.9).

Maximum Likelihood DAG (MLDAG). Similar to the MLG, but with the added requirement of acyclicity, this graph (a deterministic representation) is constructed by selecting the most probable edges and adjusting to ensure no cycles using Eq. 4.10.

$$\mathcal{G}_{\text{DAG}} \quad \approx \quad \arg\max_{\mathcal{G}} p(\mathcal{G}|f, \text{DAG}, \pi_{\text{ML}}) \tag{4.10}$$

Detailed derivations for these equations are provided in [33].

4.4 Experiments

I use the following labels for my approaches: GNN-PG (sample graph from the probability distribution), GNN-MLG (maximum likelihood estimate graph), GNN-PDAG (sample DAG from the probability distribution), and GNN-MLDAG (DAG using the maximum likelihood estimate).

I present empirical results from both synthetic and real-world datasets, comparing my methods with traditional approaches (PC [80], GES [81]), CausalPairs approaches [33] and recent methods (LiNGAM [74], DAG-GNN [26], NOTEARS-MLP [88]). Public implementations of PC, GES, and LiNGAM were used and for DAG-GNN and NOTEARS-MLP, I followed the provided implementations in their respective manuscript and git repository. Default settings and hyperparameters were used for all implementations.

4.4.1 Datasets

Synthetic Data. I generated synthetic data to train my GNN model on causal graph estimation, producing 200 graphs with 72 different combinations of nodes (d = [10,20,50,100]), edges (e = [1d, 2d, 4d]), data samples per node (n = [500, 1000, 2000]), and graph models (Erdos-Renyi and Scale-Free). Non-linear data samples were generated similarly to the NOTEARS-MLP implementation, with random graph structures and ground truth for training. The process for generating synthetic test data follows the methodology outlined in [33], where 160 types of graph combinations were considered, each with varying numbers of nodes, edges, graph types, and data samples per node.

CSuite Data. In addition to the synthetic test datasets, I employed two benchmark datasets from Microsoft CSuite, a collection designed for evaluating causal discovery and inference algorithms [97]. The CSuite data is generated from well-defined hand-crafted structural equation models (SEMs), which serve to test various aspects of causal inference methodologies. The five datasets utilized in this study are: *large_ backdoor* (9 nodes, 10 edges); *weak_ arrows* (9 nodes, 15 edges); *mixed_ simpson* (4 nodes, 4 edges); *nonlin_ simpson* (4 nodes, 4 edges); *symprod_ simpson* (4 nodes, 4 edges);. Each dataset includes 6000 data samples, and a corresponding ground truth graph, providing a basis for performance evaluation.

Real-World Data. I used the dataset from [52], based on protein expression levels. This dataset is widely used due to its consensus ground truth of the graph structure, consisting of 11 protein nodes and 17 edges representing the protein signaling network. I aggregated 9 data files, resulting in a sample size of 7466 for my experiments.

4.4.2 Metrics

I evaluated the causal graphs using True Positive Rate (TPR), False Positive Rate (FPR), and Structural Hamming Distance (SHD). A lower SHD and FPR indicate better performance, while a higher TPR is preferable. SHD, TPR, and FPR were calculated as defined for PC, GES, and NOTEARS-MLP, with GNN-based and CausalPairs-based methods following the implementation procedures used in [33].

4.4.3 Results

Table 4.1 showcases the superior performance of my GNN-based methods on 80 Scale-Free (SF) and 80 Erdos-Renyi (ER) graph structures. My methods consistently outperform traditional and recent approaches, demonstrating improved recovery of causal structures through reduced Structural Hamming Distance (SHD) and increased True Positive Rate (TPR). Key observations across both graph structures include:

1. My GNN-based methods, especially GNN-PDAG and GNN-MLDAG, con-

Table 4.1: 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) and 80 Erdos-Renyi (ER) graph structures in the test data.

Dataset type →	**Scale-Free (SF)**			**Erdos-Renyi (ER)**		
Method ↓ / Metrics →	**SHD/d**	**TPR**	**FPR**	**SHD/d**	**TPR**	**FPR**
GNN-PG	1.88±0.08	0.51±0.02	0.30±0.01	2.08±0.11	0.52±0.02	0.52±0.06
GNN-MLG	1.85±0.13	0.20±0.02	0.01±0.00	2.17±0.17	0.25±0.02	0.01±0.00
GNN-PDAG	1.55±0.07	0.56±0.02	0.19±0.01	1.75±0.11	0.61±0.03	0.28±0.03
GNN-MLDAG	1.40±0.11	0.48±0.03	0.08±0.01	1.66±0.15	0.54±0.03	0.13±0.02
CausalPairs-PG	2.02±0.12	0.31±0.01	0.26±0.02	2.38±0.14	0.39±0.02	0.72±0.10
CausalPairs-MLG	1.97±0.13	0.12±0.01	0.03±0.01	2.32±0.17	0.15±0.02	0.07±0.01
CausalPairs-PDAG	1.96±0.12	0.30±0.01	0.21±0.02	2.30±0.15	0.38±0.02	0.61±0.09
CausalPairs-MLDAG	1.88±0.13	0.20±0.01	0.09±0.01	2.18±0.16	0.28±0.02	0.29±0.05
PC	1.93±0.15	0.17±0.02	0.08±0.01	2.40±0.21	0.17±0.02	0.22±0.04
GES	1.43±0.11	0.51±0.03	0.26±0.04	1.78±0.13	0.48±0.02	0.87±0.15
LiNGAM	1.68±0.11	0.35±0.02	0.34±0.04	1.97±0.13	0.43±0.02	1.04±0.17
DAG-GNN	1.75±0.12	0.24±0.02	0.02±0.00	2.10±0.17	0.27±0.02	0.06±0.00
NOTEARS-MLP	1.36±0.11	0.47±0.02	0.12±0.02	1.33±0.10	0.58±0.02	0.32±0.06

sistently achieve lower SHD and higher TPR values compared to CausalPairs methods and traditional methods such as PC and GES. They also perform favorably or better than advanced methods such as LiNGAM, DAG-GNN, and NOTEARS-MLP. Notably, they significantly improve TPR while maintaining low SHD.

2. The GNN-MLG method significantly minimizes false positive causal relationships but at the cost of a lower TPR. Other GNN-based methods balance TPR and FPR.

3. Enforcing DAG constraints in GNN-PDAG and GNN-MLDAG improves performance metrics relative to GNN-PG and GNN-MLG, highlighting the benefit of integrating global structural information to enhance accuracy.

Figure 4.2 presents a comprehensive comparison of the Structural Hamming Distance (SHD), True Positive Rate (TPR), and False Positive Rate (FPR) performance metrics for different methods on 160 SF and ER graphs with node-to-edge ratios of 1:1 and 1:4.

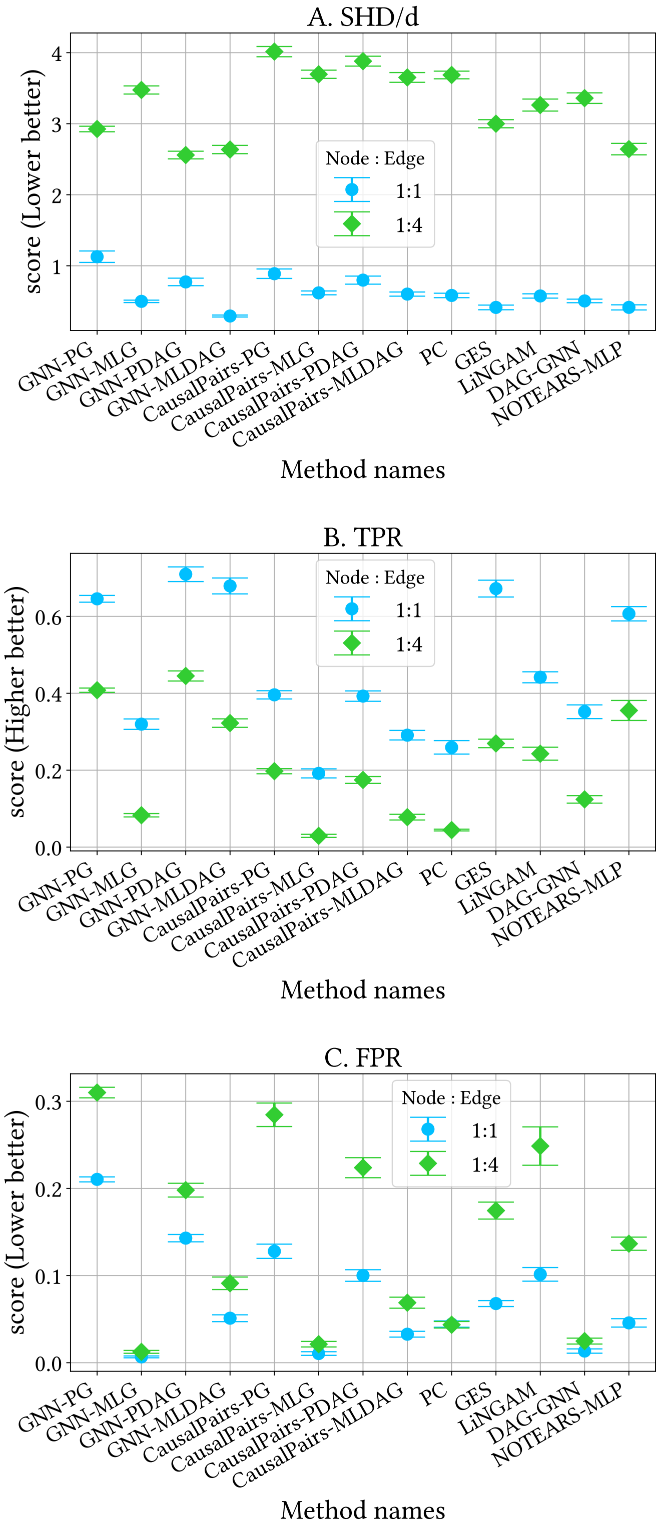


Figure 4.2: Comparison of SHD (A), TPR (B) and FPR (C) for different methods on ER and SF graph structures illustrating their mean results and standard error of metrics, plotted against varying Node-to-Edge ratios.

