COMPUTATIONAL METHODS FOR DISCOVERING AND ANALYZING
CAUSAL RELATIONSHIPS IN HEALTH DATA

Yiheng Liang, M.S.

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS
August 2015

APPROVED:

Armin R. Mikler, Major Professor
Joseph R. Oppong, Committee member
Bill P. Buckles, Committee member
JungHwan Oh, Committee member
Suhasini Ramisetty-Mikler, Committee member
Barrett R. Bryant, Chair of the Department of
Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of
Engineering and Interim Dean of
Toulouse Graduate School

Publicly available datasets in health science are often large and observational, in contrast to experimental datasets where a small number of data are collected in controlled experiments. Variables' causal relationships in the observational dataset are yet to be determined. However, there is a significant interest in health science to discover and analyze causal relationships from health data since identified causal relationships will greatly facilitate medical professionals to prevent diseases or to mitigate the negative effects of the disease.

Recent advances in Computer Science, particularly in Bayesian networks, has initiated a renewed interest for causality research. Causal relationships can be possibly discovered through learning the network structures from data. However, the number of candidate graphs grows in a more than exponential rate with the increase of variables. Exact learning for obtaining the optimal structure is thus computationally infeasible in practice. As a result, heuristic approaches are imperative to alleviate the difficulty of computations.

This research provides effective and efficient learning tools for local causal discoveries and novel methods of learning causal structures with a combination of background knowledge. Specifically in the direction of constraint based structural learning, polynomial-time algorithms for constructing causal structures are designed with first-order conditional independence. Algorithms of efficiently discovering non-causal factors are developed and proved. In addition, when the background knowledge is partially known, methods of graph decomposition are provided so as to reduce the number of conditioned variables.

Experiments on both synthetic data and real epidemiological data indicate the provided methods are applicable to large-scale datasets and scalable for causal analysis in health data. Followed by the research methods and experiments, this dissertation gives thoughtful discussions on the reliability of causal discoveries computational health science research, complexity, and implications in health science research.
ACKNOWLEDGMENTS

I sincerely express the first thank to my major professor Dr. Armin R. Mikler for his endless support and his professional guide to my Ph.D research. I also thank my parents because they have given me their persistent encouragement during the years of my Ph.D study and they have nurtured me with valuable experience in life. In addition, I thank my committee members for their precious time on my academically related activities, my coursework instructors who taught me knowledge in various fields, and all my workplace supervisors and colleagues who have provided opportunities to communicate and to gain my career experience. Last, I thank all my friends who positively influenced me as well as my enemies who once discouraged me, as they have shaped my character to be unique and make me endeavor to succeed.
# Table of Contents

## Acknowledgments

LIST OF TABLES

## Chapter 1 Introduction

1.1. Research Background in Epidemiology and Computer Science

1.2. Challenges in Health Science Research

1.3. Challenges with Data and Study Designs in Health Science

1.4. Computational Approaches in the Problem Solving

1.5. Research Contributions

## Chapter 2 Background and Literature Review

2.1. Discovering Associations

2.2. From Association to Causation

2.2.1. Graphical Models and Causal Models

2.2.2. Probabilistic Models in Epidemiological Research

2.2.3. Examples of Disease Related Issues

2.3. Bayesian Networks

2.3.1. Definition

2.3.2. Conditional Independence and Conditional Dependence

2.3.3. Graphical Representations of Dependency

2.3.4. Inference and Learning

2.4. Heuristics

2.5. Causation and Its Definitions

2.6. From Bayesian Networks to the Discovery of Causality

2.6.1. Learning Structures of Bayesian Networks

2.6.2. Inferring Causal Structures in Bayesian Networks
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparisons of Causal Modeling Methods</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Learning Methods in Bayesian Networks</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td>Number of Labeled DAGs</td>
<td>45</td>
</tr>
<tr>
<td>5.1</td>
<td>Two-by-Two Contingency Table</td>
<td>62</td>
</tr>
<tr>
<td>5.2</td>
<td>Measurement of Accuracy</td>
<td>65</td>
</tr>
<tr>
<td>5.3</td>
<td>Performance of Learning Asia Network</td>
<td>77</td>
</tr>
<tr>
<td>5.4</td>
<td>Performance of Learning ALARM Network</td>
<td>79</td>
</tr>
<tr>
<td>6.1</td>
<td>An Example of Row-Oriented Table of Data</td>
<td>88</td>
</tr>
<tr>
<td>6.2</td>
<td>An Example of Column-Oriented Table of Data</td>
<td>89</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1. Research Background in Epidemiology and Computer Science

In epidemiology, investigators are dealing with a large volume of data and are trying to obtain useful information from them. Most publicly available epidemiological datasets are large in terms of the number of individuals and the number of surveyed variables. Epidemiological datasets, such as disease surveillance data, are often observational, thereby no causal relationships are explicitly described or studied. Hence, the discovery and analysis of causal relationships in epidemiological datasets are of significant interest to health professionals [65].

In computer science, researchers are developing methods to analyze data with graphical models. Particularly, Bayesian networks are used for probabilistic reasoning and causal analysis. In order to identify possible causal relationships among variables, learning the structures of a Bayesian network is needed. However, structural learning process without background knowledge, with partially observed data, or with large samples is hard. Hence, when the dataset is large, existing methods of learning a complete Bayesian network are computationally infeasible. Such learning methods are restricted by implausible assumptions or are subject to specific data structures.

Combined with the background and research need in both epidemiology and computer science, this research provides novel computational approaches to discover and analyze causal relationships in epidemiology. In this dissertation, we discover causal structures with graphical models, analyze causal relationships from large-scale datasets, and construct causal networks with the improved efficiency.

1.2. Challenges in Health Science Research

In public health research in general and in chronic disease study in particular, investigators identify risk factors for a certain disease, assess the impact of the disease over time, and provide suggestions and methods to prevent the disease progressions. Chronic diseases
such as diabetes or breast cancer are under scrutiny in that identifying all causes or
risk factors for a certain disease is intractable and not all diseases and health problems can
be inferred from an individual’s health conditions.

One obstacle faced by medical professionals is that a large number of disease records
as well as a relatively long term of observations is required for them to evaluate all the in-
fluencing factors and impacts of the disease properly and precisely. For prospective cohort
studies and randomized controlled designs, the cost of collecting sufficient samples to con-
duct clinical research over time is not negligible. In addition, causal analysis is essential in
epidemiological studies [114]. Finding associations among factors rarely provides informa-
tion about their causal relationships. Only through controlling causal factors of a disease
can medical professionals prevent the disease and mitigate the negative effects of the disease.
A human’s judgment on causality is sometimes unfounded if no experiments are performed
to verify the assertion, but conducting each controlled study is time consuming and not cost
efficient.

1.3. Challenges with Data and Study Designs in Health Science

Due to the quantity and complexity of various medical data, researchers are striving
to develop methods to efficiently analyze those data in health science. Data mining initially
plays an important role. Especially, medical data differ from other data sources in that they
are unique in many aspects [20]. General data mining methods may not be applicable in
the specific fields of epidemiology. In order to better interpret the implications of discovered
results and understand the mechanisms of the diseases, some modifications, adjustments,
and special attentions of mining data in health science may be required.

Epidemiological studies are either descriptive such as ecological and cross-sectional
studies or analytical such as case-control, cohort, or follow-up studies. Descriptive studies
provide information of disease occurrence in populations and analytic studies have specific
focus to test causal hypotheses generated from clinical observations or laboratory studies
[67]. Data collected in descriptive studies are in large quantity but are not primarily used
for causal inference.
Despite the analysis on association of those surveyed factors, it may be sometimes possible to infer causation from statistical relationships on observational data [25] [100] [110] [129], though additional experiments are required to ascertain the causal relationships with temporality. Advantages of using publicly available observational data are manyfold. For example, surveillance data are relatively easy to obtain, as opposed to clinical data. The size of the dataset is large and the survey covers many participants and collects data on several variables, as this can be an advantage to prevent bias in epidemiological study.

Besides the difficulty of inferring causation solely from observational data, there are several drawbacks:

1. There may be issues with data quality since data are surveyed from a large sample but not precisely measured and verified [117] or due to a low participation rate [36] [43],

2. There may be several irrelevant variables and ambiguous expressions in the survey that make it hard for researchers to make any inference from the analysis of data, and

3. Observational datasets may include redundant surveyed data.

1.4. Computational Approaches in the Problem Solving

Traditional statistical methods that embody many parametric assumptions such as correlated-outcome logistic regression analysis [52] have disadvantages in analyzing epidemiological data due to several facts.

First, the dataset may be large. In order to find interesting information such as previously unnoticed associations, an efficient algorithm should be utilized because time and cost of discovering such information from large-scale data should be taken into account [79]. For example, the Behavioral Risk Factor Surveillance System (BRFSS) dataset each year has over 400,000 surveyed individuals and each with almost one thousand variables.

Second, the structure of data may be skewed. For example, when we study a certain type of a rare disease, the disease cases are limited, only being a small percentage of the population. Since most observational data such as surveillance data are not used to detect rare diseases, it is nearly impossible to get much valuable information about their prevalence or potentially related cause and effect of these diseases from the data.
Third, some traditional statistical methods have disadvantages in dealing with parametric inclusions by overestimating the importance of one parameter or underestimating it [52]. Such a situation can be that if one parameter, probably a confounding factor, is included in a regression model, or if one less noticeable parameter, but a real independent risk factor, is not considered in such kind of model.

Due to the cost of time and expense, it is impossible to perform all medical experiments. Computational methods are used to facilitate the process of discovering possible causal structures of variables in the epidemiological dataset. A computational method has an advantage of efficiently processing large amounts of data. In model scientific research, reasoning about causality is not just empirical from discovered correlations, but should be analytical by providing experimental evidence. Moreover, developed tools in computer science such as Bayesian networks are effective at representing inter-relationships among variables and superior in analyzing probabilities. All these computational approaches should eventually facilitate health professionals for their further research and provide convincing guidelines to policy makers for them to make effective and justifiable decisions.

