Causal Analysis in Spatiotemporal Sea-Surface Temperature Data

Evan Vera[1], Wei Xu[2]
[1]Department of Computer Science, Cornell University – Ithaca, NY 14853
[2]Computational Science Initiative, Brookhaven National Laboratory – Upton, NY 11973
e-mail: evera@bnl.gov, xuw@bnl.gov

January 8, 2024

Abstract **– Sea-surface temperatures are one of the important metrics to model the ocean system that is useful for predicting weather patterns and observing changes in climate on a large scale. However, machine learning-based models which seek to predict future climate patterns suffer from a lack of transparency, with no clear indication of how related factors are considered in their analysis and predictions. These models also do not make causal inferences about the variables they model. In this project, we develop two branches of causal analysis that allow us to infer how the modeled variables affect future predictions. These methods are tested using causal discovery benchmark datasets. They not only allow for a better understanding of how predictions are made but also give valuable information about the real ocean system, supporting BNL's mission of better understanding Earth's climate.**

Keywords **– Causality, Causal Inference (CI), Granger-causality (GC), Time Series, Structural Causal Model (SCM)**

1 INTRODUCTION

Predicting and understanding the relationships between sea-surface temperatures over time has enormous potential for improving the accuracy of weather forecasts, evaluating the impacts of climate change, and highlighting areas that have a large impact on future climate patterns. Finding the causal linkage among these time series is essential to guide the design of machine learning surrogates for predicting climate patterns, and can be used to compare the difference between simulations and observations.

The most commonly taken definition of causality is the process by which one event contributes to the occurrence of another event. The problem with this notion in the sciences is that it lacks a rigorous definition, and thus the tools of statistics cannot be used to infer it experimentally or observationally. In this project, we operationalize this intuitive notion using various rigorous definitions, and then attempt to infer the causal relationships among time series datasets provided by the CauseMe causal benchmark platform, with the eventual goal of these methods being used on pre-industrial control (piControl) data generated by the Community Earth System Model (CESM2).

We began our causal discovery framework by inferring causality using Granger causality, developed by economist Clive Granger. This definition uses the intuitive notions that cause should precede effect (which we will assume throughout this paper), and that causes should contain relevant information about their effects. Thus, in order to detect Granger causality we see whether our hypothesized causal variable provides predictive power to the optimal predictor of our response variable. We consider two functional forms for the vector autoregressive model (VAR) of our optimal predictor, one which is multivariate linear and the other based on general radial basis functions (GRBF).

We next used conditional independence based causal discovery tools from a structural causal model (SCM) perspective, which assumes every stochastic process is generated by a system of functions with mutually independent noise terms. This entails a probability distribution for

our random variables, which we can estimate by sampling from this population. We attempt to infer the structure of the SCM by testing for independence relations in our distribution and orienting the causal relationships based on temporal ordering.

Ultimately, we found that the Granger-causality based methods suffered from low detective power in general, though this improved dramatically when testing on true causal models of a similar functional form (i.e. linear models for the linear VAR and those with significant non-linearity for the GRBF VAR). The conditional independence based methods depended significantly on the conditional independence tests used on the data, though robust, non-linear detection methods such as nearest neighbors and distance correlation seemed to work well for a wide variety of data. Therefore, we conclude that conditional independence based models are quite versatile in their applicability, though causal discovery should be accompanied by domain knowledge as to the approximate form of the data generating process.

2 THEORETICAL BACKGROUND

Before we introduce our definitions of causality, we will first define the theoretical underpinnings of the time series data we analyze. It is assumed that the reader is familiar with the terms random variable and probability space from probability theory. The following definitions come from [1], [2].

Definition 2.1 (Stochastic Process, Time Series) *A stochastic process $(\mathbf{X}_t)_{t\in T}$ is a sequence of random vectors on a probability space $(\Omega, \mathcal{F}, P)$, indexed by a strict total order on a set T whose elements are meant to represent time indices. A time series $(\mathbf{x}_t)_{t\in T}$ is defined as a realization of a stochastic process.*

Thus, for all time series data we analyze, we assume that there is some (unknown) stochastic process which generated it. We must also restrict the class of stochastic processes we consider to ones which have time invariant joint probability distributions, or at least fixed first and second moments. The reason for this is that if our probability distribution changes for each time index, then our causal relations will have changed, and thus there is no one causal structure to explain the data we have observed. A good example to consider is that of a time series with two variables, which both have increasing means over time but no causal relationship between each other. These two time series will both seem extremely correlated, leading us to assume one causes the other or there is a common cause of the two, but in reality there may be no such mechanism and we arrive at a false causal relationship from spurious correlation.