The GNN-based methods, specifically GNN-PDAG and GNN-MLDAG, consistently achieve lower SHD values than traditional methods (PC and GES), CausalPairs methods, and advanced methods (NOTEARS-MLP and DAG-GNN). Notably, my proposed methods (GNN-PG, GNN-PDAG, and GNN-MLDAG) demonstrate significantly higher TPRs than all other methods, indicating improved accuracy in identifying true causal relationships. GNN-PDAG and GNN-MLDAG exhibit robust performance across both sparse (1:1) and dense (1:4) graphs, showcasing their ability to accurately recover causal structures with fewer errors. The improvement is more pronounced in denser graphs (1:4 node-to-edge ratio), showing promise in handling complex, highly connected networks.

Table 4.2: Comparison of GNN-based edge probability model (trained on synthetic train data) on the Microsoft CSuite datasets.

Dataset Name →	**large_backdoor**			**weak_arrows**		
Method↓ / Metrics →	**SHD/d**	**TPR**	**FPR**	**SHD/d**	**TPR**	**FPR**
GNN-PG	0.59	0.42	0.20	0.56	0.66	0.24
GNN-MLG	0.68	0.32	0.17	0.82	0.51	0.09
GNN-PDAG	0.56	0.44	0.19	0.67	0.6	0.29
GNN-MLDAG	0.55	0.44	0.18	0.66	0.6	0.28
CausalPairs-PG	2.42	0.88	0.80	2.24	0.85	0.93
CausalPairs-MLG	1.77	0.88	0.55	1.89	0.82	0.68
CausalPairs-PDAG	2.28	0.97	0.75	2.06	0.95	0.85
CausalPairs-MLDAG	2.14	0.96	0.70	1.97	0.94	0.81
PC	1.00	0.53	0.29	0.89	0.44	0.22
GES	1.33	0.67	0.67	0.88	0.88	0.37
LiNGAM	2.22	0.20	0.91	1.67	0.22	0.56
DAG-GNN	0.89	0.53	0.05	0.67	0.44	0.04
NOTEARS	1.00	0.47	0.19	0.89	0.44	0.19

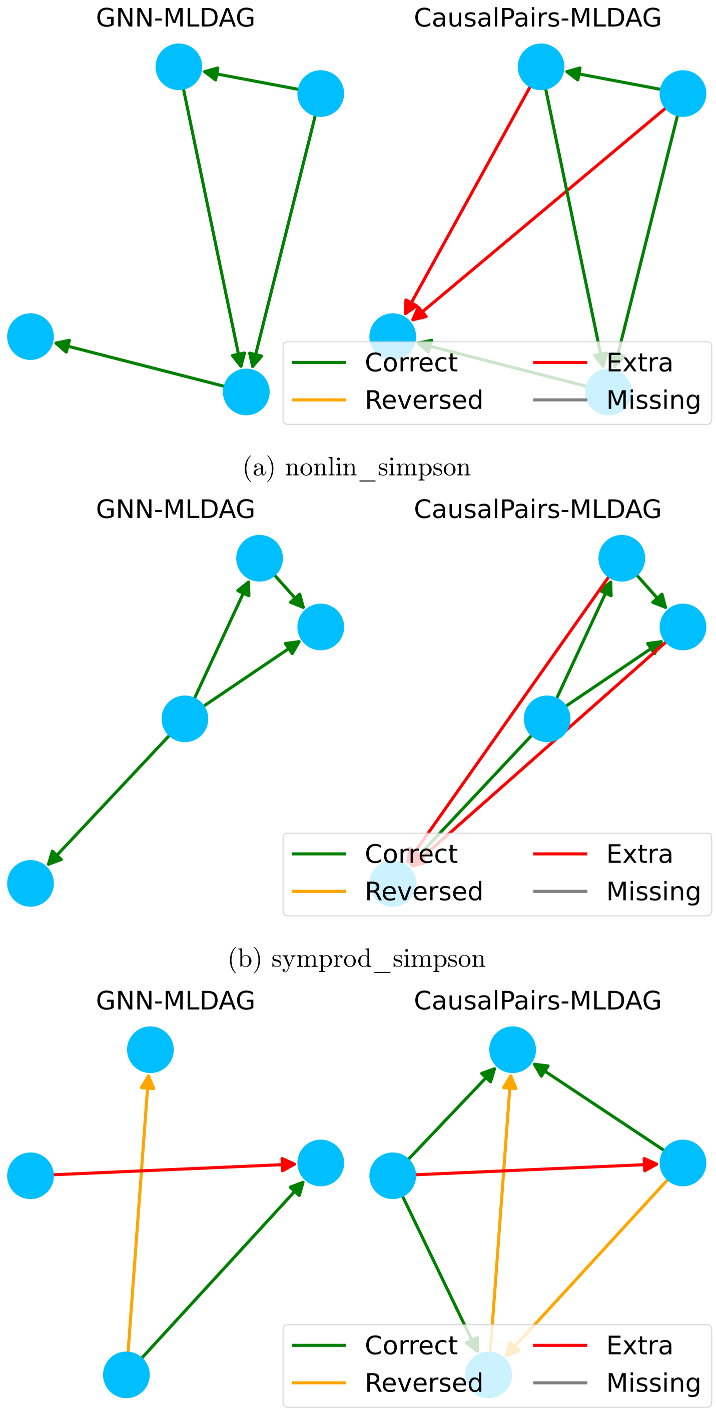
Tables 4.2 present the results of applying my methods to five datasets from the Microsoft CSuite. The GNN-based methods achieve significantly lower SHD, higher TPR, and lower FPR compared to all other methods, demonstrating the robustness and generalizability of the GNN-based framework across diverse datasets. Compared to the synthetic datasets presented in Table 4.1, the Microsoft CSuite datasets have fewer nodes and edges. Additionally, the three smaller datasets

from Microsoft CSuite allow us to demonstrate the method's capability to recover various graph structures learned directly from data.

In these datasets, which include graphs with four nodes and four edges, my methods accurately identified V structures such as $A \rightarrow B \leftarrow C$. This ability to capture fork or collider structures highlights the method's precision in determining causal directions and understanding interactions between variables. We also observed that in datasets like mixed_simpson and nonlin_simpson, with confounder structures such as $A \rightarrow B$ and $A \rightarrow C$, the methods demonstrated the ability to recognize common causes affecting multiple outcomes. Chain structures like $A \rightarrow B \rightarrow C$ were also accurately recovered, showcasing the capability to model sequential causal relationships. For instance, among two of these datasets, my GNN-based methods achieved a SHD score of 0 and a TPR score of 1, perfectly identifying the true graph, and validating the methods' effectiveness in learning complex causal structures directly from data.

Notably, as shown in Figure 4.3, the GNN-based methods not only identified the true graph structure but also avoided predicting extraneous edges. In contrast, while CausalPairs methods were able to identify the true edges, they also predicted all possible edges, leading to higher false positives. This underscores the precision of the GNN-based approach in distinguishing true causal relationships from spurious ones.

In Table 4.3, my methods, particularly GNN-PG and GNN-MLDAG, demonstrate strong performance on the real-world protein network dataset, accurately predicting edge counts. Notably, they outperform LiNGAM and GES in terms of correct edge predictions, and even match or surpass the performance of recent methods like NOTEARS-MLP and DAG-GNN. The incorporation of global structural information through GNNs enables accurate edge prediction, while my approach also shows improved directional accuracy, as evident from the lower number



(a) nonlin_simpson

(b) symprod_simpson

(c) mixed_simpson

Figure 4.3: Performance comparison between GNN-based methods and CausalPairs methods on smaller CSuite datasets: (a) nonlin_simpson, (b) symprod_simpson, and (c) mixed_simpson. The plots illustrate the number of correct, reversed, extra, and missing edges for each method with respect to ground truth graphs.

Table 4.3: Comparison of my GNN-based probabilistic methods with GES, LiNGAM, DAG-GNN 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 manuscripts [26, 88]. The edge probability model based on the GNN framework is trained on synthetic train data.

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

of reversed edges achieved by GNN-MLDAG and GNN-PG.

A notable aspect is that DAG-GNN and NOTEARS-MLP exhibit sensitivity to data scaling, with performance variations between standardized and non-standardized data. This sensitivity arises because their continuous optimization processes can be disrupted by changes in data magnitude and distribution, potentially losing important information related to the data's mean and variance. Additionally, LiNGAM, which is designed for non-Gaussian linear models, may struggle with the non-linear relationships present in the protein network dataset. In contrast, my GNN-based methods show consistent performance across both standardized and non-standardized datasets, demonstrating robustness to data scaling. This robustness is attributed to the effective capture and utilization of both local and global structural information by GNNs.

4.5 Summary

In this work, I introduce a probabilistic causal discovery framework that harnesses the power of Graph Neural Networks (GNNs). My results on synthetic and real-world datasets demonstrate the efficacy of my GNN-based approach, surpassing the previous CausalPairs methods across various graph types and densities. By leveraging global structural information, my method overcomes traditional limitations and enhances causal graph learning precision. My GNN-based methods significantly advance the state of causal discovery, effectively capturing complex dependencies through node and edge feature integration. This integration enables more accurate and reliable causal inference, showcasing GNNs' potential to enhance scalability and generalization. Moreover, the GNN-based framework represents a significant breakthrough in causal structure learning, offering improved performance. Future research directions will explore the integration of acyclicity constraints within the Graph Neural Network (GNN) framework, aiming to improve the model's robustness and accuracy by enforcing causal consistency. Exploring advanced GNN architectures may further elevate this approach's performance, expanding its applicability to diverse and complex datasets.