We use the following figure to explain the problems of this research in a data perspective (Figure 1.1). In general, epidemiologists model chronic diseases by reasoning with cause and effect from experiments. Computer scientists research in the fields of machine learning and artificial intelligence, and have interests in graphical models such as Bayesian networks to infer causal structures from data. An epidemiologist is unaware of the total number of variables if she investigates only one exposure and one outcome of the disease. Existing methods in computer science can also discover the optimal causal structures from a Bayesian network by obtaining all probability distributions from small samples. However, the increase of data in size creates new problems in epidemiological research and computer science. It is nearly impossible to conduct all experiments by analyzing each pair of variables. Even if experiments are made possible, experimental data may not exist. Observational data may be available or helpful but they are relatively large in size and non-temporal. In computer science, there are many problems solvable only when the input size of data is small.
but intractable in general (i.e. not solvable in polynomial time) since they are NP-hard [26]. In a word, this research provides solutions to alleviate the difficulty of learning causal relationships from large data to facilitate epidemiological research.

![Diagram](image)

**Figure 1.1. Explanation of Research in a Data Perspective**

1.5. Research Contributions

This research contributes to both computer science and epidemiology. In the computational part, we provide the methods of locally discovering causal and non-causal factors and the learning methods with partially known background knowledge. In the epidemiological part, we show how our computational applications facilitate causal discoveries and analytics in the multivariate dataset. We then show how our models are specifically suitable for health science research.

An epidemiological dataset may contain many causal relationships among variables but discovering all of them is computationally hard. To obtain a portion of useful information where variables may be causally related, we focus on the development of local causal discoveries. The first contribution in the dissertation shows the following:

1. the design of a framework of solving the computational problem,
(2) the extended algorithms from currently existing ones for local causal discoveries,

(3) the methods of learning network structures through subgraph decompositions for parallel processing, and

(4) the reductions of computations feasible of learning causal structures in a disease-related network.

In health science search, facts about a few causal relationships are known, yet many more still have to be explored. Followed by the contributions in the area of local causal discoveries, we explore computational approaches of learning causal structures combined with partially known background knowledge. The second contribution of in the dissertation shows the following:

(1) the methods of graph decompositions from given ordering of variables,

(2) the theoretical improvements to existing constraint based learning methods, and

(3) the efficiency of learning causal structures with our methods.

The structure of this dissertation is organized as follows. First we introduce the theoretical background, show the research motivation, and provide a brief literature review related to our work. Next, we explain our methods and analyze the algorithms. In addition, we argue the necessity of implementing these methods, which are potentially beneficial to the analysis of data in the public health domain. Followed by the research methods, the experimental design as well as analytical discussions of the experiments are provided. After the discussions of advantages and limitations in the research methods, we conclude and leave questions and suggestions for future work.
CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1. Discovering Associations

Association rule learning is a process that discovers previously unnoticed or underestimated relations among multiple variables in data mining. By definition [54], given a whole set of items \( I = \{ i_1, i_2, ..., i_n \} \), one task is to find such frequent patterns \( X \Rightarrow Y \) where \( X \) and \( Y \) are two sub-sets of \( I \) and \( X \cap Y = \emptyset \), with which support of \( X \Rightarrow Y \) and confidence of \( X \Rightarrow Y \) are all above the previous set standard.

Support of a rule \( A \Rightarrow B \) is the probability of the itemset \( A, B \): 
\[
support(A \Rightarrow B) = P(A,B).
\]

Confidence of a rule \( A \Rightarrow B \) is the conditional probability of \( B \) given \( A \): 
\[
confidence(A \Rightarrow B) = P(B|A) = \frac{support(A,B)}{support(A)}.
\]

Rule support and confidence are two measures of rule interestingness. When both support and confidence are above the previously set minimum threshold, such an association rule is considered strong.

To explain the above association rule, an assumptive example of mining association rules in an epidemiological dataset can be as follows:

Let the set of items \( I = \{ \text{gender, alcohol consumption, obesity, snoring, smoking, diabetes, lung cancer} \} \) in the database, find all association rules that meet the standard of support and confidence.

The results of such findings can be “obesity” \( \Rightarrow \) “diabetes” and “smoking” \( \Rightarrow \) “lung cancer”. If more variables are included, the finding may be expressed as: “obesity, snoring” \( \Rightarrow \) “diabetes”, and the association between “diabetes” and the occurrence of both “obesity” and “snoring” may have an even higher support than that of “obesity” \( \Rightarrow \) “diabetes”.

Motivated by the need to ascertain different risk factors that may contribute to a chronic disease, and how much influence those factors have for the disease [78], we initially aimed to study multiple variables with association rule learning. However, discovering asso-
ciations from an exceptionally large number of factors for a certain disease requires massive computing resources with respect to time and space. Moreover, the reality of disease dynamics is often more complex than the numerical model of associations itself.

There is ample research on data mining with the scope of medical-related data [6] [34] [69] [127]. Besides the issues with cost of time in mining associations with multiple factors, there are also several other problems with the general process of association rules discovery. These problems have been addressed or only partially solved.

One of these problems is that a general association rule learning method only considers the frequency of items [80]. However, it is not always effective to use frequency to analyze epidemiological data because the number of cases for a certain disease may only account for a small portion of the whole dataset. For example, in the Behavioral Risk Factor Surveillance System (BRFSS) data, the reported percentage of diabetes cases is just around ten percent of the total population. Unless the research scientists set the support and confidence low enough, they will not be able to retrieve all cases of diabetes or associated risk factors of diabetes effectively. This problem is due to the low frequency of the number of disease cases in the dataset. However, if they set the support and confidence with low frequency, several noises will come up in the association rules discovery. Discoveries based on frequency may result in finding many “good conditions”. For example, researchers may find \{ “No smoking”, “No sleeping disorder”, “No eye problems”\} ⇒ “No cancer”; however, they are more interested in finding conditions that result in the disease. Solutions to such kinds of problems and analysis of data in epidemiology are different from most conventional association rule discovery based on the frequency count, though there is a possible solution of developing a method to mine the infrequent dataset [63].

Another problem with the association rules mining is that causal relationships among variables in a given set are not taken into account but should be identified. Assume that we have discovered associations such as \{ “Obesity”, “No alcohol consumption”\} ⇒ “Diabetes”, or “Alcohol consumption” ⇒ “No diabetes” [62]. These findings will not make sense unless both “Obesity” and “No alcohol consumption” are two distinct causes of diabetes. It does
not even make sense biologically to conclude that people who drink often are less likely to have diabetes. In this case, “Obesity” is supposedly a cause of “Diabetes”, while “No alcohol consumption” could possibly be a result of “Diabetes”: People who have been diagnosed with diabetes may be refrained from drinking, as advised by doctors.

Given the nature and limitations on the association rules learning in epidemiological dataset, our design and the development of new methods and models are propelled by all these discussed problems.

2.2. From Association to Causation

2.2.1. Graphical Models and Causal Models

Graphical models [74] are used to study relationships of multiple factors and disease [32] [87] [131], in etiologic inference [109], or to reduce bias in epidemiological research [119]. These models include Bayesian networks, Markov networks [99], chain graphs [75] [77], and influence diagrams [92]. A web-of-causation is traditionally used to visually representing causation of diseases [71].

Models for causal discoveries in epidemiology include potential-outcome models, structural-equations models (SEM) [13], and sufficient-component cause models. Causal modeling methods in epidemiology vary [91] and have their own advantages and disadvantages respectively. Table 2.1 is a brief summary of major four models [51]:

In this research, we focus on graphical models such as Bayesian networks for causality research. A Bayesian network can be used to infer causal structures among variables. Note the difference between a Bayesian network and a causal network. A causal network (i.e. a causal diagram shown in Table 2.1) is considered to be a graphical representation based on the fact of causality, which is qualitative and factual. A Bayesian network that represents probabilistic independence, only defines a directional link between the conditions of variables and such relationships are quantitative and not always factually causal. For instance, if there is a directed edge from node A to node B in the Bayesian network, A is the precursor of B (A should or is likely to happen before B), but A does not have to be a real cause of B. A causal network, or a causal diagram [52] however, is theoretically relevant to Bayesian
### Table 2.1. Comparisons of Causal Modeling Methods

<table>
<thead>
<tr>
<th>Models</th>
<th>Features and limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical Model (Causal</td>
<td>Representation of qualitative assumptions, easy to understand; Potential problems for quantitative analysis.</td>
</tr>
<tr>
<td>Diagrams)</td>
<td></td>
</tr>
<tr>
<td>Potential-outcome (Counterfactual) Models</td>
<td>Quantitative assumptions about responses of units in study population; Causes may be mistakenly claimed.</td>
</tr>
<tr>
<td>Structural-equations Model</td>
<td>Widely used in social science, providing analysis of effects; Not accurate quantified representation of causations.</td>
</tr>
<tr>
<td>(SEM)</td>
<td></td>
</tr>
<tr>
<td>Sufficient Component Cause Model (SCC)</td>
<td>Originated in epidemiology, requiring specification of mechanisms within the individual units. Mostly in theory.</td>
</tr>
</tbody>
</table>

network but edges do represent causation. Our research on Bayesian networks is acting as a vehicle to discover potential causal structures. We will formally describe the details of Bayesian Networks in the rest of this chapter.

The reason of making such a distinction in our research is that researchers are not able to completely exclude non-causal factors such as confounders. Moreover, the constructed Bayesian network is based on parameters and inferred structures. Information provided by this Bayesian network may be falsely represented due to the lack of background knowledge, the missing of latent variables or the misrepresentations of reality.

#### 2.2.2. Probabilistic Models in Epidemiological Research

We adopt a probabilistic model in this research to evaluate associations, causations, or the risks of a disease. Researchers are not always one hundred percent certain about their knowledge or absolutely confident with observational results. The inclusion of probability in the disease models is necessary because it is advantageous for the health science researchers to reason under uncertainty.

Probabilistic reasoning is widely used in expert systems[88] and multiagent systems[130]. Methods using probabilistic reasoning has played a central role in artificial intelligence with a long history and has recently been adopted into particular research fields such as psychology,
Several types of logic including propositional logic, first-order logic, and temporal logic have only three statuses: on, off, and unknown [124]. Since people are not certain about the fact(verity), a reckless classification to the status either “True” or “False” will often lead to mistakes. However, “Unknown” status provides no more information than telling us that such a subject should be further scrutinized. In reality, people’s judgment on a matter is usually somewhere between completely uncertain and completely affirmative.