Definition 2.2 (Stationarity, Strict Stationarity) *The stochastic process $(\mathbf{X}_t)_{t\in\mathbb{Z}}$ is said to be stationary if*

1. $\mathbb{E}\left[\|\mathbf{X}_t\|^2\right] = m < \infty, \forall t \in \mathbb{Z}$ *and*
2. $Cov(\mathbf{X}_r, \mathbf{X}_s) = Cov(\mathbf{X}_{r+t}, \mathbf{X}_{s+t}), \forall r, s, t \in \mathbb{Z}$,

and strictly stationary if for $k \geq 0$

$$P(\mathbf{X}_{t_1}, \ldots, \mathbf{X}_{t_k}) = P(\mathbf{X}_{t_1+h}, \ldots, \mathbf{X}_{t_k+h}), \forall k, t_1, \ldots, t_k, h \in \mathbb{Z}.$$

Unless explicitly stated, we will assume that all time series we analyze are generated by either a stationary or strictly stationary stochastic process

Now that we've introduced the theoretical background of the data we will study, we will move on to our discussion of causality. We will assume that every random process found in the world is generated by a SCM, which is a system of assignments for our random variables which depend on their causal parents and independent noise terms. We term these equations assignments as they cannot be algebraically manipulated. We will assume familiarity with graph theory terms such as graph, edge, and acyclicity (for a reference, see the appendix). All of the upcoming definitions and theorems are from [4], [3], [6].

Definition 2.3 (SCM) *A structural causal model (SCM)* $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ *consists of a collection* $\mathbf{S}$ *of* d *(structural) assignments for random variables* $\mathbf{X} = (X_1, \ldots, X_D)$

$$X_j := f_j(\mathbf{pa}_j, N_j), \quad 1 \leq j \leq d, \quad \mathbf{pa}_j \subseteq \mathbf{X} \setminus X_j, \quad N_j \in \mathbf{N}$$

and a set $\mathbf{N} = \{N_1, \ldots, N_d\}$ *of independent noise variables with distribution* $P_{\mathbf{N}}$*. The graph* $\mathcal{G}_{\mathcal{C}} = (\mathbf{X}_{\mathcal{C}}, \mathcal{E}_{\mathcal{C}})$ *of* $\mathcal{C}$ *is defined as the graph such that*

$$(\mathbf{pa}_j, X_j) \in \mathcal{E}_{\mathcal{C}} \qquad \forall 1 \leq j \leq d.$$

The SCM is said to be acyclic if $\mathcal{G}_{\mathcal{C}}$ *is.*

We will assume throughout this paper that all SCMs are acyclic.

Theorem 2.1 (Distribution for S) *A SCM* $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ *determines a unique distribution* $P_{\mathbf{X}}$ *for* $\mathbf{X} = (X_1, \ldots, X_d)$*, called its entailed distribution.*

We now have the proper definitions to define causality.

Definition 2.4 (Causality) *For an acyclic SCM* $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ *with DAG* $\mathcal{G}_{\mathcal{C}}$*, we define causality as a homogeneous relation* $\xrightarrow{\mathcal{C}}$ *on the set of random variables* $\mathbf{X} = (X_1, \ldots, X_d)$ *such that*

$$X_i \xrightarrow{\mathcal{C}} X_j \iff X_j \subseteq \mathbf{DE}_i^{\mathcal{G}} \qquad 1 \leq i, j \leq d$$

Theorem 2.2 (Properties of Causality) *The causality relation is irreflexive, asymmetric, and transitive.*

This definition fits with our intuitive notions that an event should not be able to cause itself, that two events cannot simultaneously cause each other, and that if one event cause another event, than it is an indirect cause of any event which that event causes.