CHAPTER 5: CAUSAL MODELING OF SOCIAL MEDIA POLARIZATION: QUANTIFYING INFLUENCER EFFECTS ON AFFECTIVE POLARIZATION

5.1 Introduction

The transformative impact of social media on political communication cannot be overstated. As a platform for instantaneous information exchange, social media has redefined the way individuals engage with political content, debate policy, and form community bonds [98, 99]. However, this digital revolution has also given rise to a less auspicious phenomenon: affective polarization [28–30]. The term, affective polarization refers to the phenomenon where individuals feel more positively towards members of their political group while simultaneously harboring negative sentiments towards those of opposing groups [100, 101].

Affective polarization on social media is particularly pernicious due to its potential to erode the foundations of democratic discourse, replacing reasoned debate with hostile confrontation. It transforms disagreement into contempt, making compromise and consensus-building increasingly elusive [102, 103]. The implications of this trend extend beyond the digital realm, influencing real-world political engagement and the broader societal fabric [31, 32].

At the heart of this polarization are influential usersâindividuals and entities with significant followings and the ability to shape public discourse. These influencers can act as catalysts, amplifying existing divides or bridging gaps in understanding through their engagement with contentious topics [104, 105]. Nowadays, the reach and impact of such figures are greater than ever before, making it essential to closely examine their role in the process of affective polarization. [106].

In this study, we aim to measure the impact of conversations started by influential users on the polarization seen in Twitter/X discussions. Our paper makes the following contributions to understand how these influencers shape public sentiment and contribute to polarization in online communities.

- This research presents a new framework using counterfactual analysis to quantitatively measure how conversations led by influencers affect polarization on Twitter/X. By comparing polarization scores with and without these influential conversations, the study reveals their impact on public discourse.
- It offers a detailed analysis of how influential users shape emotional dynamics within contentious topics (e.g. gun control and climate change), providing quantifiable measures of their influence on affective polarization.

By identifying the influential figures and the factors that exacerbate or mitigate polarization, platform designers and policymakers can better navigate the challenges posed by this phenomenon and work towards a more informed and less divided public discourse.

5.2 Related Work

The phenomenon of affective polarization has been extensively documented in political psychology, with recent studies revealing its escalation on social media platforms [107, 108]. Affective polarization extends beyond ideological disagreements, encapsulating emotional responses that manifest as mutual dislike and distrust among those with opposing political allegiances [109].

Research highlights a complex interplay between media, political figures, and entrenched ideologies that magnifies societal divisions. Social media platforms enhance these divisions through algorithmically curated content that often promotes inflammatory and polarizing material [110–112]. Echo chambers, a topic of much debate, are often criticized for reinforcing ideological conformity and shielding users from opposing viewpoints, which can intensify affective polarization [113]. Conversely, some research suggests that even when individuals are exposed to contrary opinions, this exposure does not necessarily mitigate polarization and may, under certain conditions, actually exacerbate it [114, 115].

Advancements in quantifying affective polarization have led to the development of metrics that capture the emotional content and toxicity in social media interactions [31, 116]. Studies now classify users by political partisanship and analyze the emotions and language used in their communications, providing a subtle understanding of the affective component of polarization [100, 117]. This line of research highlights that negative emotions and toxicity are not randomly distributed but correlate with the network distance in social media interactions, suggesting structural properties of these networks influence the emotional tone of online discourse [118].

Kramer et al. [119] assert that emotional states are contagious so whether expressed by other people or appears on Newsfeed (a personalized stream of content provided by social media platforms) has a direct impact on our emotions. Another finding of their study is that as opposed to prevalent perception, non-verbal language, and inter-personal interactions are not required for contagion of emotions. The study by Cha et al. [120] explores influence patterns on social media and identifies the most important role-playing factors. Betts and Bliuc [121] shows that an influencer with extreme opinions will invariably accelerate the pace of polarization, and this impact grows in proportion to their influence and level of engagement. The impact of a neutral influencer, however, depends on the society's openness to differing viewpoints.

Despite extensive research on affective polarization, much of the current literature focuses on the consequences of polarization without a clear methodology to quantify its emergence and escalation directly from influencer interactions. In addition, there is a limited exploration into the specific role of influential users within these divisive dynamics, particularly in how their conversations shape public sentiment over time [106].

This research builds upon these findings by specifically examining how influential figures impact the affective landscape of online discourses. By systematically evaluating the presence and absence of high-profile conversations, I offer a unique perspective on the role influencers play in either mitigating or exacerbating affective polarization. Through this lens, I contribute to a deeper understanding of the dynamics at play in social media's political discourse to provide insights into digital communication strategies aimed at reducing polarization.

5.3 Methodology

The goal of this study is to methodically quantify the impact of influencers on affective polarization on social media platforms. To achieve this, I implement a counterfactual analysis framework, which involves constructing hypothetical scenarios to understand what might happen if certain influencer-led conversations did not occur. This approach, illustrated in Figure 5.1, allows us to isolate the effects of these conversations on polarization dynamics, providing a clearer picture of their influence. The subsequent subsections will provide an in-depth examination of the methodological approach and analytical techniques utilized in this study, including data collection strategies, interaction network construction, sentiment analysis, and polarization metric quantification, offering a detailed understanding of the procedures employed to assess the impact of influencers on online social dynamics.

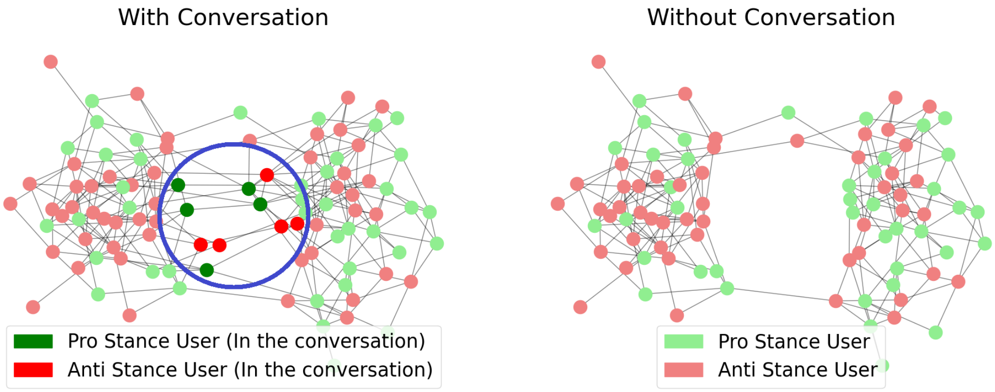

With Conversation
Without Conversation
Pro Stance User (In the conversation)
Anti Stance User (In the conversation)
Pro Stance User
Anti Stance User

Figure 5.1: Interaction networks with and without a specific conversation. The left panel displays the network including the conversation (circled), while the right panel shows the network without it, highlighting the dense areas where ongoing interactions among the influencer's followers persist through other conversations.

5.3.1 Data

5.3.1.1 Data Collection

For this research, I assembled a large-scale dataset from Twitter, focusing on tweets related to two contentious political issues that spark intense debate: climate change and gun control. I used specific Hashtags relevant to each topic to scrape the initial set of tweets using the Twitter API before its restrictions were imposed in 2023. To ensure a thorough capture of conversation cascades and user interactions, I expanded this data collection by recursively retrieving all referenced tweets linked to the initial set. This method allowed us to include all pertinent discussions, capturing the depth and breadth of conversations. User metadata such as the number of likes, replies, and retweets was also retrieved, as shown in Table 5.1.

Table 5.1: Dataset details

Dataset	**Timeline**	**Total Tweets (million)**	**Total Conversations (million)**	**Total Users (million)**	**#likes (billion)**	**#replies (million)**	**#retweets (billion)**
Gun Control	2022.01.01 - 2022.12.31	2.65	2.26	0.94	1.46	120.69	7.49
Climate Change	2021.06.01 - 2022.05.31	7.24	6.46	2.04	30.39	149.05	2.64

5.3.1.2 Data Exploration

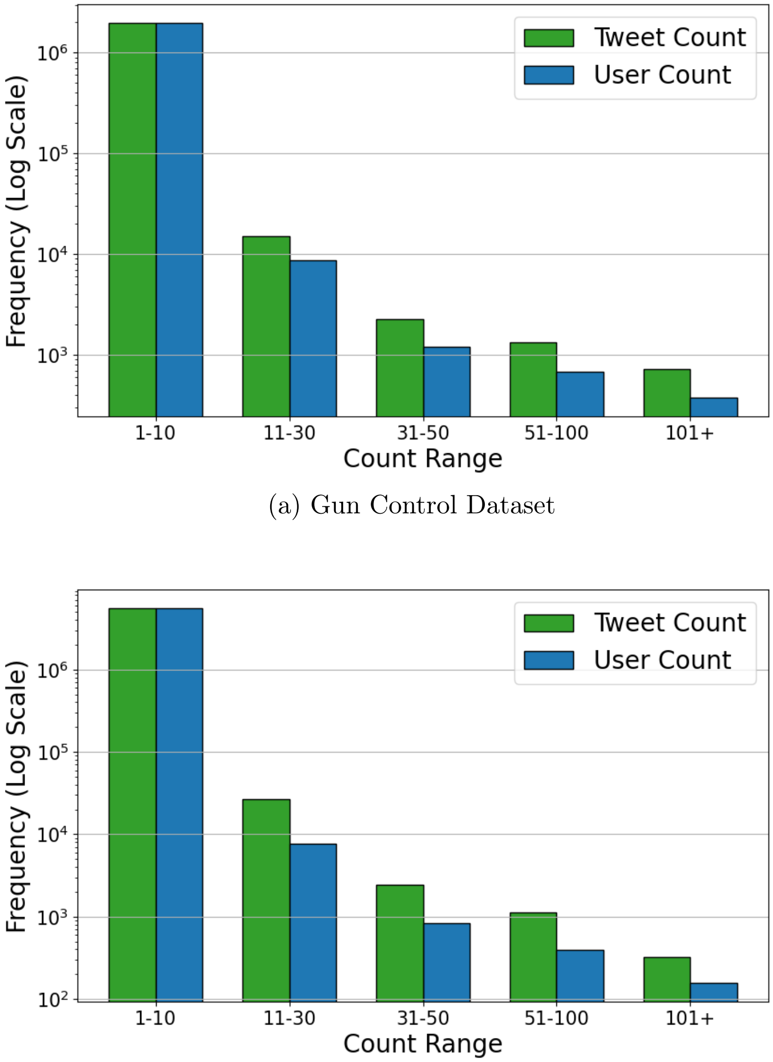
In analyzing the dataset, I identified a significant number of conversation threads with limited active engagement, defined as the number of distinct users replying within each thread. This measure focuses on direct interactions rather than passive activities like retweets and likes, aiming to capture meaningful exchanges. Many threads exhibited minimal interaction, often involving less than ten participants (Figure 5.2). To ensure the analytical robustness of this study, I established minimum engagement criteria based on discussions with domain experts. Only threads with at least 20 tweets and participation from at least 10 distinct users were included.