Fuzzy logic is sometimes used though [48], which has a degree of truth from 0 (completely false) to 1 (completely true). Fuzzy logic may not be suitable in this research since we are not measuring the degree of truth about a certain fact such as causality. In other words, the fact of a causal relationship itself is yes or no.

To avoid using these terms interchangeably, we make a distinction between probability theory and fuzzy logic. Fuzzy logic in nature focuses on the ambiguities in describing events, rather than the uncertainty about the occurrence or non-occurrence of events [99]. In this research, we are evaluating the degree of belief in our reasoning (more subjectively) such as how much confidence we think if there is a causal relationship.

Most traditional methods and general regression models for epidemiological research [83] [85] do not properly incorporate probability theory for causal inference. Though there are some models such as decision tree models in which probabilistic measurements can be added, they do not serve as universal or convenient as the graphical models, in which probabilistic measurements can be conveniently expressed and explicitly represented.

2.2.3. Examples of Disease Related Issues

A given disease could be often associated with multiple risk factors and disease symptoms are affected by several environmental conditions [106]. In order to control the disease progression and to provide effective treatments to the disease, medical professionals and policy makers need to evaluate all associated factors and weigh in with their decisions. However, the problem is how they will balance their judgments about uncertainties with all possible outcomes in decision making [66].
One issue for analyzing multiple risk factors is whether we treat those factors as independent or dependent. For example, if there are two potential risk factors “snoring” and “obesity”, and both are associated with “diabetes”, we may find that both the association between “snoring” and “diabetes” and the association between “obesity” and “diabetes” are relatively strong. For instance, according to the Behavioral Risk Factor Surveillance System (BRFSS) data in 2010, the odds ratio for snoring and diabetes is 1.5, and the odds ratio for obesity (BMI \( \geq 25 \)) and diabetes is 3.3. There is also a finding that the association between the occurrence of “snoring” and “obesity” both and the presence of “diabetes” is even stronger and has an odds ratio of 4.4. These values of odds ratios shown are raw results, which are only used to show that three variables are correlated with each other.

General association rules mining with a more sensible evaluation such as the odds ratio instead of the frequency measurement in general, is still problematic. Odds ratio is the odds of disease among the exposed number divided by the odds of disease among the unexposed number. We rewrite the associations as:

\[ A \geq C, \text{ when odds ratio for } A \text{ and } C \text{ is above our pre-set support}; \]
\[ B \geq C, \text{ when odds ratio for } B \text{ and } C \text{ is above our pre-set support}; \]

The research goal is to find such \((A+B) \Rightarrow C\), when the odds ratio for “both \(A\) and \(B\)” and \(C\) is above our pre-set support and has an even higher value than that of \(A \Rightarrow C\) or \(B \Rightarrow C\). There can be different assertions: variables are either dependent from each other or independent.

Assertion One: variables are dependent.

If we assert that obesity and snoring have a synergistic effect based on the criterion of association rules mining \((A \Rightarrow C, B \Rightarrow C \text{ and } (A+B) \Rightarrow C\) for a higher value of association), we may incorrectly classify at least one risk factor as a real, independent cause of the disease, given the fact that “snoring” and “obesity” are strongly correlated (the odds ratio is 2.6, according to data). This conclusion seems to suggest that “snoring” and “obesity” are dependent on each other. When two factors are supposedly independent but are actually not, the situation could possibly be due to an interaction or effect modifier, which is because
of another effect, or *confounding*, which is associated with cause and effect but not a real cause of the disease.

Assertion Two: variables are independent.

Suppose we have two independent risk factors for a certain disease. The odds ratio of the first risk factor and the disease is $A$ and the odds ratio for the second risk factor and the same disease is $B$. Assuming there is no statistical error, no hidden variables, or any noises in data, we shall expect that the odds ratio after combining these two risk factors and the disease is $A \times B$:

$$\text{OR} \ (A+B, C) = \text{OR} \ (A, C) \times \text{OR} \ (B, C),$$

where OR stands for odds ratio.

If the equation does not hold given the actual results, i.e. $\text{OR} \ (A+B, C)$ is not equal to $\text{OR} \ (A, C) \times \text{OR} \ (B, C)$, the condition suggests that two variables are not completely independent.

Neither of these two assertions is true. An arbitrary classification of variables being independent or dependent is inadequate. Regardless of the presence of “obesity”, “snoring” may or may not be a risk factor of “diabetes” [4].

To visualize the directional relationships among snoring, obesity, and diabetes, and assuming snoring and obesity are possible causes or influencing factors of diabetes and obesity influencing snoring, we can represent them with the following graph:

![Figure 2.1](image.png)

**Figure 2.1.** Obesity, Snoring, and Diabetes

*Figure 2.1* shows a directed graph. A directed graph is a graph where edges have directions. Each node in the graph represents a variable, which could be a risk factor, a condition, a symptom, or a type of disease. The directed edge from $A$ to $B$ represents $A$ causes $B$ or $A$ influences $B$. The premise is that an association is a necessary condition of
a causation, and two variables cannot be causally related or influenced with each other if there is no directed edge between two nodes.

On the decision of how to evaluate if two factors are correlated, or if they are marginally dependent, we may use the odds ratio as the measurement. A directed graph rather than an undirected graph is used to represent causal relationships because it is more specific to depict multiple causes and effects, independent risk factors, or even confounding factors.

The above example of the problems occurred in the disease modeling and the evaluations of risk factors have been shown. For the rest of this chapter, it is organized as follows. First, we introduce the Bayesian network and causal inference from the Bayesian network through learning data and structure. Second, we show how to overcome the difficulties from existing research methods by introducing heuristics. Third, we discuss the definitions of causality and how the causality is specifically related epidemiological research. Finally, we show our solutions to the problems of causal discoveries from learning Bayesian networks.

2.3. Bayesian Networks

We introduce the Bayesian network and show how the Bayesian network is applied to epidemiological research. Then we present the general research pertaining to inference and learning from a Bayesian network.

2.3.1. Definition

A Bayesian network (or a Bayesian belief network) is a directed acyclic graph where nodes represent variables and directed edges represent conditional probabilities [98] [101]. The direction of an edge, such as from A to B, represents that A has directional influence on B. Directional influence means there is a statistical association between A and B, and A happens before B. The fact that A may directly or indirectly cause or influence B is represented by “→” (i.e. A → B).
2.3.2. Conditional Independence and Conditional Dependence

We define the terms of conditional independence and conditional dependence. The notation of upper-class letters denotes random variables and the notation of lower-class letters denotes their values.

*Conditional Independence:*

Let $V_1$, $V_2$, and $V_3$ be random variables with a joint distribution. For three variables $V_1$, $V_2$, and $V_3$, suppose that the distributions of $V_1$ and $V_2$ are statistically independent conditioned on $V_3$, we say that $V_1$ and $V_2$ are conditionally independent given $V_3$.

When $V_1$, $V_2$, and $V_3$ are discrete random variables, The following form of conditional independence can be simplified as $V_1 \perp V_2 \mid V_3$:

$$P(V_1 = v_1, V_2 = v_2 \mid V_3 = v_3) = P(V_1 = v_1 \mid V_3 = v_3) \cdot P(V_2 = v_2 \mid V_3 = v_3)$$

The above equation holds for all $z$ with the probability $P(V_3 = v_3) > 0$.

Since this research deals with discrete variables only, if not specified elsewhere, we use the simplified form of conditional independence throughout all chapters.

*Conditional Dependence:*

Suppose that the distributions of two variables $V_1$ and $V_2$ are statistically independent from each other but these two variables become statistically dependent conditioned on a third variable $V_3$, we say $V_1$ and $V_2$ are conditionally dependent given $V_3$. Such a type of structure is referred as a v-structure or a collider in graphical models and in epidemiology.

Conditional independence can explain that the concurrence of individuals’ behaviors is due to the presence of a third condition. For example, people who can legally vote can also legally drive. The fact that “Vote” and “Drive” are associated is because of the condition “Age over eighteen”.

Conditional dependence can explain that previously independent factors are associated only when a third condition is observed. For example, the alarm may be triggered by either an earthquake or a burglary. “Earthquake” and “Burglary” are associated only when the condition “Alarm” is present; however, they are independent events and share no common reason to happen together.
More often in the real world, two factors are neither completely conditionally independent nor completely conditionally dependent under observation of any third variable. We may still use the concept of conditional independence, but in addition, we consider a threshold to determine if two variables are independent or dependent given a third variable. For instance, if two variables are marginally dependent (i.e. the condition of dependence is so trivial that it can be ignored), these two variables are independent.

2.3.3. Graphical Representations of Dependency

Conditional independence and dependence are mathematical forms. Rather than a pure numerical approach, a graphical representation of variables’ relationships visualizes the conditions of dependency. Suppose we have a directed acyclic graph representing a Bayesian network shown below (Figure 2.2): 

![Figure 2.2. An Example of a Bayesian Network with Five Variables](image)

In such an example with five variables, if two nodes have no direct link that connects them, then these two nodes are considered to be independent. In this figure, $A$, $B$, $D$, and $E$ are all independent from each other, because $C$ separates them. Under the condition of the presence of variable $C$, if the statistical results suggest that $A$ and $B$ are associated, i.e. the coexistence of $A$ and $B$, $A$ and $B$ will be conditionally dependent given $C$.

Any single change in one variable should not directly influence any change on the other variable and should not be instantly affected by an independent variable. For example in this graph, any changes originated from $A$ should not have any direct impact on $D$. The reason that $D$ changes is the changes from $A$ or $B$ through $C$: As $A$ and/or $B$ changes, they
affect the variable $C$ first, and the changes of $C$ is the reason that $D$ changes. Without any changes of $C$, $D$ will remain unchanged whatever the status of $A$ or $B$ is. For example, the Centers for Disease Control and Prevention (CDC) suggests that diabetes (e.g. $A$) and influenza (e.g. $D$) may be associated and the presence of diabetes increases the risk of getting influenza complications by individuals (http://www.cdc.gov/flu/diabetes/ and http://www.flu.gov/at-risk/health-conditions/diabetes/), but the mechanism behind the association is that individuals with diabetes has a relatively poor immune system (e.g. $C$), which will have less ability to fight with infected viruses. Consequently, these people have at high risk of getting influenza and serious complications.