The problem in causal inference is, given a probability distribution $P_{\mathbf{X}}$ over random variables $\mathbf{X} = \{X_1, \ldots, X_d\}$, to describe the structural causal model $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ with graph $\mathcal{G}_{\mathcal{C}}$ of these random variables. This problem is ill-posed, meaning for every DAG there exists an SCM which entails the distribution $P_{\mathbf{X}}$. Thus, we need to make additional assumptions on the distribution in order to identify the correct causal structure and graph. However, first we must introduce some additional definitions.

Definition 2.5 (d-Separation) *In a DAG* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$*, a path* $p = \langle i_1, \ldots, i_k, \ldots, i_m \rangle$ *is d-separated by a set* $\mathbf{S}$ *(where* $i_1, i_m \notin S$*) if either*

1. $i_k \in \mathbf{S}$ *and*

$$i_{k-1} \to i_k \to i_{k+1} \text{ or}$$
$$i_{k-1} \leftarrow i_k \leftarrow i_{k+1} \text{ or}$$
$$i_{k-1} \leftarrow i_k \to i_{k+1}$$

2. *neither* i_k *nor any of its descendants is in* $\mathbf{S}$ *and*

$$i_{k-1} \to i_k \leftarrow i_{k+1}.$$

We also say that two disjoint subsets $\mathbf{A}, \mathbf{B} \subseteq \mathbf{V}$ *are d-separated by another disjoint set* $\mathbf{S}$ *if every path between nodes in* $\mathbf{A}$ *and* $\mathbf{B}$ *is d-separated by* $\mathbf{S}$*, which we denote*

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{S}.$$

Definition 2.6 (Markov Property) *Given a DAG* $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ *and a joint distribution* $P_{\mathbf{X}}$ *with a density* $p_{\mathbf{x}}$*, the distribution is said to satisfy the Markov property with respect to* $\mathcal{G}$ *if*

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{S} \implies \mathbf{A} \perp\!\!\!\perp_{P_{\mathbf{X}}} \mathbf{B} \mid \mathbf{S}$$

for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$.

This relationship between d-separation and conditional independence allows us to encode the independence relation of a distribution in a DAG. Thus, if we are given a probability distribution $P_{\mathbf{X}}$, we can encode the independence information in that distribution in a DAG $\mathcal{G}_{\mathcal{C}}$ of the SCM $\mathcal{C}$ of $\mathbf{X}$. The following theorem tells us this is always possible.

Theorem 2.3 (Markov Property of SCM) *If* $P_{\mathbf{X}}$ *is the entailed distribution of the SCM* $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ *with DAG* $\mathcal{G}_{\mathcal{C}}$*, then* $P_{\mathbf{X}}$ *is Markov relative to* $\mathcal{G}_{\mathcal{C}}$.

However, the specific DAG $\mathcal{G}_{\mathcal{C}}$ of a SCM $\mathcal{C}$ may be impossible to recover from distributional data alone, and thus we may need to settle for an equivalence class of DAGs.

Definition 2.7 (Markov Equivalence) *We define the set of distributions Markovian with respect to a DAG* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ *as*

$$\mathcal{M}(\mathcal{G}) = \{P : P \text{ satisfies the Markov property with respect to } \mathcal{G}$$

Two DAGS $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *are Markov equivalent if*

$$\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2).$$

The set of all DAGs Markov equivalent to $\mathcal{G}$ *is called the Markov equivalence class of* $\mathcal{G}$*, which can be represented by a PDAG*

$$CPDAG(\mathcal{G}) = (\mathbf{V}, \mathcal{E}_C),$$

where $(i, j) \in \mathcal{E}_C$ *if and only if one of the DAGs equivalent to* $\mathcal{G}$ *has this directed edge.*

Theorem 2.4 (Graphical Markov Equivalence) *Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and set of v-structures.*

Finally, we arrive at the conditions needed to identify a DAG from a distribution $P_{\mathbf{X}}$.