5.3.1.3 Influential Users

This methodology involved identifying influential users within each dataset based on their ability to engage substantial audience interactions, primarily measured by the total number of likes their posts received. Influential users were determined by sorting all users in descending order based on their like counts and selecting those at the top of this list (Table 5.2), which indicates a significant impact on online discourse. Although Table 5.2 showcases only the top five for illustrative purposes, this analysis included a broader set of influential figures, with 2,103 users for the dataset related to gun control and 3,701 users for climate change. I examined 5,500 gun control and 8,421 climate change conversations initiated by these users over a period of one year.

5.3.2 Quantifying Polarization with E/I Index

This study employs a multi-model sentiment analysis approach using VADER, BERTweet, and RoBERTa to evaluate the emotional content of tweets. This comprehensive sentiment profiling forms the basis for the subsequent analysis of affective polarization.



(a) Gun Control Dataset

(b) Climate Change Dataset

Figure 5.2: Bar chart distributions of conversation characteristics: (a) Gun Control and (b) Climate Change topics, displaying the frequency of conversations by number of tweets (left) and number of users (right).

Table 5.2: Top 5 influential users for each dataset

Dataset	Twitter Account	Tweet Count	Total Likes (million)	Total Retweets (million)	Total Replies (million)	Stance
Gun Control	User 1	1,241	272.40	25.55	19.69	anti
	User 2	1,665	66.79	10.35	10.77	pro
	User 3	1,750	37.74	4.75	2.75	pro
	User 4	1,760	31.75	5.07	4.15	pro
	User 5	746	22.81	5.25	1.15	pro
Climate Change	User 1	1,766	190.47	16.64	11.84	believe
	User 2	76	185.99	48.82	9.99	believe
	User 3	1,945	57.79	8.05	6.79	believe
	User 4	33	26.43	7.53	0.65	believe
	User 5	55	24.12	5.61	1.14	believe

I utilize the E/I Index methodology, adopted from Tyagi et al. [31], to quantify affective polarization by measuring the ratio of external (between-group) to internal (within-group) interactions within Twitter discourse networks. This method allows us to assess the degree of in-group cohesion and out-group engagement. The E/I Index is derived by evaluating the balance of positive and negative interactions within and between groups. I compute the difference between the E/I indices of positive and negative interactions to discern the polarization valence. For further details on the computations and equations underlying the E/I Index, I refer readers to the original work by Tyagi et al. [31].

This approach is akin to the P-Index proposed by Guerra et al. [122], albeit tailored specifically for Twitterâs unique interaction dynamics. Through this framework, I focus on:

- **In-group Solidarity:** Gauged through the frequency and sentiment of interactions within a group.

- **Out-group Engagement:** Characterized by the nature of interactions across different stance groups, often highlighting conflict or disagreement.

In practical applications, such as in gun control debates, this index helps us identify and describe patterns like Pro-Anti and Anti-Pro, indicating shifts from passive support or opposition to active advocacy or contention. While Bestvater et al. [123] caution against using sentiment as a standalone measure for polarization or stance, this study integrates these insights as part of a broader, multi-faceted approach to understanding affective polarization.

5.3.3 User Stance Labeling using Graph Neural Networks (GNNs)

My approach to user stance labeling employs a two-stage pipeline that integrates textual and social interaction data, as established in [124].

Initial Label Generation. The stance labeling process begins with the construction of a user-hashtag bipartite graph, where one set of nodes represents users and the other set represents hashtags used in their tweets. This graph forms the basis for applying a reciprocal label propagation algorithm, initially assigning stance labels based on users' engagements with specific hashtags. I identify seed labels from a subset of users who frequently use hashtags that are strongly associated with known stances. This preliminary method generates labels for approximately 500,000 gun control dataset users and 1.6 million climate change dataset users, providing an initial but incomplete picture of user stances. Social interactions among users are not factored into this analysis.

Expanding Stance Labeling with GNNs. To enhance and refine this stance labeling, I construct a comprehensive user-user interaction network. In this network, nodes represent users and edges signify direct interactions such as retweets or replies. Leveraging BERTweet, I embed the textual content of tweets into node features to capture linguistic nuances. Subsequently, I trained a GNN classifier (GraphSAGE) on the interaction graph with seed labels derived from the initial hashtag-based method. This two-stage approach integrates both textual content and interaction patterns to predict user stances more accurately, resulting in expanded stance labels for 2.6 million users in the gun control context and 4.7 million users in the climate change context. Moreover, the GNN predicts node labels by generating a probability distribution over possible stances for each user. Detailed information on the GNN's architecture, parameter settings, and performance evaluation can be found in the work by Melton et al. [124].

Optimization of Classification Thresholds. To more accurately delineate users into distinct stance groups, I adopted a dual-threshold strategy for classifying user stances based on their stance probability scores. I identified optimal thresholds, Threshold_1 and Threshold_2, through a comprehensive grid search methodology designed to refine the decision boundary by optimizing the F1 score. This process entailed meticulously comparing predicted labels against a benchmark set comprising both heuristic stances derived from the initial hashtag-based labeling and a subset of users manually annotated by domain experts. This dual-source validation approach ensured robustness in threshold selection by integrating empirical data with expert judgment. The final thresholds were determined as follows: Threshold_1 (≤ 0.40) for identifying 'pro' or 'believers' stances, and Threshold_2 (≥ 0.60) for 'anti' or 'disbelievers'. Users with probabilities falling between these thresholds were classified as 'undecided', and subsequently, their data were excluded from further analysis in both datasets.

5.3.4 Approach to Quantifying Influencer Impact on Affective Polarization

I center my methodology around the concept of a counterfactual scenario within a subgraph of influence, where polarization scores are calculated in scenarios both with and without specific influencer-led conversations, as illustrated in Figure 5.1. This comparative method is crucial for isolating the shifts in polarization attributable to these conversation networks, providing insights into how individual conversations can sway public sentiment.

To enhance my analysis, I focus on constructing subgraphs around specific influencers and their follower networks, recognizing that the comprehensive interaction graphs on platforms like Twitter, which involve millions of daily interactions, can obscure the effects of smaller-scale conversation networks. For instance, in my detailed case study of 'User X's Twitter interactions, I selectively identified a subset of approximately 1,000 followers from a larger pool of 15,000 active users. I constructed a subgraph that included these followers and their adjacent connections, encompassing around 2,000 users in total.

This subgraph methodology provides a focused lens through which we can observe and analyze the polarization dynamics more clearly. Initially, I compute the polarization within this subgraph by including 'User X's conversation. Subsequently, to distinctly understand the conversation's specific impact, I recalibrated the polarization after removing the 400 users directly involved in the conversation. This technique of contrasting polarization scores with and without the conversation offers a transparent view of how specific conversations influence the network's polarization dynamics.

By focusing on influencer-centric subgraphs, I uncover the subtle yet significant impacts of specific conversations on affective polarization, avoiding the dilution of findings by the broader network's noise.

5.4 Results

My analysis reveals a temporal sensitivity in polarization, underscored by fluctuations that correspond closely with real-world events and influencer-led conversations. Specifically, the temporal examinationâillustrated in Figures 5.3 and 5.4 highlights notable shifts in polarization within the contexts of gun control and climate change discourse.

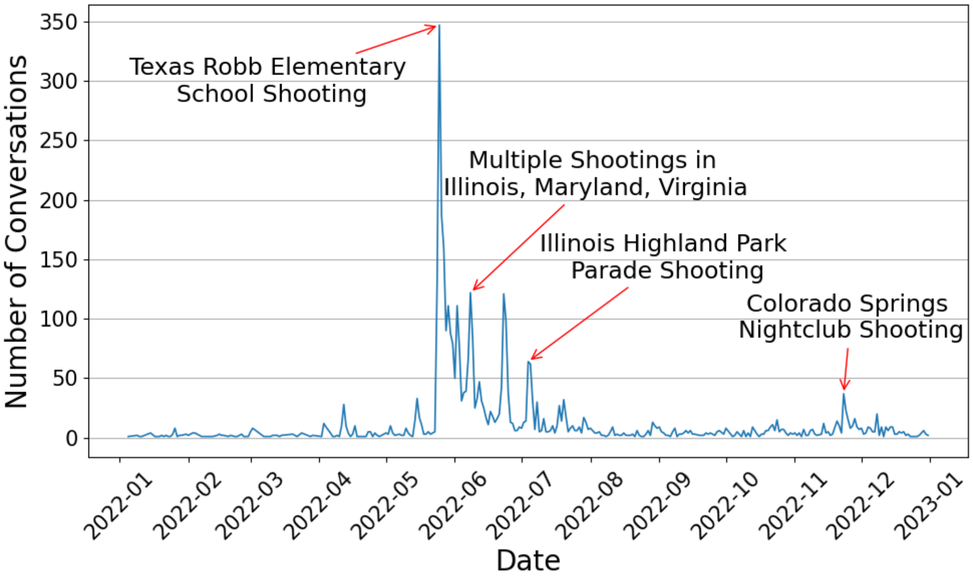
For instance, in Figure 5.3a during the summer of 2022, the gun control dataset exhibited significant spikes in the daily Twitter conversation, correlating with high-profile shooting incidents in the United States. This period, notably marked by the tragic Texas Robb Elementary School shooting, saw intensified influencer activity leading conversations that either amplified or attempted to bridge divides in public sentiment. Similarly, in the climate change discourse (Figure 5.3b), pivotal events such as the release of the Intergovernmental Panel on Climate Change (IPCC) report and the United Nations Climate Change Conference were mirrored by peaks, pointing to the reactive nature of online discourse to global climate events.