The example shows how to reason cause and effect with the Bayesian network model. The bottom-up approach, which we infer the mostly likely cause from the evidence of the effect, is referred as diagnostic reasoning. The top-down process, which we compute the probability of the effect given evidence, is referred as causal reasoning.

2.3.4. Inference and Learning

There are different types of inference problems [47] in Bayesian networks from prior knowledge, data, or a combination [57]. One problem is learning the probability from a Bayesian network. Values of the probability may be unknown. Such a learning process is categorized as “learning the data”. Another problem is learning the relationships among variables. The existence of variables’ relationships is unknown. Such learning process is categorized as “learning the structure”. In real situations, data or structure may be completely known, partially known, or completely unknown. In summary, there will be four categories of learning from Bayesian networks:

(1) Complete data, known structure,

(2) Incomplete data, known structure,

(3) Complete data, unknown or partially known structure, and

(4) Incomplete data, unknown or partially known structure.

Table 2.2 shows the possible learning methods based on the observability of Data and Structure in each category.
Table 2.2. Learning Methods in Bayesian Networks

<table>
<thead>
<tr>
<th>Structure</th>
<th>Data</th>
<th>Learning Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>Complete</td>
<td>Maximum Likelihood Estimation</td>
</tr>
<tr>
<td>Known</td>
<td>Incomplete</td>
<td>Expectation–Maximization (EM)</td>
</tr>
<tr>
<td>Unknown</td>
<td>Complete</td>
<td>Search over Space</td>
</tr>
<tr>
<td>Unknown</td>
<td>Incomplete</td>
<td>EM and Search over Space</td>
</tr>
</tbody>
</table>

The most difficult category is learning from unknown structure where data are partially known. However, even if we consider the so-called “easiest” category, in which data and structure are all known to us, it is still computationally difficult and time-consuming for a complete inference of the Bayesian network.

We use a Bayesian network with three variables only to show the complexity of learning data. Three variables are named as A, B, and C. We calculate all possibly different probabilities from this three-variable structure. These probabilities include (a) marginal, (b) joint, and (c) conditional.

(a) A marginal probability considers a single variable with the state of either presence or non-presence. There are $3 \times 2 = 6$ marginal probabilities:

$P(A), P(\neg A), P(B), P(\neg B), P(C), P(\neg C)$.

(b) A joint probability considers more than one variables with the states of either presence or non-presence for each variable.

(b.1) For 2 variables, there are a total of $3 \times 2^2 = 12$ joint probabilities:

$P(A, B), P(A, C), P(B, C), P(\neg A, B), P(\neg A, C), P(\neg B, C), P(A, \neg B), P(A, \neg C), P(B, \neg C), P(\neg A, \neg B), P(\neg A, \neg C), P(\neg B, \neg C)$

(b.2) For 3 variables, there are $2^3 = 8$ joint probabilities:

$P(A,B,C), P(\neg A,B,C), P(A,\neg B,C), P(A,B,\neg C), P(\neg A,\neg B,C), P(\neg A,B,\neg C), P(A,\neg B,\neg C), P(\neg A,\neg B,\neg C)$
(c) A conditional probability considers the one or more variables under the specific state of other variables. For example, $P(B|A)$ is the conditional probability of $B$ given $A$.

(c.1) For a single variable conditioned on one variable only, there are $2^2 \times 2 \times 3 = 24$:

Conditioned on $A$:
$P(B|A), P(\neg B|A), P(C|A), P(\neg C|A),$
$P(B|\neg A), P(\neg B|\neg A), P(C|\neg A), P(\neg C|\neg A),$

Conditioned on $B$:
$P(A|B), P(\neg A|B), P(C|B), P(\neg C|B),$
$P(A|\neg B), P(\neg A|\neg B), P(C|\neg B), P(\neg C|\neg B),$

Conditioned on $C$:
$P(A|C), P(\neg A|C), P(B|C), P(\neg B|C),$
$P(A|\neg C), P(\neg A|\neg C), P(B|\neg C), P(\neg B|\neg C),$

(c.2) For two variables conditioned on one variable, there are $2^2 \times 2 \times 3 = 24$:

Conditioned on $A$:
$P(B,C|A), P(\neg B,C|A), P(B,\neg C|A), P(\neg B,\neg C|A),$
$P(B,C|\neg A), P(\neg B,C|\neg A), P(B,\neg C|\neg A), P(\neg B,\neg C|\neg A),$

Conditioned on $B$:
$P(A,C|B), P(\neg A,C|B), P(A,\neg C|B), P(\neg A,\neg C|B),$
$P(A,C|\neg B), P(\neg A,C|\neg B), P(A,\neg C|\neg B), P(\neg A,\neg C|\neg B),$

Conditioned on $C$:
$P(A,B|C), P(\neg A,B|C), P(A,\neg B|C), P(\neg A,\neg B|C),$
$P(A,B|\neg C), P(\neg A,B|\neg C), P(A,\neg B|\neg C), P(\neg A,\neg B|\neg C).$

(c.3) For one variable conditioned on two variables, there are also $2 \times 2^2 \times 3 = 24$.

Conditioned on $A$ and $B$:
$P(C|A,B), P(C|\neg A,B), P(C|A,\neg B), P(C|\neg A,\neg B),$
$P(\neg C|A,B), P(\neg C|\neg A,B), P(\neg C|A,\neg B), P(\neg C|\neg A,\neg B),$

Conditioned on $A$ and $C$:
$P(B|A,C), P(B|\neg A,C), P(B|A,\neg C), P(B|\neg A,\neg C),$
\[ P(\neg B|A,C), P(\neg B|\neg A,C), P(\neg B|A,\neg C), P(\neg B|\neg A,\neg C), \]

Conditioned on B and C:
\[ P(A|B,C), P(A|\neg B,C), P(A|B,\neg C), P(A|\neg B,\neg C), \]
\[ P(\neg A|B,C), P(\neg A|\neg B,C), P(\neg A|B,\neg C), P(\neg A|\neg B,\neg C). \]

Learning from Bayesian network highlights computations on conditional probabilities, and computing conditional probabilities take up most of our calculations. We can hardly get all the results of calculations in a manageable time. When the Bayesian network structure is unknown, every probability may be calculated in order to get the best estimation for the optimal structure. However, not all calculations are necessary and not all results are interesting to us in a specific domain of research if there is a known structure.

Exact inference from large-scale data is a NP-hard problem \[16\] \[18\]. Researchers consider feasible approaches for approximate inferences. Several methods have been developed \[58\]. These include but are not limited to Monte-Carlo simulation, Monte-Carlo Markov Chain (MCMC), Gaussian approximation, Expectation Maximization Algorithm, Maximum likelihood, and so on.

This research mainly focuses on learning structures of the Bayesian network, we will show the details of the problems, the complexity of learning structures with large-scale data, and the solutions to the domain-specific fields of knowledge in the subsequent sections and chapters.

2.4. Heuristics

**Heuristic**, as an adjective, by definition, means “serving to discover” \[105\]. As it was stated, “**Heuristic reasoning is reasoning not regarded as final and strict but as provisional and plausible only, whose purpose is to discover the solution of the present problem.**” Heuristics can also be viewed as uncertainty of outcome, basis in incomplete knowledge, improvement of performance, and guidance of decision making \[112\]. They represent compromises of requirements between the need of making the criteria to be simple and the desire to see them discriminate correctly between choices \[97\]. Such criteria are determined to provide the most effective actions to reach the research goal. In our words, we may trade completeness,
accuracy, and reaching the optimal solution for a fast speed, a relatively satisfactory solution and yielding several reasonable results.

Heuristic approaches are used in a variety of fields [21]. For example, a heuristic search method of a traveling sales person can be implemented by traversing most of the large cities in a quick time but ignoring smaller cities. In the area of artificial intelligence such as in chess games, a heuristic algorithm by machine can achieve several “good steps” even though such steps may not be optimal if we consider all possibilities of subsequent steps. An antivirus application may also use heuristic method to detect malware by analyzing the behaviors of the programs.

There are also potential uses in epidemiological studies for heuristics. Heuristics play an essential role through reducing computational complexity of learning causal structures from the Bayesian network. [40]. Several reasonable assumptions such as combining background knowledge [35] are made and approximate learning algorithms such as local causal discoveries are developed to obtain the near optimal solutions for variables’ causal relationships.

2.5. Causation and Its Definitions

There are different definitions and inference methods of causality in literature [60] [68]. In epidemiology [95] [113], especially for studying chronic disease, Bradford Hill [61] listed the following criteria:

(1) Strength. The stronger an association, the more likely to be causally related. However, the conclusion based on strength of association only may be problematic. Strength of association can be measured by the value of odds ratio.

(2) Consistency. Findings in different scenario such as by different researchers, with different methods or at different places are consist regardless.

(3) Specificity. A cause should result in one effect. However, this condition does not stand since effects may be caused by a common factor.

(4) Temporality. A cause should always precede the outcome or consequence.
(5) Biological gradient. One example is that an increase of the disease rates corresponds with the increase of the exposure.

(6) Plausibility. By human’s judgement, plausibility means that the association between two factors should make sense.

(7) Coherence. This term is similar to plausibility, the explanation of a causal relationship should fit with facts or biological mechanisms.

(8) Experiment. It is essential for observing a change in the outcome with the change in exposure under the controlled experiment.

(9) Analogy. Finding from the assertion should have similar situations in the past.

Among those criteria, only temporality is the necessary condition that a causal relationship can be established. However, it is hard to oversee a causation in most observational dataset based on the criterion of temporality, since there is no temporal relationship explicitly stated in those datasets.

The definition by Rothman [114] says that “A cause of a specific disease event as an antecedent event, condition, or characteristic precedes the disease event and without such conditions, that event either would not have occurred at all or would not have occurred until some later time.” Rothman’s definition does not provide a complete causal mechanism but only a component of it.

A common question is asked, as not all smokers develop lung cancer but some non-smokers do, why do we still correlate smoking as the cause of lung cancer? The question may be simply answered by treating it as a probabilistic issue. Moreover, if the research shows that lung cancer does not cause or prevent smoking, the statement that smoking causes lung cancer will be more convincing.