Definition 2.8 (Causal Faithfulness and Minimality) *For a distribution $P_{\mathbf{X}}$ and a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$,*

1. *$P_{\mathbf{X}}$ is faithful to $\mathcal{G}$ if*
$$\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B}|\mathbf{C}$$
for all disjoint sets $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbf{X}$
2. *A distribution satisfies causal minimality with respect to $\mathcal{G}$ if it is Markovian with respect to $\mathcal{G}$, but not to any proper subset of $\mathcal{G}$*

Theorem 2.5 (Identifiability of Markov Equivalence Class) *If we have a joint distribution $P_{\mathbf{X}}$ which is Markovian and faithful to a dag $\mathcal{G}^0$, then for each $\mathcal{G} \in CPDAG(\mathcal{G}^0)$ we can find a SCM $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ that entails $P_{\mathbf{X}}$. Furthermore, there is no graph $\mathcal{G} \notin CPDAG(\mathcal{G}^0)$ such that $P_{\mathbf{X}}$ is Markov and faithful relative to $\mathcal{G}^0$*

Thus, the goal of causal inference is to use distributional data to determine the equivalence class of DAGs which the distribution is Markovian and faithful to.

Independence based methods assume that the distribution $P_{\mathbf{X}}$ is faithful to the DAG $\mathcal{G}_{\mathcal{C}}$ of its generating SCM $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$, which means there is a one-to-one correspondence between d-separation and conditional independencies. These methods first estimate the skeleton, which relies on the following theorem.

Theorem 2.6 (Adjacency and D-Separation) *For a DAG $G = (\mathbf{X}, \mathcal{E})$,*

1. *Two vertices $X, Y \in \mathbf{X}$ are adjacent if and only if they cannot be d-separated by any subset $S \subseteq \mathbf{X} \setminus \{X, Y\}$*
2. *If two vertices $X, Y \in \mathbf{X}$ are not adjacent, then they are d-separated by either $\mathbf{PA}_X^{\mathcal{G}}$ or $\mathbf{PA}_Y^{\mathcal{G}}$.*

These methods then orient all undirected edges based on the fact that all v-structures entail the structure
$$X - Y - Z \implies X \to Y \leftarrow Z,$$
and other consistency rules.

Since we only have finite data, we assume that for a stochastic process $(\mathbf{X}_t)_{t \in T}$ the random variables at a time t, $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,d})$, only depend on the preceding $q \in \mathbb{N}$ lags of the time series. In other words, if the SCM of this stochastic process is $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$, then
$$\mathbf{pa}_{t,j} \subseteq \bigcup_{i=0}^{q} \mathbf{X}_{t-i} \setminus X_{t,j} \qquad \forall 1 \leq j \leq d.$$

We say that the SCM of a stochastic process has no instantaneous effects if

$$\mathbf{pa}_{t,j} \subseteq \bigcup_{i=1}^{q} \mathbf{X}_{t-i} \setminus X_{t,j} \qquad \forall 1 \leq j \leq d,$$

or in other words each variable does not depend on the values of other variables at the same time.

The conditional independence based methods analyze the distribution of $P_{\mathbf{X}_t,\ldots,\mathbf{X}_{t-q}}$ to generate a corresponding CPDAG, using the results presented here. Granger-causality, on the other hand, attempts to detect conditional independence through predictive power of causal variables, and orients causation based on time ordering. Granger-causality fits models of the form

$$X_{t,j} := f(\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-q}) + \epsilon_{t,j}$$

and

$$X_{t,j} := f(\mathbf{X}_{t-1} \setminus X_{t-1,i}, \ldots, \mathbf{X}_{t-q} \setminus X_{t-q,i}) + \tilde{\epsilon}_{t,j}$$

using ordinary least squares for each $1 \leq j \leq d$, where $(\epsilon_{t,j})_{t\in T}$ and $(\tilde{\epsilon}_{t,j})_{t\in T}$ are assumed to be independent identically distributed time series. The variable $X_{t,i}$ is said to Granger-cause $X_{t,j}$ if $\mathrm{Var}\left[\epsilon_{t,j}\right] < \mathrm{Var}\left[\tilde{\epsilon}_{t,j}\right]$, which in the case of normally distributed error terms corresponds to conditional independence.

3 METHODS

To demonstrate the workings of each method, we will suppose that we are analyzing a time series $(\mathbf{x}_t)_{t\in T}$ where $T = \{1, \ldots, \tau\}$ generated by stochastic process $(\mathbf{X}_t)_{t\in T}$ with SCM $\mathcal{C} = (\mathbf{S}, P_{\mathbf{N}})$ and graph $\mathcal{G}_{\mathcal{C}}$, which has a maximum time lag q.