Following these events, Figure 5.4 illustrates how influential-led conversation changes the dynamic of polarization. For both, gun-control and climate-change datasets, we observe that when a certain event occurs, the polarization score increases compared to the situation it was before the event happened. Statistical significance tests confirm that many of these observed shifts are significant, reinforcing polarization's temporal alignment with external events. These observations highlight the critical role of influencers in steering the conversation and the powerful impact that influencers and real-world events can have affecting the polarization landscape.

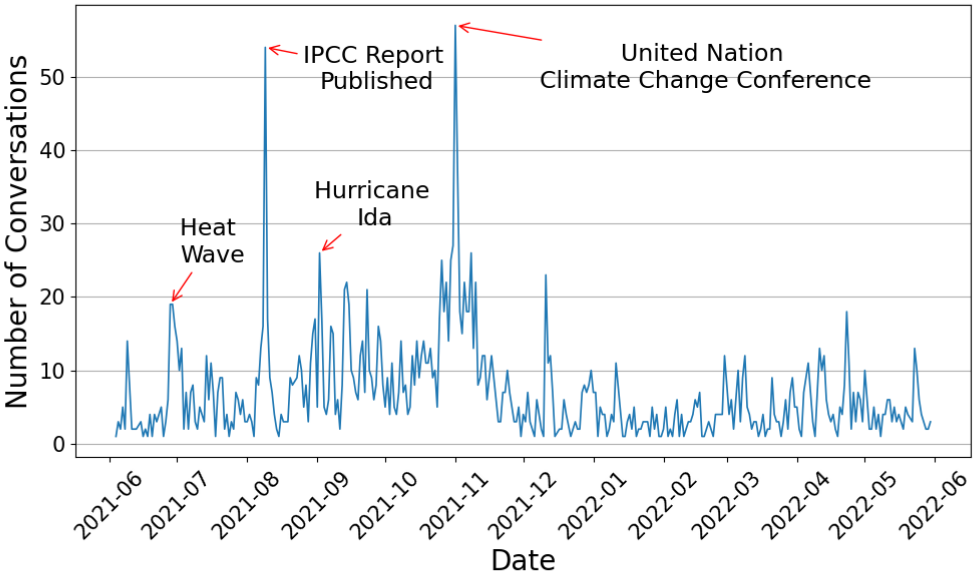
5.4.1 Effect of an Influential-led Conversation

The shifts in polarization are quantified by examining the changes in polarization scores when specific influencer-led conversation is either included or excluded from the analysis. This approach helps isolate the direct impact of a single conversation on the overall polarization score that day.

For example, as illustrated in Table 5.3, the removal of a conversation by 'User 3', a prominent disbeliever in climate change, resulted in a notable shift in the polarization scores. Specifically, the polarization score for the believer-to-disbeliever direction increased from 0.471 to 0.633, indicating a more pronounced divide when the disbeliever's influence was absent from the conversation. This change, representing a magnitude of 0.162 in the polarization score, underscores the significant

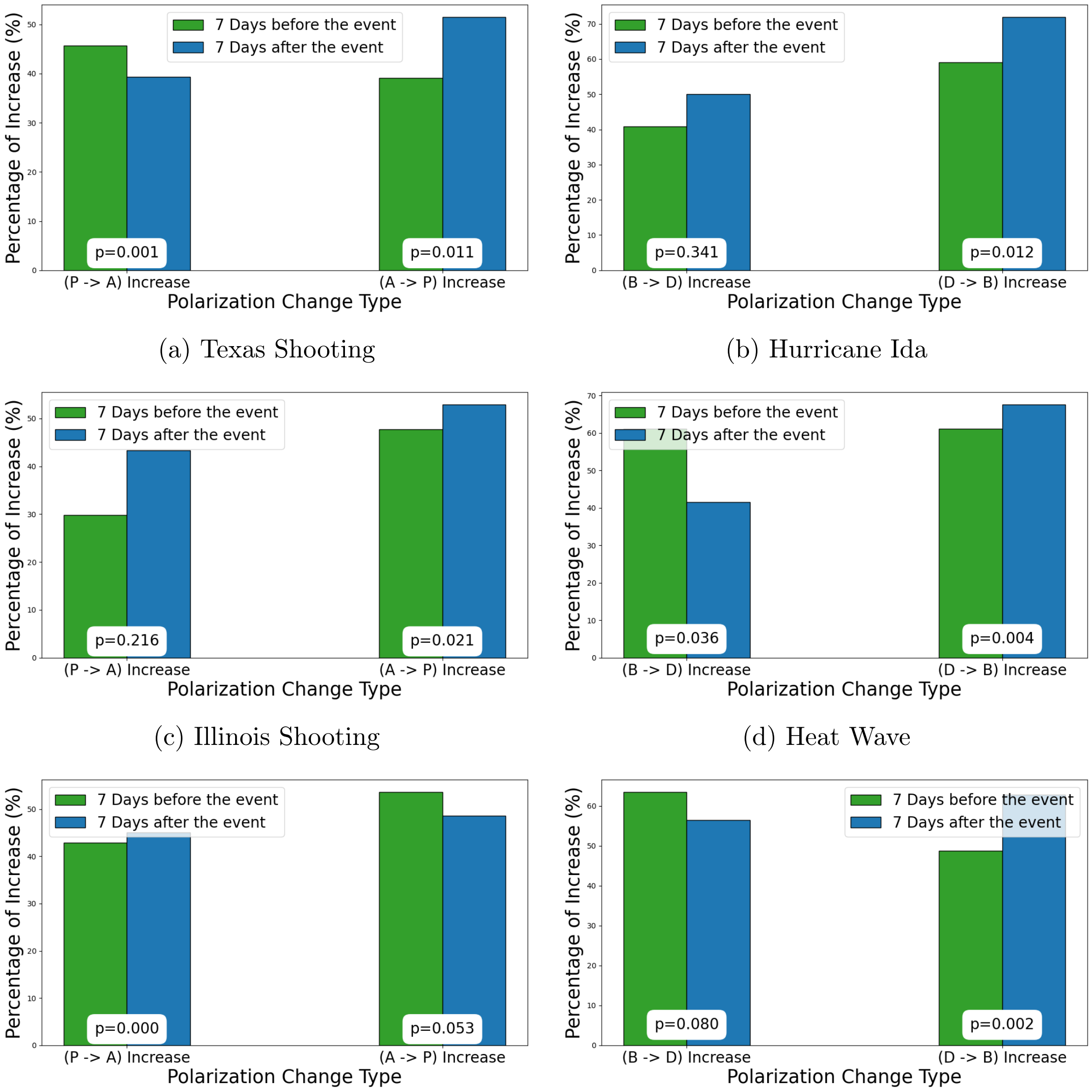


(a) Gun Control Dataset



(b) Climate Change Dataset

Figure 5.3: Frequency distributions of daily conversation counts over the specified time frame, stratified by topic: (a) Gun Control and (b) Climate Change.



(a) Texas Shooting

(b) Hurricane Ida

(c) Illinois Shooting

(d) Heat Wave

(e) Colorado Shooting

(f) IPCC Annual Report

Figure 5.4: Temporal dynamics of polarization scores in response to events. The left trio of subfigures (a,c,e) illustrates changes in polarization preceding and following events related to gun control, with '(P → A)' denoting shifts from pro-to-anti gun control stance and '(A → P)' denoting shifts from anti-to-pro stance. The right trio (b,d,f) depicts similar changes for climate change-related events, with '(B → D)' representing shifts from believe-to-disbelieve in climate change and '(D → B)' for shifts from disbelieve-to-believe stance direction. P-values indicate the statistical significance of changes, confirming that these shifts are not random but are influenced by the events.

role that influencer-led interactions play in shaping public sentiment and polarization.

Similarly, in the context of gun control as detailed in Table 5.4, the exclusion of a conversation led by an anti-gun control user (i.e. 'User 2') resulted in a decrease in the anti-pro polarization score. This demonstrates how conversations led by users opposed to gun control amplify polarization towards anti-gun stances when present, and reduce it when removed.

Table 5.3: Climate Change Dataset: The effect of a single conversation by influential users on daily polarization score.

User	User's Stance	Number of Tweets in the Conversation	Number of Followers in the Conversation (#disbelieve / #believe)	Number of Followers Interacting on That Day (#disbelieve / #believe)	Stance Direction	Polar. Score Without Conv.	Polar. Score With Conv.
User 1	believe	104	41 / 50	904 / 1,332	disbelieve → believe	-0.091	-0.135
					believe → disbelieve	0.463	0.398
User 2	believe	187	131 / 24	457 / 333	disbelieve → believe	0.115	0.185
					believe → disbelieve	0.378	0.402
User 3	disbelieve	157	68 / 14	302 / 100	disbelieve → believe	-0.137	-0.136
					believe → disbelieve	0.633	0.471
User 4	believe	134	51 / 37	356 / 356	disbelieve → believe	0.149	0.181
					believe → disbelieve	-0.228	-0.247
User 5	disbelieve	66	43 / 2	395 / 86	disbelieve → believe	0.144	0.152
					believe → disbelieve	0.248	0.237

Table 5.4: Gun Control: The effect of a single conversation by influential users on daily polarization score.