According to the definition of causality by Pearl [102], the causation is expressed as:

\[ P(\text{outcome}|do(\text{cause})) > P(\text{outcome}|do(\text{not cause})). \]

\[ P(\text{outcome}|do(\text{cause})) \] is the probability of the outcome under the condition where there is a manipulated action, the potential cause. \( P(\text{outcome}|do(\text{not cause})) \) is the probability of the outcome under the condition where there is no such action. The causal relationship
can be established from the cause to the outcome when the probability of the outcome under one manipulated exposure is greater than the same outcome where there is no manipulated exposure. The do-operator emphasizes that it is an intervention, and it compels the happening of the cause. In the real world, people may not be able to strictly follow this procedure by conducting all the underlying experiments not only because of the moral issues but also due to various internal reasons in the analytical stage such as limits of population size, difficulties of measurements, and biased observations.

In this research, we do not use the definition based on the abovementioned do-operator’s expression, because in an observational dataset there is no record that can enforce the intervention. We consider causality as a condition of possibility to happen, when there is a strong relationship between two variables and when it is more likely a variable precedes another variable. Accordingly, a causal structure discovered from this research should tell the indication rather than the affirmation of causal relationships among variables. A causal network depicts the functional relations and a causal model is viewed as a system of structural equations that show how variables are inter-related.

Even though to this point, we do not fully address or solve the issues of variable interactions, confounding-related biases, erroneous data problems, or any other situations that may make our assertions less plausible in epidemiology. As in an observational dataset most causal relationships are previously unknown, this research can still be practical in causal reasoning if researchers are able to find something interesting. If computer scientists could delineate a scope of possible causal relationships, or discover several possible or highly likely causal links from data without external clinical experiments, results would be helpful for medical professionals to conduct further experimental studies by saving time and cost.

2.6. From Bayesian Networks to the Discovery of Causality

2.6.1. Learning Structures of Bayesian Networks

In machine learning and artificial intelligence, researchers analyze Bayesian networks by learning data or by learning structures\[57\]. For structural learning of Bayesian networks, there are score based methods such as Bayesian Information Criterion(BIC)\[118\],
constraint based methods such as PC algorithm[122], and hybrid methods such as Max-Min-Hill-Climbing(MMHC)[126]. Score based learning methods calculate scores for possible network structures and output the optimal structure with the highest score through search. Constraint based learning methods check conditional independence among variables given a valid statistical testing and backed by d-separation properties[99]. A hybrid approach absorbs the merits from both score based and constraint based learning methods.

We recall the important features in the Bayesian network, conditional independence and conditional dependence shown below:

**Conditional Independence:** For three variables $V_1$, $V_2$, and $V_3$, suppose that the distributions of $V_1$ and $V_2$ are statistically independent conditioned on $V_3$, we say that $V_1$ and $V_2$ are conditionally independent given $V_3$ and write the form as $V_1 \perp V_2 | V_3$.

**Conditional Dependence:** Suppose that the distributions of two variables $V_1$ and $V_2$ are statistically independent from each other but these two variables become statistically dependent conditioned on a third variable $V_3$, we say $V_1$ and $V_2$ are conditionally dependent given $V_3$. Such a type of structure is referred as a v-structure or a collider in graphical models and in epidemiology.

The statistical phenomenon of the v-structure or the collider which was originally observed in medical statistics known as Berkson's bias [10], shows how two independent variables explain away the outcome influenced by both variables. The feature of a v-structure, however, enables researchers to discover possible causal structures among these three variables.

No matter which method one chooses to learn the structure of Bayesian networks, the learning process is time consuming and requires abundant memory space if the database contains a large amount of data with many variables. Generally, exact learning is a computationally hard problem [16] [18] and there are many existing approaches to alleviate the difficulty of this problem. These approaches include Greedy Bayesian Search Algorithm [123], Sparse Candidate [42], Greedy Equivalence Search [17], Three Phase Dependency Analysis [15], Optimal Reinsertion [86] and many more. Furthermore, the efficient learning from
Bayesian networks is propelled by saving both runtime and search space. Solutions such as recursive conditioning [30] or the implementation of parallel learning method [94] take both time and space complexity into account for the trade-off to balance time and space.

Instead of any methods for exact learning [70], several heuristic methods mentioned above are considered giving approximate or near optimal results to avoid ultra exponential computations. Assumptions can be made such as pre-setting maximum number of in-degree, out-degree, or the depth of the graph, eliminating edges according to the strength of association, and assuming all parents of the target variables are independent from each other, or with some known ordering of variables.

Information or constraints pertaining to the prior structures may be assumed in a general learning scenario. For example, Chow-Liu tree [19] or a polytree [107] assumes that there is no more than one parent of a variable in any of the directed graph structure. Others may be reasonably assumed in domain-specific areas. For example, in a causal structure of the gene network, it is more likely that a gene causes many other genes than that a gene is caused by many others. With this assumption, there will be no more than a specific number of parents for any one of the variables in the structure [93], thus the size of conditional probability table used for structural learning is decreased.

2.6.2. Inferring Causal Structures in Bayesian Networks

Bayesian networks are discussed and analyzed in a variety of fields such as data mining [56], bioinformatics and computational biology [38] [41] [90], medical service [3], biomedical applications [82], decision support systems [33], medicine [5] [27] and epidemiology [2] [46]. Moreover, initiated by a renewed interest for understanding causal concepts with modeling complex stochastic systems [7], Bayesian networks have natural causal interpretations and thus causal relationships can also be discussed and analyzed. However, a Bayesian network is not a causal network. Issues related to the differences of two networks have been theoretically discussed [104] and methods with a Bayesian approach for learning causal networks have also been studied [55]. To establish a truly causal relationship, experimental study is needed, all variables must be unconfounded, relationships among variables
should be non-parametric, and causal relationships in the network should be represented correctly with no incomplete data and selection bias. However, inspired by methods of inferring possible causal structures through the learning of Bayesian network, we state here that our methods and research process are intended for discovering statistically possible causations rather than obtaining any truly causal relationships in an experimental setting.

A constraint-based method for structural learning tries to discover v-structures of variables in order to orient edge directions from the undirected graph. For example, if there are three variables A, B, and C, and suppose A and B are independent given no other variable, but conditionally dependent given a third variable C, then we obtain a v-structure as $A \rightarrow C \leftarrow B$. Such a structure showing the directions of edges represents causal relationships where both A and B are possible causes of C. If there is no edge between a pair of two variables in the graph, there is no causal relationship between these two variables.

2.6.3. Research of Causality in Health Data

A public health dataset contains data from biomedical research, demographics, patients in hospitals, medical experiments, surveillance, epidemiology, and so on. Most of publicly available datasets obtained from disease surveillance and cross-sectional study are observational. There are several advantages of using observational data rather than experimental data. For observational data, there are relatively large number of records and they are easily accessible. However, due to their nature, observational data are mostly non-temporal, consist of a large number of irrelevant variables within a particular research focus (e.g. if they are not associated with a specific disease), and often have a relatively low quality compared with experimental data since erroneous and incomplete records exist.

The first problem regarding to temporality creates an obstacle for causal inference. The second problem regarding to the number of irrelevant variables results in computational difficulties. The third problem regarding to poor data quality such as data completeness may reduce the accuracy of any research outcome.

This research only focuses on the first two problems and provides solutions to them. To overcome the first problem, we utilize previously stated methods to discover possible
causal structures from the data. To alleviate the difficulty of the second problem, we propose heuristic learning methods. One consideration is to use reasonable assumptions and existing background knowledge to reduce computational complexity of structural learning [59]. Especially in health science research, there are many established temporal relationships and confirmed causal relationships among variables. This research will show how we exploit existing knowledge of variables’ relationships to learn the causal structures efficiently.

Causal inference in health science study is necessary and of interests to epidemiologists and medical professionals to study diseases. Motivated by the need and research advances [103] of discovering and analyzing causal relationships from observational data, we provide heuristically novel methods for local causal discoveries and for learning causal structures combined with background knowledge in health science research.

2.7. Research with Big Data

Big Data becomes a popular term from computer scientists to the general public in the twenty-first century. With the growing complexity of data management and processing, Big Data problems cover all aspects such as acquisition, storage, search, sharing, and analysis. They differ from traditional data problems in that Big Data needs novel methods to capture, store, extract, protect, and process them effectively and efficiently.

Even before “Big Data” was coined, computer scientists had already realized the problems due to the growth of data storage and the need of efficiency for data processing in the relational database management systems. Consequently, three “V”s of data: volume, velocity, and variety were mentioned and discussed first[73]. In addition to the original three “V”s, several other “V”s have been embraced into the category of Big Data to represent its features, such as validity, value, viability, and even visualization.

To the current knowledge, Big Data is most often referred to five “V”s: volume, variety, velocity, veracity, and variability.

(1) Volume: Volume refers to the feature that there is need to provide sufficient space and efficient methods to store or retrieve information in the database. It reflects the growing scale of data.
(2) Variety: Data can be stored in multiple forms and from different locations, such as in the table of a file or in a cloud database. Inconsistent forms of data is accompanied with the difficulty and the cost of processing or rearranging unstructured data.

(3) Velocity: From a batch file to the real-time information of data, researchers need methods to capture, store, curate, and analyze data in a time-sensitive environment. Issues concerning the effectiveness of processing data are important when researchers are dealing with streaming data.

(4) Veracity: Veracity reflects the problem with data due to their poor quality or people’s uncertainty of data. Extra work is needed for researchers to evaluate and ensure the quality of results when there is a large amount of misinformation of data.

(5) Variability: Inconsistency of data at times or different data sources selections may limit researchers providing reliable results and making affirmative conclusions.

Besides these five “V”s, the complexity of managing and analyzing data cannot be overlooked when we are performing certain tasks. Process of solving a specific computational problem itself may already be challenging, not to mention the situation when data are changing, unstructured, and in large volumes.