The first causal discovery method developed was linear VAR Granger-causality. This consisted of fitting a linear VAR model to the data, such that we assume our SCM takes the form

$$\mathbf{X}_t := \sum_{k=1}^{q} A_k \mathbf{X}_{t-k} + \epsilon_t.$$

Given our definition of causality, we would infer a causal relationship between a variable X_i and X_t if for our coefficient matrices $A_k = [a_{k,ij}]$, $[a_{k,ij}] \neq 0$ for some $1 \leq k \leq q$ using a Wald test. This is equivalent to our previous definition, as this would imply a lesser error variance by minimization of least squares.

The second causal discovery method developed was GRBF VAR Granger-causality, which began by fitting a model to the time series data of the form

$$X_{t,j} := \sum_{k=q}^{\tau} a_k \phi_{\mathbf{B}_k}(\mathbf{A_t}),$$

where

$$\mathbf{A}_t = (\mathbf{X}_{t-1}, \ldots, \mathbf{X}_{t-q}) \text{ and } \mathbf{B}_k = (\mathbf{x}_k, \ldots, \mathbf{x}_{k-q}),$$

where $\phi_{\mathbf{B}_k} : \mathbb{R}^{dq} \to \mathbb{R}$ is a GRBF, meaning $\phi_{\mathbf{B}_k}(\mathbf{A_t}) = f(\|\mathbf{A}_t - \mathbf{B}_k\|)$ for some $f : \mathbb{R} \to \mathbb{R}$. The implementation of this method was based off of [7].

For the conditional independence based causal discovery, we used the PCMCI algorithm with partial correlation and distance correaltion tests for conditional independence. For descriptions of these algorithms, see the works of Jakob Runge in the references [5].

4 RESULTS

The results of the difference causal discovery methods are detailed below, tested using data provided by the CauseMe causal discovery benchmark. The relevant score for each method is the reciever operating curve (ROC) score which is a measure of the power of a discovery method as a function of its type 1 error rate.

Table 1: ROC Scores of Causal Discovery Methods

Data	**VAR**	**GRBF**	**PCMCI (PC)**	**DC**
LVAR (10, 300)	0.97	0.88	0.99	0.82
LVAR (20, 300)	0.97	0.83	0.98	0.82
LLNS (10, 150)	0.65	0.77	0.64	0.92
LLNS (10, 300)	0.72	0.95	0.72	0.96
NVAR (10, 300)	0.84	0.83	0.82	0.83
NVAR (20, 600)	0.88	0.88	0.87	0.92

5 CONCLUSION

The results show that the GC based on the VAR model performed best when used on data generating processes with a linear functional form, while still being able to detect some causal relations in the nonlinear case. The GRBF model performed worse on linear data, though outperformed the VAR model on non-linear data, as to be expected. The PCMCI based methods performed consistently well on both linear and non-linear stochastic processes, though it depended heavily on the conditional independence test used. Thus, we can conclude that the method of causality inference used should heavily depend on background knowledge of the underlying process. In future work, we plan to test the performance of these methods with our own climate datasets.

6 ACKNOWLEDGEMENTS

This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships Program (SULI), as well as the mentorship and direction of Dr. Wei Xu.

7 REFERENCES

[1] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, 1991.

[2] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2007.

[3] Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2013.

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

[5] Jakob Runge et al. "Detecting and quantifying causal associations in large nonlinear time series datasets". In: *Science Advances* 5.11 (2019), eaau4996. DOI: 10.1126/sciadv.aau4996. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aau4996. URL: https://www.science.org/doi/abs/10.1126/sciadv.aau4996.

[6] Peter Spirtes. *Causation, prediction, and Search*. Springer, 2012.

[7] Chad Vernon. *Regularized linear regression with radial basis functions*. URL: https://www.chadvernon.com/blog/rbf/.

7.1 CCI

The following people participated in the research presented in this paper:

Table 2: Participants

Name	Institution	Role
Evan Vera	Department of Computer Science, Cornell University	Student researcher
Wei Xu	Computational Science Iniative, Brookhaven National Lab	Mentor

All research anctivities were completed within the Computational Science Initiative facility at Brookhaven National Lab, and as so far there have been no major outcomes of this research.