User	User's Stance	Number of Tweets in the Conversation	Number of Followers in the Conversation (#anti / #pro)	Number of Followers Interacting on That Day (#anti / #pro)	Stance Direction	Polar. Score Without Conv.	Polar. Score With Conv.
User 1	pro	746	334 / 325	1,211 / 1,040	anti → pro	0.147	0.252
					pro → anti	0.247	-0.200
User 2	anti	1,333	497 / 391	856 / 707	anti → pro	0.079	0.229
					pro → anti	0.220	-0.087
User 3	pro	57	20 / 22	312 / 356	anti → pro	0.085	0.099
					pro → anti	-0.022	0.121
User 4	anti	164	37 / 100	176 / 286	anti → pro	0.247	0.121
					pro → anti	0.204	0.241
User 5	pro	129	16 / 50	867 / 1,879	anti → pro	0.043	0.041
					pro → anti	0.069	0.085

5.4.2 Insight of Influential Users

In this section, I examine the collective influence of conversations initiated by influential users on shaping public sentiment and polarization across various themes. I employ a systematic approach to quantify the impact of these users by assessing the frequency and effect of their conversations on polarization shifts.

Tables 5.5 and 5.6 present data on influential users who actively engage in discussions and initiate several conversations related to gun control and climate change, respectively. These tables reveal the percentage of conversations that increase or decrease polarization, given a specified number of conversations within the context of these topics, thereby illuminating the effects of influential users on social media dynamics.

Table 5.5: Gun Control Dataset: Impact of conversation initialized by influential authors (Twitter users) on polarization scores. The '% increase' or '% decrease' denotes the percentage of cases, (how) polar score changes upon adding a conversation compared to the total number of conversations.

Twitter User	**User Type**	**Stance (Gun Control)**	**Number of Conversations**	**Polarization: Change (pro → anti)**	**Polarization: Change (anti → pro)**
User 1	Media Outlet	anti	5	100.0 % - decrease	100.0 % - increase
User 2	Media Outlet	pro	6	66.67 % - decrease	100.0 % - decrease
User 3	Commentator	pro	9	88.89 % - decrease	77.78% - decrease
User 4	Journalist	anti	18	88.89 % - decrease	61.11% - increase
User 5	Politician	pro	7	85.71 % - increase	85.71% - decrease
User 6	Politician	pro	14	64.29 % - decrease	78.57% - decrease
User 7	Legal Analyst	pro	38	76.32 % - decrease	71.05% - increase
User 8	Media Outlet	anti	12	66.67 % - increase	75.00% - increase
User 9	Attorney	anti	7	71.43 % - increase	71.43% - increase
User 10	Politician	pro	26	53.85 % - decrease	61.54% - decrease

Table 5.6: Climate Change Dataset: Impact of conversation initialized by influential authors (Twitter users) on polarization scores. The '% increase' or '% decrease' denotes the percentage of cases, (how) polar score changes upon adding a conversation compared to the total number of conversations.

Twitter User	**User Type**	**Stance (Climate Change)**	**Number of Conversations**	**Polarization: Change (believe → disbelieve)**	**Polarization: Change (disbelieve → believe)**
User 1	Scientific Org.	believe	5	60.00% - decrease	100.0 % - increase
User 2	Public Figure	disbelieve	5	80.00% - decrease	100.0 % - increase
User 3	Media Personality	disbelieve	8	50.00% - increase	87.5 % - decrease
User 4	Politician	believe	7	85.71% - decrease	57.14 % - decrease
User 5	Media Outlet	believe	41	51.22% - decrease	85.37 % - increase
User 6	Actor	believe	6	83.33% - decrease	50.0 % - increase
User 7	Climate Activist	believe	9	77.78% - increase	55.56 % - decrease
User 8	Media Personality	disbelieve	8	75.00% - increase	62.5 % - increase
User 9	Climate Advocate	believe	7	57.14% - decrease	71.43 % - decrease
User 10	Politician	believe	19	57.89% - decrease	63.16 % - increase

5.4.3 Insight of Conversations Within Stance-Group

Table 5.7 provides a more comprehensive view by analyzing all the conversations initiated by users within a certain stance. Interestingly, it is observed that conversations led by users with anti or disbeliever stances often result in decreased polarization compared to their opposing stance group. However, it's important to note the observational nature of our approach; while it reveals significant insights into the dynamics of influencer conversations, it does not assert direct causation. Overall, the insights gleaned from these influencer-led conversations across different thematic areas demonstrate the critical role of influential users in modulating the affective landscape of online communities.

Table 5.7: Details of conversations' effects based on stances for both gun control and climate change. The '% increase' or '% decrease' denotes the percentage of cases, (how) polar score changes upon adding a conversation compared to the total number of conversations.

Dataset	**Stance**	**Number of Conversations**	**Polarization Score: Change**	
Gun Control			**(pro → anti)**	**(anti → pro)**
	pro	1,940	63.40% - decrease	52.99% - increase
	anti	1,704	55.52% - decrease	53.28% - decrease
Climate Change			**(believe → disbelieve)**	**(disbelieve → believe)**
	believe	1,561	50.99% - decrease	64.25% - increase
	disbelieve	596	50.67% - decrease	58.22% - decrease

5.5 Summary

This study provides a detailed analysis of the mechanics of affective polarization on social media, particularly highlighting the crucial role of influential users in modulating public sentiment on platforms like Twitter. By implementing a counterfactual framework, which evaluates scenarios with and without specific influencer-led conversations, this work offers a unique methodological contribution that allows us to isolate and quantify the influence of such conversations on the dynamics of polarization. The findings demonstrate that influencers have the potential to amplify or mitigate divisiveness across polarizing topics like gun control and climate change. It reveals how subtle shifts in influencer-driven dialogues can significantly affect public discourse by employing a comparative analysis coupled with computational techniques, such as subgraph construction. This approach deepens our understanding of how specific conversations impact polarization that are critical to social media dynamics.

Despite these promising results, this study faced methodological challenges, particularly due to the limitations inherent in calculating the polarization scores based on sentiment analysis. This depends on the presence of positive or negative words in tweets and it is negatively impacted by the sporadic nature of meaningful social media interactions. These challenges underscore the complexity of capturing the sophisticated landscape of online discourse and point to the need for developing more refined metrics and methodologies that leverage current advancements in language understanding that go beyond sentiment analysis.

CHAPTER 6: CONCLUSIONS

In this dissertation, I focused on the field of causal structure learning, which is a crucial aspect of causal discovery. Specifically, I proposed to go beyond the naive approach of generating graph probabilities from causal pair probabilities by enforcing acyclicity and approximating its solution using the maximum spanning directed acyclic graph approach. I showed that leveraging GNNs and enforcing acyclicity improved performance on both synthetic and real datasets compared with the causal pairs and traditional approach and had statistically better and/or similar performance than some state-of-the-art methods. I discussed the challenges associated with this task, such as the high dimensionality of data, computational complexity, and the presence of confounding variables.

Furthermore, I introduced the causal feature selection (CFS) method to select informative and relevant features from observational data, using causal graphs to provide a unique advantage over traditional correlation-based metrics. I demonstrated the effectiveness of the proposed CFS method with new evaluation criteria on synthetic and real-world datasets of various domains and compared its performance with other baseline and traditional approaches. This approach can help reduce the dimensionality of the data and eliminate confounding variables, leading to more accurate and reliable causal models.

Overall, the contributions of this dissertation include:

- A comprehensive review of the field of causal structure learning, including causal inference and causal discovery.

- A novel causal graph learning approach with probabilistic information using cause-effect pairs

- A proposed approach of causal feature selection method that incorporates the causal relationships between variables in the selection process.
- A novel evaluation criteria for causal feature selection using causal metrics
- The introduction of a GNN-based probabilistic framework for causal discovery.
- An application of counterfactual analysis to study affective polarization on social media.

6.1 Future Work

One promising area of research is to investigate the combination of causal feature selection with other correlation-based methods, such as LiNGAM and mRMR, to improve the accuracy and efficiency of selecting causal features, especially in high-dimensional datasets. It would also be interesting to explore the applications of causal structure learning and causal feature selection in various domains e.g. public health, to inform decision-making processes.

Specifically, I plan to apply my machine learning and causal modeling expertise to cancer research. Building on the methodologies developed in this dissertation, I will use causal discovery techniques to analyze patient data can identify causal links between genetic mutations, lifestyle factors, and cancer progression. This aligns with my upcoming role at the University of Tennessee Health Science Center (UTHSC), where I will focus on predicting cancer outcomes by integrating clinical and social determinants of health datasets. These efforts aim to enhance personalized treatment strategies and improve equity in cancer care.

Furthermore, extending from my research on polarization, I aim to develop methods to analyze how misinformation spreads and influences public opinion, identifying causal relationships between influencer activities and the dissemination of misinformation. To address this, I will collect and analyze data from social media posts, shares, comments, and user engagement metrics to identify patterns and sources of misinformation. Additionally, I will gather data from news articles, fact-checking organizations, and user reports. Using techniques such as large language models and graph neural networks, I plan to develop algorithms to detect misinformation based on linguistic cues, network analysis, and source credibility. This research will provide actionable insights for mitigating the harmful impacts of misinformation and fostering healthier online discourse.

Overall, there is still significant potential for research in the field of causal modeling and causal structure learning. Future research in these areas promises to make substantial contributions to various fields and deepen our understanding of causality.

REFERENCES

[1] J. Pearl, *Causality.* Cambridge university press, 2009.