Beyond the applications in a variety of fields such as database management, information retrieval, and social network analysis, Big Data also stretches to the research in medicine and health [64] [128]. Big Data is closely related to epidemiology in terms of processing data to facilitate decision making. This research, however, does not address all issues in Big Data such as features of variety and velocity due to the nature of collected observational dataset. For the volume of data, there is not a problem regarding to the physical storage of raw epidemiological datasets but there is a problem for storing and searching the generated data structures from the dataset. There are also veracity and variability problems in the observational dataset due to either researcher’s biased selections of samples and their survey preference or survey takers’ mishandling. Nevertheless, to identify every possible causal relationship by testing all combinations of variables is computationally infeasible. Discovering the optimal structure from all variables requires massive searching space and abundant stor-
age. Irrelevant information and incomplete data also hamper the process for efficient causal discoveries.
3.1. Research Assumptions

Several assumptions are made to facilitate our research. We include fixed criteria, such as what characteristics the dataset has to incorporate in general and what hypothesis we have to make in order to fit the models throughout the entire research. We also include expansible criteria that are only related to particular datasets in epidemiology, implementations and research focus. By adding and adjusting requirements in epidemiological study and combining previous research assumptions in literature[23] [120], we summarize them as follows:

Assumption 3.1. Database completeness and data integrity.

We assume that each record is stored in one single database, and such recorded data are uncorrupted. Though it is also possible to combine data from different sources, we do not consider processing different databases with multiple resources. We use a single database with complete data only. All types of incomplete data such as “non-response”, “refused to answer” or “classified as confidential” are excluded and all issues related to data quality problems (e.g. missing data or erroneous data) are resolved prior to processing for causal reasoning. Our research does not focus on any improvements of data quality as well [8].

Assumption 3.2. Bayesian network causal model.

Conditions of Bayesian network causal model are met. They include that there is no cycle in the model, whereas in reality a cause and an effect could be bi-directional. In a word, all variables under this assumption should be displayed as either a cause or a consequence of the other variable, given that two variables are correlated. Other requirements for a Bayesian network should also be met. These include but are not limited to a directed acyclic graph where nodes represent variables and edges represent probabilistic dependence and two
variables being independent if there is no edge between them.

**Assumption 3.3.** No selection bias.

Probability distribution over the dataset should be equal to the distribution over our model. This assumption may not always hold in all datasets. For instance, some surveys do not cover the whole population, but rather they are designed to focus on a specific group, an arbitrarily selected region, or a randomly chosen community. A selection bias may occur anyway. For example, we study multiple diseases and their possible associations but select the surveyed people in hospitals only. The use of surveillance dataset can make the selection bias less likely to happen. However, even though most national surveillance surveys cover the entire population and the surveyed cases represent the proportion of the population, surveys themselves are not designed for causal discovery and analysis primarily. For the use of any real datasets and only for the purpose of our computational approaches, we assume the perfection of data with respect to the methods of sampling and the consistency of probability distribution in both entire population and the surveyed samples.

**Assumption 3.4.** Causal Faithfulness.

If one variable is the cause of the other, these two variables will be correlated. Correlation should be a necessary condition for causation in our model regardless though it does not imply causation. No correlation between two variables in this research implies there is no causation. A condition can be either presence (as positive) or non-presence (as negative) as related to the other condition. Variables X and Y are conditionally independent given a set Z with respect to a probability distribution P if \( P(X, Y | Z) = P(X | Z) \times P(Y | Z) \). Causally faithful relationships meet the Markov condition or the d-separation condition [99].

**Assumption 3.5.** Valid statistical testing.

The test results for independence or dependence are consistent with the real condition for dependencies. A proper threshold to decide the level of independence or dependence may
be determined with the construction of our causal network.

**Assumption 3.6.** Discrete variables or Boolean variables.

We do not provide any solutions to handle continuous data in this research. If the original dataset contains continuous data, a conversion to several different discrete categories is needed. For instance, the measurement of age or height is continuous, but we can group the numbers into several categories such as Age between 11-20, Age between 21-30, Height between 150 cm to 159 cm, Height between 160 cm to 169 cm. Analysis through categorizing data is for the feasibility purpose.

Moreover, for the convenience of calculations we make the assumption that there will be no polytomous data, as we consider our simplified model with Boolean data only. Polytomous data are the type of data in which the exposure or outcome has more than two levels. [114]. For example, a symptom of disease can be described as “none,” “slight,” “moderate,” and “severe”. We use dichotomy by comparing only two variables such as “none” versus “severe” in a simplified model. This assumption is not true in epidemiological research used in practice though, we may use Boolean values to get a raw estimate on variables’ relationships with simplified calculations.

**Assumption 3.7.** Structural models for simplicity.

Researchers may consider unmeasured or unobserved variables, whereby these variables may influence measured variables [96]. Although structural equation models (SEM) are widely used in social science, in this research, a graphical model does not include the structural equation model. A corresponding graph here only shows the structural or systematic relations among observed variables. Under this assumption, there should be no measurement error or any variations of randomness. This research defines a structural model such as a Bayesian network where there are only variables and their probability distributions.

All above assumptions are followed in our research method. In this chapter, we describe the methods by showing a framework of learning causal relationships in an epidemio-
logical dataset. We give a novel approach of locally learning the structure of a causal network with background knowledge. Then we describe the process of building a large causal network from local causal structures. The process consists of methods for local causal discoveries and subgraph decompositions.

3.2. A framework of Learning the Causal Network

We present a framework of data processing and causal discoveries, practical to analyzing epidemiological data. Steps of the workflow are shown below (Figure 3.1).

![Workflow of Data Processing and Causal Discoveries](image)

**Figure 3.1.** Workflow of Data Processing and Causal Discoveries

In order to obtain the filtered data that can fit our computational model, we include one step named as data pre-processing (step a) such as filtering ambiguous data, eliminating variables with very low frequencies, and excluding all missing, erroneous and invalid data. Variable eliminations and conversion to discrete or boolean data may be necessary for analyzing data in our model. The purpose of performing such transformations on data is to let us obtain cleaner and more simplified data for causal inference, as well as to reduce subsequent complexity of computations. We do not describe all these problems and their solutions in detail since this research focuses on learning causal structures based on the assumptions after we already have the desired structural data from the data.
Combined with background knowledge, we construct a multi-layer structure of the entire network (step b). Variables on top of the layer are known acausal factors, while variables at the bottom of the layer are known consequences which do not cause or directionally influence other variables (e.g. death). More details of the multi-layer structure will be explained with examples in the illustration section.

Algorithms for local causal discoveries are implemented (step c), following the construction of the multi-layer structure. A local causal discovery is defined in our study as a heuristic approach through testing three chosen variables repeatedly, and this approach can be solved in linear time. The output results will show several potential causal and non-causal conditions from underlying data, even though future examinations for causal relationships are necessary. We will describe the methods in more detail later in this section.

A large causal network is then constructed from prior local causal structures (step d), in which all statistical testing and constraints of independence hold. A larger network provides researchers graphical representations of relationships between conditions and diseases, and give them easily identifiable and understandable information for their analysis and external validations.

Step a is for data processing and data mining, which we do not elaborate details due to the focus of this research. Steps b, c, and d are more specifically related to causal discovery and network structure learning. Step e is left for bio-statisticians and medical professionals. In the following section, we emphasize on the introduction to the current methods for local causal structure learning, the expansions of one existing algorithm, and several attempts to build larger global network structures.

3.3. A Knowledge-Based Approach of Structural Learning

We design a novel method of learning the structure of a causal network of diseases. This approach is constraint-based and background knowledge-based. Rather than score-based methods [39] such as calculating Bayesian Information Score to find the optimal structure of the network [118], constraint-based methods test conditional independence and ultimately direct the edges to build the structure. This approach is for previously mentioned
Constraint based methods for learning the structure of Bayesian networks have been introduced and developed, with several as heuristic approaches that do not require exponential time [31]. These methods rely on several assumptions such that there is a known cause to other variables, there is no other known cause of these variables, or there is no hidden or intermediate variables between known causal relationships. This approach incorporates some variables known to us as factors who either have no known causes or have no known consequences.

There are three main reasons to incorporate background knowledge in the model. First, this research has a specific focus on epidemiology, not on the learning process in a generic network model. Second, it is silly to ignore what we have known and what knowledge we are almost certain about. We utilize the background knowledge such like “gender of one individual is not determined by a condition of the disease,” “death cannot cause one person’s disease symptoms,” or “accidental body injuries does not affect previously long-lasting malnutrition”. Third, when the background knowledge is incorporated in our computational model, the modeling and causal inferring process will produce both efficiency and accuracy. Factual correctness provided by background knowledge warrants accuracy of the output structure from the learned network.

Our research on the structural learning of a causal network differs from a generic modeling in that a theoretical learning process does not yet incorporate, specify, or process any background knowledge. The reason why incorporating background knowledge is necessary to achieve accuracy in reality for causal structural learning is pertaining to the concept of equivalence relations in the Bayesian network. Structures of variables in the same equivalent class in a Bayesian network will have d-separation properties [45] [100]. Given that multiple structures that are in the same equivalence class (e.g. the same probability distributions, score equivalent), we are unable to infer the exact structure from them only through computations. For example, if we have two variables A and B only, we will not be able to compute the value and tell the correct structure (if A causes B, or if B cases A)
unless certain prior background knowledge is incorporated (such as we already know A has no known causes, so the direction will be from A to B only).

3.4. Process of Discovering Causal Relationships

A Bayesian network is not a causal network [102]. We take the concepts of conditional independence and graphical representations to indicate causal or influencing relationships among variables in Bayesian networks. As stated in previous sections, we do not fully distinguish a causal network and a Bayesian network, while we take a weaker form of definition of causality where a directional relationship in a Bayesian network suggests causal relationships in a graphical model. Given that all variables are unconfounded, and a directed edge from one node of variable to the other suggests directional influence, the following process, derived from structural learning of a general Bayesian network [92] [122] as well as the adjustments made by us, shows the process of learning a causal network in an epidemiological dataset.

(1) Orienting undirected edges based on simple associations

We add an edge between two nodes of variables if they are correlated. This step does not perform any conditional independence test due to time complexity of computations. In this research, we use “correlation,” “association,” and “dependence” interchangeably, which are all related without the specification of directions. The association tests can be completed in linear time by comparing all pairs of two different variables. Suppose we have n variables, it will take O(n^2) to get values of all associated pairs. We draw an undirected edge between these two variables if they are correlated.