7.2 GRAPH THEORY

Definition 7.1 (Graph) *A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of a finite set of vertices $\mathbf{V}$ and a set of edges $\mathcal{E} \subseteq \mathbf{V} \times \mathbf{V}$ such that $\forall v \in \mathbf{V}.(v, v) \notin \mathcal{E}$. A graph $\mathcal{G}_S = (\mathbf{V}_S, \mathcal{E}_S)$ is defined as a subgraph of a graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ if $\mathbf{V}_S = \mathbf{V}$ and $\mathcal{E}_S \subseteq \mathcal{E}$, and a proper subgraph if $\mathcal{E}_S \subset \mathcal{E}$.*

Definition 7.2 (Parent, Child, Adjacent) *A vertex $i \in \mathbf{V}$ of a graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is called a parent of vertex $j \in \mathbf{V}$ if $(i, j) \in \mathcal{E}$*

$$i \to j$$

and $(j, i) \notin \mathcal{E}$

$$j \nrightarrow i,$$

and a child of j if $(j, i) \in \mathcal{E}$

$$j \to i$$

and $(i, j) \notin \mathcal{E}$

$$i \nrightarrow j.$$

The set of parents of vertex i *is denoted* $\mathbf{PA}_i^{\mathcal{G}}$*, and its set of children is denoted* $\mathbf{CH}_i^{\mathcal{G}}$*. Vertices* i *and* j *are defined as adjacent if either* $(i, j) \in \mathcal{E}$

$$i \rightarrow j$$

or $(j, i) \in \mathcal{E}$

$$j \rightarrow i,$$

and $\mathcal{G}$ *is fully-connected if* $\forall v, w \in \mathbf{V}$*.* v *and* w *are adjacent.*

Definition 7.3 (Directed, Undirected) *There is an undirected edge between two adjacent vertices* $i, j \in \mathbf{V}$ *of graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ *if* $(i, j) \in \mathcal{E}$

$$i \rightarrow j$$

and $(j, i) \in \mathcal{E}$

$$j \rightarrow i.$$

An edge between two adjacent nodes is directed if it is not undirected. We define $\mathcal{G}$ *as directed if all its edges are directed.*

Definition 7.4 (V-Structure, Skeleton) *Three nodes are called a v-structure if one node is a child of the other two, and the other two nodes are not adjacent. The skeleton* $\tilde{\mathcal{G}} = (\mathbf{V}, \tilde{\mathcal{E}})$ *of a graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ *is the graph such that* $(i, j) \in \tilde{\mathcal{E}}$ *if* $(i, j) \in \mathcal{E}$

$$i \rightarrow j,$$

or $(j, i) \in \mathcal{E}$

$$j \rightarrow i.$$

Definition 7.5 (Path, Descendant) *A path* $p = \langle i_1, \ldots, i_m \rangle$ $(m \geq 2)$ *of a graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ *is a sequence of distinct vertices such that* $(i_k, i_{k+1}) \in \mathcal{E}. \forall 1 \leq k < m$*, or visually*

$$i_k \rightarrow i_{k+1}.$$

A subsequence $\langle i_{k-1}, i_k, i_{k+1} \rangle$ *where* $(i_{k-1}, i_k) \in \mathcal{E}$*, meaning*

$$i_{k-1} \rightarrow i_k$$

and $(i_{k+1}, i_k) \in \mathcal{E}$

$$i_{k+1} \rightarrow i_k$$

is called a collider relative to p*. If* $(i_k, i_{k+1}) \in \mathcal{E}$

$$i_k \rightarrow i_{k+1}$$

for all $1 \leq k < m$*, then we say that* p *is directed, and that* i_1 *is an ancestor of* i_m *and* i_m *is a descendant of* i_1*. We denote the set of all descendants of* i *as* $\mathbf{DE}_i^{\mathcal{G}}$ *and the set of all its non-descendants as* $\mathbf{ND}_i^{\mathcal{G}}$*.*

Definition 7.6 (PDAG, DAG) *A graph* $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ *is called a partially directed acyclic graph (PDAG) if there are no directed cycles, that is there are no pairs* $(j, k) \in \mathbf{V} \times \mathbf{V}$ *such that there is a directed path from* j *to* k *and a directed path from* k *to* j*. The graph* $\mathcal{G}$ *is called a directed acyclic graph (DAG) if it is a PDAG and all its edges are directed.*