[2] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, p. e253, 2017.

[3] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

[4] T. Farrand, "How to understand the world of causality," Feb 2023.

[5] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search.* MIT press, 2000.

[6] J. Pearl and T. S. Verma, "A theory of inferred causation," in *Studies in Logic and the Foundations of Mathematics*, vol. 134, pp. 789–811, Elsevier, 1995.

[7] D. Colombo, M. H. Maathuis, *et al.*, "Order-independent constraint-based causal structure learning.," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3741–3782, 2014.

[8] X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu, "A kernel-based causal learning algorithm," in *Proceedings of the 24th international conference on Machine learning*, pp. 855–862, 2007.

[9] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *arXiv preprint arXiv:1202.3775*, 2012.

[10] A. Medvedovsky, V. Bafna, U. Zwick, and R. Sharan, "An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks," in *Algorithms in Bioinformatics: 8th International Workshop, WABI 2008, Karlsruhe, Germany, September 15-19, 2008. Proceedings 8*, pp. 222–232, Springer, 2008.

[11] K. Singh, G. Gupta, L. Vig, G. Shroff, and P. Agarwal, "Deep convolutional neural networks for pairwise causality," *arXiv preprint arXiv:1701.00597*, 2017.

[12] O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, K. Srinivas, M. Perrone, S. Sohrabi, and M. Katz, "Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts.," in *IJCAI*, pp. 5003–5009, 2019.

[13] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[14] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, pp. 31–78, 2006.

[15] S. Zhu, I. Ng, and Z. Chen, "Causal discovery with reinforcement learning," *arXiv preprint arXiv:1906.04477*, 2019.

[16] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–36, 2020.

[17] T. Gao and Q. Ji, "Efficient markov blanket discovery and its application," *IEEE transactions on Cybernetics*, vol. 47, no. 5, pp. 1169–1179, 2016.

[18] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation," *Journal of Machine Learning Research*, vol. 11, no. 7, pp. 171–234, 2010.

[19] X. Zhang, Y. Hu, K. Xie, S. Wang, E. Ngai, and M. Liu, "A causal feature selection algorithm for stock prediction modeling," *Neurocomputing*, vol. 142, pp. 48–59, 2014.

[20] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[21] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[22] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[23] L. Waikhom and R. Patgiri, "A survey of graph neural networks in various learning paradigms: methods, applications, and challenges," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6295–6364, 2023.

[24] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[25] G. Lacerda, P. L. Spirtes, J. Ramsey, and P. O. Hoyer, "Discovering cyclic causal models by independent components analysis," *arXiv preprint arXiv:1206.3273*, 2012.

[26] Y. Yu, J. Chen, T. Gao, and M. Yu, "Dag-gnn: Dag structure learning with graph neural networks," in *International Conference on Machine Learning*, pp. 7154–7163, PMLR, 2019.

[27] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21865–21877, 2020.

[28] O. Yair, "A note on the affective polarization literature," *Available at SSRN 3771264*, 2020.

[29] X. Yu, M. Wojcieszak, and A. Casas, "Affective polarization on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users," 2021.

[30] T. J. Rudolph and M. J. Hetherington, "Affective polarization in political and nonpolitical settings," *International Journal of Public Opinion Research*, vol. 33, no. 3, pp. 591–606, 2021.

[31] A. Tyagi, J. Uyheng, and K. M. Carley, "Affective polarization in online climate change discourse on twitter," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 443–447, IEEE, 2020.

[32] J. Beel, T. Xiang, S. Soni, and D. Yang, "Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 32–42, 2022.

[33] R. Rashid, J. Chowdhury, and G. Terejanu, "From causal pairs to causal graphs," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 802–807, 2022.

[34] R. Rashid, J. Chowdhury, and G. Terejanu, "Causal feature selection: Methods and a novel causal metric evaluation framework," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–9, IEEE, 2023.

[35] R. Rashid and G. Terejanu, "Graph neural networks for probabilistic causal discovery," in *9th Causal Inference Workshop at UAI 2024*, 2024.

[36] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International journal of data science and analytics*, vol. 3, pp. 121–129, 2017.

[37] P. Nandy, A. Hauser, and M. H. Maathuis, "High-dimensional consistency in score-based and hybrid structure learning," *The Annals of Statistics*, vol. 46, no. 6A, pp. 3151–3183, 2018.

[38] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[39] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, "Learning sparse nonparametric dags," in *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425, PMLR, 2020.

[40] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and dag constraints for learning linear dags," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17943–17954, 2020.

[41] I. Guyon, O. Goudet, and D. Kalainathan, "Evaluation methods of cause-effect pairs," *Cause Effect Pairs in Machine Learning*, pp. 27–99, 2019.

[42] J. A. Fonollosa, "Conditional distribution variability measures for causality detection," *Cause Effect Pairs in Machine Learning*, pp. 339–347, 2019.

[43] O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf, "Probabilistic latent variable models for distinguishing between cause and effect," *Advances in neural information processing systems*, vol. 23, 2010.

[44] N. Schluter, "On maximum spanning dag algorithms for semantic dag parsing," in *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pp. 61–65, 2014.

[45] R. McDonald and F. Pereira, "Online learning of approximate dependency parsing algorithms," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 81–88, 2006.

[46] A. Reisach, C. Seiler, and S. Weichwald, "Beware of the simulated dag! causal discovery benchmarks may be easy to game," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27772–27784, 2021.

[47] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[48] S. Ranka and V. Singh, "Clouds: A decision tree classifier for large datasets," in *Proceedings of the 4th knowledge discovery and data mining conference*, vol. 2, 1998.

[49] R. Jin and G. Agrawal, "Communication and memory efficient parallel decision tree construction," in *SDM*, 2003.

[50] Y. Hua, “An efficient traffic classification scheme using embedded feature selection and lightgbm,” in *2020 Information Communication Technologies Conference (ICTC)*, pp. 125–130, IEEE, 2020.

[51] M. Ali, “Pycaret: An open source, low-code machine learning library in python,” *PyCaret version*, vol. 2, 2020.

[52] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[53] Y. Zhai, Y.-S. Ong, and I. W. Tsang, “The emerging" big dimensionality",” *IEEE Computational Intelligence Magazine*, vol. 9, no. 3, pp. 14–26, 2014.

[54] V. Kumar and S. Minz, “Feature selection: a literature review,” *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[55] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[56] H. Liu and H. Motoda, *Computational methods of feature selection.* CRC press, 2007.

[57] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: a study on high-dimensional spaces,” *Knowledge and information systems*, vol. 12, pp. 95–116, 2007.

[58] I. Guyon, C. Aliferis, *et al.*, “Causal feature selection,” in *Computational methods of feature selection*, pp. 79–102, Chapman and Hall/CRC, 2007.

[59] K. Yu, M. Cai, X. Wu, L. Liu, and J. Li, “Multilabel feature selection: a local causal structure learning approach,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[60] K. Yu, L. Liu, J. Li, W. Ding, and T. D. Le, “Multi-source causal feature selection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2240–2256, 2019.

[61] J. Peters, P. Bühlmann, and N. Meinshausen, “Causal inference by using invariant prediction: identification and confidence intervals,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 947–1012, 2016.

[62] M. Paul, “Feature selection as causal inference: Experiments with text classification,” in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 163–172, 2017.

[63] A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205, Ieee, 2015.

[64] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, “Using causal discovery for feature selection in multivariate numerical time series,” *Mach. Learn.*, vol. 101, p. 377â395, oct 2015.

[65] A. Limshuebchuey, R. Duangsoithong, and T. Windeatt, “Redundant feature identification and redundancy analysis for causal feature selection,” in *2015 8th Biomedical Engineering International Conference (BMEiCON)*, pp. 1–5, IEEE, 2015.

[66] Y.-W. Chang and C.-J. Lin, “Feature ranking using linear svm,” in *Causation and prediction challenge*, pp. 53–64, PMLR, 2008.

[67] P. Panda, S. S. Kancheti, and V. N. Balasubramanian, “Instance-wise causal feature selection for model interpretation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1756–1759, 2021.

[68] J. Pearl, *Causality.* Cambridge university press, 2009.

[69] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, “Finding a" kneedle" in a haystack: Detecting knee points in system behavior,” in *2011 31st international conference on distributed computing systems workshops*, pp. 166–171, IEEE, 2011.

[70] W. Webber, A. Moffat, and J. Zobel, “A similarity measure for indefinite rankings,” *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 4, pp. 1–38, 2010.

[71] Z. Zhao, R. Anand, and M. Wang, “Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform,” in *2019 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 442–452, IEEE, 2019.

[72] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, “Mifs-nd: A mutual information-based feature selection method,” *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371–6385, 2014.

[73] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[74] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, “A linear non-gaussian acyclic model for causal discovery.,” *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[75] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. P. Pellet, P. Spirtes, and A. Statnikov, “Causality workbench,” in *Causality in the sciences*, Oxford University Press, 2011.

[76] A. Asuncion and D. Newman, “Uci machine learning repository,” 2007.

[77] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[78] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[79] S. Ott, S. Imoto, and S. Miyano, “Finding optimal models for small gene networks,” in *Biocomputing 2004*, pp. 557–567, World Scientific, 2003.

[80] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search.* MIT press, 2001.

[81] D. M. Chickering, “Optimal structure identification with greedy search,” *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[82] D. Heckerman, D. Geiger, and D. M. Chickering, “Learning bayesian networks: The combination of knowledge and statistical data,” *Machine learning*, vol. 20, pp. 197–243, 1995.

[83] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, “Dags with no tears: Continuous optimization for structure learning,” *Advances in neural information processing systems*, vol. 31, 2018.

[84] R. R. Bouckaert, “Probabilistic network construction using the minimum description length principle,” in *European conference on symbolic and quantitative approaches to reasoning and uncertainty*, pp. 41–48, Springer, 1993.