(2) Directing edges from the undirected graph

We repeatedly test conditional independence from the preliminary structure with all undirected edges. We remove edges between two variables if they are conditionally independent given any third variable. Originally, the skeleton of the Bayesian network (i.e. the graph in which no more edges will be added and directions of edges are pending examined) is drawn through complete conditional independence test. For any edges A_1 and A_2, A_1-A_2 is part of the skeleton if and only if A_1 and A_2 are not conditionally independent given any X, for all X not A_1 or A_2.

In this step, rather than picking up nodes with lowest degree in the dataset, we
select variables from top layer (known as causal factors) or from bottom layer (known as consequences that do not cause other factors) first. Then we test conditional independence between each selected variables and other ones in the middle layer. The directed edges will be drawn based on existing partially known causal relationships from the skeleton. For example, if there is a variable “born with low weight”, which is placed on the top layer as one of the acausal factors, then conditional independence tests will be given for all variables (such as an individual’s all kinds of health problems at age of six) originally associated with this variable, even the node denoting “born with low weight” has many edges connected to it (the out-degree is high). Similarly, conditional independence tests will be conducted for all variables on bottom of the layer as well. After these procedures, we will get a partially directed graph from the skeleton. Next we choose the nodes with already established directed edges for conditional independence tests, and finally direct all edges from the skeleton.

Next step is to introduce v-structures. A v-structure is defined as the structure for variables A, B, and C, A causes C and B cause C, but A and B are previously independent from each other \((A\rightarrow C \leftarrow B)\). The existence of v-structures allows us to infer causation from observational data. If two previously independent variables become conditional dependent given a third variable, then there is a v-structure containing A, B, and C, where A causes C, and B causes C.

After applying tests for discovering v-structures, we avoid any new v-structures such that if variable A causes B while B and C are correlated \((A\rightarrow B\rightarrow C)\), the direction of the edge between B and C can only be from B to C \((B\rightarrow C)\).

Next we avoid any cycles such that if A causes B \((A\rightarrow B)\), B causes C \((B\rightarrow C)\), while A and C are correlated \((A\rightarrow C)\), then the direction of the edge between A and C will be from A to C \((A\rightarrow C)\).

Last, rather than choosing a direction randomly for all remaining edges (for complete learning only) which cannot be deployed in any of the previous steps, we may leave the questionable edges as their causal relationships are still unclear.
3.5. Conditional Independence Tests and Local Causal Discoveries

3.5.1. First-order Conditional Independence Test

Complete conditional independence tests require far more than polynomial time. Hence, we consider first order conditional independence for two variables’ testing only. First order conditional independence is shown as $A_1 \perp A_2 | B_1$, whereas nth order is $A_1 \perp A_2 | (B_1, .. , B_n)$. We may still obtain some useful information from this constraint. The computations can be reduced to $O(n^3)$ in the worst case.

**Algorithm 1** First order conditional independence test

1: For all n variables $v_1 \rightarrow v_n$:
2: \textbf{for} $i \leftarrow 1$ to $n$ \textbf{do}
3: \textbf{for} $j \leftarrow 1$ to $n$ \textbf{do}
4: \textbf{for} $k \leftarrow 1$ to $n$ \textbf{do}
5: \textbf{if} ($i \neq j$ AND $j \neq k$) \textbf{then}
6: Test CI ($i, j | k$)

CI($i,j|k$) denotes conditional independence for variables $i$ and $j$, given a variable $k$.

3.5.2. Algorithms for Local Causal Discoveries

To further reduce computations, algorithms for local causal discoveries can be implemented. Two algorithms for local causal discoveries have been developed, namely the CCU rule and the CCC rule[23] [120].

**Algorithm 2** CCU Rule

1: For all variables $A, B, \text{ and } C$, assuming that neither $B$ nor $C$ is the cause of $A$:
2: \textbf{if} D($A, B$) and D($B, C$) and I($A, C$) \textbf{then}
3: \textbf{for} all variables $B \notin [C,A]$ \textbf{do}
4: \textbf{if} CD($A, C; B$) \textbf{then}
5: \textbf{output} $C \rightarrow B$. 
Algorithm 3 CCC Rule

1: For all variables A, B, and C, assuming that neither B nor C is the cause of A:
2: if D(A, B) and D(B, C) and D(A, C) then
3: for all variables B $\notin [C,A]$ do
4: if CI(A, C; B) then
5: output B $\rightarrow$ C.

D($x, y$) indicates that variables $x$ and $y$ are dependent while I($x, y$) indicates that variables $x$ and $y$ are independent. CD($x, y; z$) indicates that variables $x$ and $y$ are conditionally dependent given variable $z$.

We have to assume, for both rules, that A has no known cause (i.e. acausal factor) under the testing dataset. It is likely that A may be caused by different variables depending on other three-variable selections. Under this algorithm, variable A is assumed to be the root which has no known causes within the set of three variables A, B, and C. Since A has no causes, and both algorithms state that A and B are dependent or correlated, we can conclude that A causes or directionally influences B from assumptions stated previously.

In reality, however, people may be more interested in finding causal factors for a certain disease. People’s knowledge of choosing the right acausal factors from a given dataset may be limited, but we do know that one condition or one particular disease, has no known consequences or effects from all other variables. For example, a variable denoting “death” does not have its influenced variables but may only have several conditions that causes death. We make use of both rules for local causal discoveries as well as the extension to CCC rule, namely CCCx rule, to exclude variables for non-causal connections. Inspired by the abovementioned situation where a certain condition has no known influence on other conditions, we extend the current CCC rule, namely the CCCx rule:

All abovementioned cases follows the feature of Markov condition in a causal network, and the CCCx rule has an extensive use for filter non-causal factors. The algorithm shows how we can exclude non-causal factors given the premise that conditioning on any one of the variables makes two previously dependent variables conditionally independent from each
Algorithm 4 CCCx Rule, an extended version of CCC:

1: For all variables A, B, and C, assuming that neither A nor B is caused by C:

2: if D(A, B) and D(B, C) and D(A, C) then

3: for all variables B $\notin [C,A]$ do

4: if CI(A, C; B) then

5: output $A \not\rightarrow C$.

6: if A has no known cause then

7: output $A \rightarrow B, B \rightarrow C$.

8: if B has no known cause then

9: output $B \rightarrow A, B \rightarrow C$.

other. If two variables in the local structure are conditionally independent, there will be no causal relationship between these two variables.

We show how to expand the version of CCCx rule to all n variables, for one of the integral steps towards structural learning of a global network.

Let the variable c represent a condition or a disease which does not cause any other variables in a given finite set S. S represents the finite set of all n variables in which each variable is correlated with variable c and at least one other variable. Let R be the set of variables in which variables remains correlated with c after conditioning process.

Input: $S= v_1, v_2, ..., v_n$, where $v_1, v_2, ..., v_n$ are variables in set S.

Algorithm 5 Discovery of non-causal factors

1: Init: $R \leftarrow S, i, j$.

2: // while R isn’t empty

3: for any variable c, (ultimate consequence) do

4: for $i \leftarrow 1$ to n-1, $v_i \in R$, do

5: for $j \leftarrow 1$ to n-1, $v_j \in R$ do

6: if D($v_i, v_j$) then

7: if $v_j \perp c \mid v_i$ then

8: $R \leftarrow R - v_j$
R represents the set of all remaining variables from the original set S. These remaining variables could be further examined by medical professionals through experiments. The excluded set, E = S - R, contains all variables that are not related to the target condition or disease, i.e., variable c, after conditioning process for some intermediate variables. None of the variables in the excluded set E will be the casual factor of variable c.

In the best case, this algorithm can execute to finish in linear time O(n), i.e., conditioning on the first variable $v_1$ can make all other variables from $v_2$ to $v_n$ independent from variable c, and then there will be no need to test other variables since they are already acausal to c. In the worst case, this algorithm may run as much as $O(n^2)$ time, if no variable can be excluded from the dataset S.

This algorithm incorporates three predetermined conditions. First, it does not condition on a variable again which was excluded from the remaining set in any previous steps (shown in line 4, such that $v_i$ must belongs to R). Second, it does not test for conditional independence for a variable again which was excluded from the remaining set in any previous steps, given any other third variable (shown in line 5, such that $v_j$ must belong to R). Third, it does not test for conditional independence of two variables if these two variables were previously independent (shown in line 6, such that $v_i$ and $v_j$ must be dependent with each other).

If two variables are independent conditioned on any third variable being observed, these two variables will not be directly connected to each other in a graph. Since two variables are separated by other variable, a causal relationship cannot be established.

This rule cannot identify all causes of a specific condition, but can exclude several factors that may not directly cause such a condition. Exclusions of variables that are not related to a particular disease are as important as discoveries of possible causal factors. Excluded variables may not be considered as real causes of diseases or conditions. Only variables that are still included in the post-screening dataset have to be examined by medical professionals for experimental studies.

The proofs of correctness will be shown in appendix section.
3.5.3. Subgraph Decomposition

We take another approach, besides implementations of local causal discovery algorithms, to improve the performance of structural learning and analysis of causal relationships. We name it subgraph decomposition.

The idea is to break the big graph containing all variables in the middle layer into several smaller components. Each component of the graph is viewed as a subgraph where any nodes are connected by path but no additional edge is connected to these nodes by any path in a supergraph.

**Definition 3.8.** Given a dataset D, for all variables excluding variables from both top layer and bottom layer in the dataset, if there is subgraph containing some variables in the middle layer of our three-layer network, where the removal of all edges coming from the top layer variables (if they exist) and going out to the bottom layer variables (if they exist) can make such subgraph disconnected with any other variables outside the subgraph through any path, then this subgraph is one component of the entire graph in our disease network.

We use the following (Figure 3.2) to exemplify the definition.

![Layered Causal Network of Disease](image)

**Figure 3.2.** An example of Layered Causal Network of Disease

**Example 3.9.** If there is an isolated variable (e.g. variable M), then it is already a compo-
Example 3.10. Variables H and I consist of one component, since they are disconnected with any other variables after edges from C (top layer) to H and from I to G (bottom layer) have been removed. Variables J, K, and L consist of another component.

The procedure of learning causal structure on subgraph decomposition is as follows.

a. Break the entire graph in the middle layer into several components followed by the definition after the construction of three-layered network structure.

b. For each component, apply all local causal discovery rules and test conditional independence for all variables, combining with existing background knowledge from two other layers.

c. Repeat the process of breaking a component into smaller components if several conditional independence tests can make some variables in one component conditionally independent given other variables after applying all mentioned methods, and repeat local causal discovery rules, until finishing the construction of the structure in each subgraph.