[85] J. A. Gámez, J. L. Mateo, and J. M. Puerta, “Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood,” *Data Mining and Knowledge Discovery*, vol. 22, pp. 106–148, 2011.

[86] A. Mohammadi and E. C. Wit, “Bayesian structure learning in sparse gaussian graphical models,” 2015.

[87] K. Mohan, M. Chung, S. Han, D. Witten, S.-I. Lee, and M. Fazel, “Structured learning of gaussian graphical models,” *Advances in neural information processing systems*, vol. 25, 2012.

[88] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, “Learning sparse nonparametric dags,” in *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425, Pmlr, 2020.

[89] A. Reisach, C. Seiler, and S. Weichwald, “Beware of the simulated dag! causal discovery benchmarks may be easy to game,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27772–27784, 2021.

[90] H. Gao, C. Yao, J. Li, L. Si, Y. Jin, F. Wu, C. Zheng, and H. Liu, “Rethinking causal relationships learning in graph neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 12145–12154, 2024.

[91] S. Zhao, I. Prapas, I. Karasante, Z. Xiong, I. Papoutsis, G. Camps-Valls, and X. X. Zhu, “Causal graph neural networks for wildfire danger prediction,” *arXiv preprint arXiv:2403.08414*, 2024.

[92] M. Zečević, D. S. Dhami, P. Veličković, and K. Kersting, “Relating graph neural networks to structural causal models,” *arXiv preprint arXiv:2109.04173*, 2021.

[93] H. Li, Q. Xiao, and J. Tian, “Supervised whole dag causal discovery,” *arXiv preprint arXiv:2006.04697*, 2020.

[94] I. Ng, S. Zhu, Z. Chen, and Z. Fang, “A graph autoencoder approach to causal structure learning,” *arXiv preprint arXiv:1911.07420*, 2019.

[95] W. Lin, H. Lan, and B. Li, “Generative causal explanations for graph neural networks,” in *International Conference on Machine Learning*, pp. 6666–6679, PMLR, 2021.

[96] R. McDonald and F. Pereira, “Online learning of approximate dependency parsing algorithms,” in *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 81–88, 2006.

[97] T. Geffner, J. Antoran, A. Foster, W. Gong, C. Ma, E. Kiciman, A. Sharma, A. Lamb, M. Kukla, N. Pawlowski, M. Allamanis, and C. Zhang, “Deep end-to-end causal inference,” *arXiv preprint arXiv:2202.02195*, 2022.

[98] P. Dahlgren, *Media and political engagement: Citizens, communication and democracy.* Cambridge University Press, 2009.

[99] D. Chambers, *Social media and personal relationships: Online intimacies and networked friendship.* Springer, 2013.

[100] D. Feldman, A. Rao, Z. He, and K. Lerman, “Affective polarization in social networks,” *arXiv preprint arXiv:2310.18553*, 2023.

[101] J. Serrano-Puche, “Digital disinformation and emotions: exploring the social risks of affective polarization,” *International Review of Sociology*, vol. 31, no. 2, pp. 231–245, 2021.

[102] T. O. Harel, J. K. Jameson, and I. Maoz, “The normalization of hatred: Identity, affective polarization, and dehumanization on facebook in the context of intractable political conflict,” *Social Media+ Society*, vol. 6, no. 2, p. 2056305120913983, 2020.

[103] L. Jenke, “Affective polarization and misinformation belief,” *Political Behavior*, pp. 1–60, 2023.

[104] B. K. Johnson, R. L. Neo, M. E. Heijnen, L. Smits, and C. van Veen, “Issues, involvement, and influence: Effects of selective exposure and sharing on polarization and participation,” *Computers in Human Behavior*, vol. 104, p. 106155, 2020.

[105] S. Balietti, L. Getoor, D. G. Goldstein, and D. J. Watts, “Reducing opinion polarization: Effects of exposure to similar people with differing political views,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 52, p. e2112552118, 2021.

[106] R. Recuero, G. Zago, and F. Soares, “Using social network analysis and social capital to identify user roles on polarized political conversations on twitter,” *Social media+ society*, vol. 5, no. 2, p. 2056305119848745, 2019.

[107] P. Törnberg, C. Andersson, K. Lindgren, and S. Banisch, “Modeling the emergence of affective polarization in the social media society,” *PLoS One*, vol. 16, no. 10, p. e0258259, 2021.

[108] N. Gillani, A. Yuan, M. Saveski, S. Vosoughi, and D. Roy, “Me, my echo chamber, and i: introspection on social media polarization,” in *Proceedings of the 2018 World Wide Web Conference*, pp. 823–831, 2018.

[109] J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan, “(mis) estimating affective polarization,” *The Journal of Politics*, vol. 84, no. 2, pp. 1106–1117, 2022.

[110] M. Suárez Estrada, Y. Juarez, and C. Piña-García, “Toxic social media: Affective polarization after feminist protests,” *Social Media+ Society*, vol. 8, no. 2, p. 20563051221098343, 2022.

[111] K. A. Heatherly, Y. Lu, and J. K. Lee, “Filtering out the other side? cross-cutting and like-minded discussions on social networking sites,” *New Media & Society*, vol. 19, no. 8, pp. 1271–1289, 2017.

[112] L. Mason, “A cross-cutting calm: How social sorting drives affective polarization,” *Public Opinion Quarterly*, vol. 80, no. S1, pp. 351–377, 2016.

[113] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, “The echo chamber effect on social media,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, p. e2023301118, 2021.

[114] S. Lee, H. Rojas, and M. Yamamoto, “Social media, messaging apps, and affective polarization in the united states and japan,” *Mass Communication and Society*, vol. 25, no. 5, pp. 673–697, 2022.

[115] S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood, “The origins and consequences of affective polarization in the united states,” *Annual review of political science*, vol. 22, pp. 129–146, 2019.

[116] A. Tyagi, J. Uyheng, and K. M. Carley, “Heated conversations in a warming world: affective polarization in online climate change discourse follows real-world climate anomalies,” *Social Network Analysis and Mining*, vol. 11, pp. 1–12, 2021.

[117] E. C. Connors, “Social desirability and affective polarization,” *Public Opinion Quarterly*, vol. 87, no. 4, pp. 911–934, 2023.

[118] M. Nordbrandt, “Affective polarization in the digital age: Testing the direction of the relationship between social media and usersâ feelings for out-group parties,” *New media & society*, vol. 25, no. 12, pp. 3392–3411, 2023.

[119] A. D. Kramer, J. E. Guillory, and J. T. Hancock, “Experimental evidence of massive-scale emotional contagion through social networks,” *Proceedings of the National academy of Sciences of the United States of America*, vol. 111, no. 24, p. 8788, 2014.

[120] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” in *Proceedings of the international AAAI conference on web and social media*, vol. 4, pp. 10–17, 2010.

[121] J. M. Betts and A.-M. Bliuc, “The effect of influencers on societal polarization,” in *2022 Winter Simulation Conference (WSC)*, pp. 370–381, IEEE, 2022.

[122] P. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg, “A measure of polarization on social media networks based on community boundaries,” in *Proceedings of the international AAAI conference on web and social media*, vol. 7, pp. 215–224, 2013.

[123] S. E. Bestvater and B. L. Monroe, “Sentiment is not stance: Target-aware opinion classification for political text analysis,” *Political Analysis*, vol. 31, no. 2, pp. 235–256, 2023.

[124] J. Melton, S. Reid, G. Terejanu, and S. Krishnan, “Two-stage stance labeling: User-hashtag heuristics with graph neural networks,” *arXiv preprint arXiv:2404.10228*, 2024.

APPENDIX A: Implementation of Methods Used in Numerical Results

To implement the PC and GES algorithms in my study, I have considered the publicly available implementations for both algorithms. I used the code from the following git repositories for the implementations.

- PC: https://github.com/keiichishima/pcalg
- GES: https://github.com/juangamella/ges

In addition, for the NOTEARS-MLP [39] approach, I followed the implementation stated in the paper and git repository: https://github.com/xunzheng/notears.

As for LiNGAM, I followed the Python package from the LiNGAM library: https://lingam.readthedocs.io/en/latest/index.html

Furthermore, for the DAG-GNN approach, I followed the implementation from the git repository: https://github.com/fishmoon1234/DAG-GNN

I used the default settings and default hyper-parameters for all these three implementations.

To implement the InfoGain and Lasso regularization, I have followed the scikit learn implementation.

APPENDIX B: Sachs Protein Network Dataset

The protein network presented by Sachs et al. [52] aims to investigate the signaling interactions among proteins involved in a cellular immune response. The authors measured the phosphorylation levels of 11 proteins over time, which reflect the degree of protein activation. This phosphorylation is a process that can impact protein activity, stability, and interactions with other molecules. In the network, each node represents a protein, and the connections between nodes depict the causal relationships among them.

To identify the causal relationships among the proteins the authors performed a combination of experimental perturbations and Bayesian structure learning. These perturbed experiments involved manipulating the activity or expression of proteins to observe the resulting changes in phosphorylation levels to identify causal influences. The authors collected data on the phosphorylation levels of the proteins before and after the perturbations. Next, the author applied Bayesian network analysis to infer the causal relationships. Based on the inferred causal relationships, the authors constructed a causal network that represents how the activation of one protein influences the activation of others. Therefore when I refer to one node causing another within the causal network, it means that the activation of a particular protein (the cause) leads to observable changes in the activation of another protein (the effect) within the protein signaling network.

This protein network dataset is considered a trusted benchmark in causality due to its rigorous experimental design, comprehensive measurements, and extensive validation. The network inferred from the data represents a model average from 500 high-scoring results with high-confidence arcs that appear in at least 85% of the networks. Besides, the authors were able to correctly infer the direction of causal influences in almost all cases (one exception, in which case the direction was inferred in the reverse order) that align with the consensus domain expertise.