We have made assumptions that variables from the top layer and bottom layer are known as background knowledge layers. If there is an edge connecting a variable $C_1$ from top layer to another variable $C_i$ not in that layer, then the direction of the edge can only be from $C_1$ to $C_i$, as $C_1$ causes $C_i$. Similarly, if there is an edge connecting a variable $C_2$ from bottom layer to another variable $C_j$ not in that layer, then the direction of the edge can only be from $C_j$ to $C_2$, as $C_j$ causes $C_2$. Given that there is no unknown path coming into or going out of the subgraph, excluding known causal directions, variables inside our defined component are causally independent from all variables in other components. By breaking the graph into several components, we can learn the structure for each subgraph separately using the rules and methods mentioned in this research.

The reason why it is of importance is that each component of graph can be learned independently for the construction of network structure. When the structures of each component have been learned, they can be integrated into bigger graphs, therefore a global
causal network will be constructed timely. A single machine or only one processor may fail to handle the complete graph in a reasonable time and space because the number of variables is huge; however, by introducing parallel processing for solving subgraph problems, we will be able to assign tasks into multiple job pools and the structural learning process for each job can be executed in parallel with the premise of causal independence of each subgraph.

Time complexity for learning structures of all subgraphs will largely depend on recovering the structure of giant components. This method may be extended to all known directed edges in a graph for decomposition.
4.1. Growth of the Number of Directed Acyclic Graphs

Inference from a Bayesian network can be computationally infeasible. Exact learning is a computationally hard problem [16] [18]. It has been shown that the number of different labeled directed acyclic graphs (DAGs) grows more than exponentially with the increase of nodes[111]:

\[
f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)!i!} 2^{i(n-i)} f(n-1)
\]

Table 4.1 shows the results how the number of different structures grows with the number of variables [92].

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Number of DAGs</th>
<th>Number of Nodes</th>
<th>Number of DAGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>13</td>
<td>1.9E31</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>14</td>
<td>1.4E36</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>15</td>
<td>2.4E41</td>
</tr>
<tr>
<td>4</td>
<td>543</td>
<td>16</td>
<td>8.4E46</td>
</tr>
<tr>
<td>5</td>
<td>29281</td>
<td>17</td>
<td>6.3E52</td>
</tr>
<tr>
<td>6</td>
<td>3.8E6</td>
<td>18</td>
<td>9.9E58</td>
</tr>
<tr>
<td>7</td>
<td>1.1E9</td>
<td>19</td>
<td>3.3E65</td>
</tr>
<tr>
<td>8</td>
<td>7.8E11</td>
<td>20</td>
<td>2.35E72</td>
</tr>
<tr>
<td>9</td>
<td>1.2E15</td>
<td>21</td>
<td>3.5E79</td>
</tr>
<tr>
<td>10</td>
<td>4.2E18</td>
<td>22</td>
<td>1.1E87</td>
</tr>
<tr>
<td>11</td>
<td>3.2E22</td>
<td>23</td>
<td>7.0E94</td>
</tr>
<tr>
<td>12</td>
<td>5.2E26</td>
<td>24</td>
<td>9.4E102</td>
</tr>
</tbody>
</table>

The number of directed acyclic graphs grows more than exponentially in the number of nodes. Figure 4.1 illustrates the table.
Even for ten variables only (n=10), there are about $4.2 \times 10^{18}$ different graphs. Constructing all structures and discovering the optimal one among all possible graphs require tremendous amount time and space.

The exact learning algorithm such as PC algorithm[122] require exponential time. Approximate inference strategies can be used in order to obtain partially useful information and relatively reliable results with fast speed. Several approaches for approximate inference have been developed and studied [31][115]. While heuristic methods for learning local structures do not require exponential time of computations, they may not be suitable for constructing a global network, because they rely on the assumption that there is no unobserved variable in the causal path within their structure. If there is, the entire output could possibly misrepresent the true relationships of variables in the global network. Furthermore, it may not be practical if no background knowledge has been assumed in general methods [110], when we model diseases and associated factors on a specific dataset. Such problems in modelling process affect both efficiency and accuracy.
4.2. Ordering of Directed Acyclic Graphs

There is a portion of variables’ precedence known by health science researchers. For instance, “smoking” precedes “lung cancer” and “bacteria infections” precedes “fever”. Such kind of information is maintained by their background knowledge or acquired from the temporal relationships in the stage of study design. Information with respect to variables’ precedence is useful because researchers can reason cause and effect with confidence and it is also crucial in practice to discover previously unknown causal relationships by not having to re-examine all known relationships.

A partial order \((\preceq)\) can represent the precedence of variables. The linear form of ordering from the directed acyclic graph is a topological ordering. A partially order set consists of variables whose order satisfies properties of reflexivity, antisymmetry, and transitivity.

Let \(a\), \(b\), and \(c\) are variables in the partially order set,

1. Reflexivity: \(a \preceq a\).
2. Antisymmetry: if \(a \preceq b\) and \(b \not\preceq a\), then \(a = b\).
3. Transitivity: if \(a \preceq b\) and \(b \preceq c\) then \(a \preceq c\).

For a strict partial order \((\prec)\), we remove the property of reflexivity and convert the antisymmetry and transitivity into the following forms:

1. if \(a \prec b\), then \(b \not\prec a\);
2. if \(a \prec b\) and \(b \prec c\) then \(a \prec c\).

The use of partial orders in this research, if not specified, is referred as strict partial orders.

An algorithm with an assumed known ordering of variables such as K2 algorithm[24] can greatly improve the performance of structural learning from the Bayesian network. Evidently, in a partially ordered set we only need to choose variables before a target variable to be the parents of such target. If there is a sequence of \(N\) variables and the target variable \(T\) is in the position of \(K\), there will be \(K-1\) candidate parent variables. The total number of structures that consist both the target variable \(T\) and all its parents will be \(\binom{N}{K-1}\). Figure 4.2 illustrates the description. The number represents the position in the partially ordered set.
Candidate parents for a target variable will be chosen from those whose position number is smaller than the target variable.

\[
\begin{array}{cccccccc}
1 & 2 & 3 & \ldots & K & \ldots & N-1 & N \\
\end{array}
\]

Choosing parents from K-1

**Figure 4.2.** Choosing Candidate Parents of One Variable in the Ordered Set

If \( K = 1 \), there will be no parent for itself. If \( K = N \), there will be \( \binom{N}{N-1} \) parents for \( K \). For all binary variables, finding the optimal parents among all possible structures will take time \( \propto \sum_{k=0}^{N-1} \binom{N-1}{K} 2^K \). Ordering of all variables significantly reduces the number of candidate structures even though the learning problem per se remains computationally challenging.

There is an assumption of variables’ ordering in K2 algorithm (Algorithm 6). An assumed maximum number of parents for each variable can also be set in this heuristic algorithm.

4.3. Learning Causal Relationships with Prior Knowledge

In reality, prior knowledge with ordering of all variables rarely exists. However, clinical scientists, public health professionals, and medical experimentalists may have already discovered or have confirmed with several causal relationships or non-causal relationships such as “smoking causes lung cancer” or “Newborn defects influence many abnormal behaviors in one’s life.”. Existing knowledge either in common sense or in an explicitly temporal relationship is also important in causal analysis such as “lung cancer does not cause one’s smoking” or “A diagnosed ADHD of an eight-year-old child does not cause his newborn defects”. In this research, we provide methods to discover previously unknown causal structures from partially known knowledge.

We may have known a directional relationship of two variables, but this relationship is established through a few intermediate variables. For example, as it has long been known
Algorithm 6 K2 algorithm [24]

1: **Input:** A set of \( n \) nodes, an ordering on the nodes, an upper bound \( u \) on the number
of parents a node may have, and a database \( D \) containing \( m \) cases.

2: **for** \( i \leftarrow 1 \) to \( n \) **do**

3: \( Parent_i \leftarrow \emptyset \)

4: \( P_{old} \leftarrow g(i, Parent_i) \) // a specific function of choosing optimal parents

5: OK-to-Proceed \( \leftarrow \) TRUE

6: **while** OK-to-Proceed AND number of \( Parent_i < u \) **do**

7: let \( z \) be the node in \( Pred(x_i) - Parent_i \) that maximizes \( g(i, Parent_i \cup \{z\}) \)

8: \( P_{new} \leftarrow g(i, Parent_i \cup \{z\}) \)

9: **if** \( P_{new} > P_{old} \) **then**

10: \( P_{old} \leftarrow P_{new} \)

11: \( Parent_i \leftarrow Parent_i \cup \{z\} \)

12: **else** OK-to-Proceed \( \leftarrow \) FALSE

13: Print: Node \( x_i \) and parents of this node \( Parent_i \)

14: **Output:** For each node, the printout of the parents of the node.

that smoking causes lung cancer in general, but more precisely speaking, smoking causes
the accumulation of tar, which is the cause of lung cancer. Our method allows observed
intermediate variables incorporated in the causal path, but we only consider strict partial
orders in this research. In other words, if we have prior knowledge of either temporal or
causal relationship of variables \( V_s, V_t \) and know that \( V_s \) must happen before \( V_t \) or that \( V_s \)
directly or indirectly influences \( V_t \), we write \( V_s \prec V_t \), i.e. \( V_s \) precedes \( V_t \).

Figure 4.3 illustrates the relationships of different research scopes and shows the
computational difficulty of each. The more background knowledge about the ordering of
variables, the easier the learning problem to solve but the less often such information to exist
in reality. Our contributions related to existing research methods are shown in the middle
of figure: learning the network structure by incorporating partially known knowledge.
4.4. Learning Causal Structures with Information of Partial Orders

PC algorithm[122] is a constraint based learning method suitable for sparse graphs, for a large size of the dataset, and for a dataset with no missing data items[89]. Its extended versions includes FCI, SGS, and so on. Those algorithms for exact learning are computationally infeasible when the number of variables are large. When background knowledge is added, we will show the algorithms of graph decompositions reduce the total number of computations.

Derived from the basis of conditional independence test for constraint-based learning and combined with certain assumptions, a local causal learning method such as CCU[23] and CCC[120] takes three variables in each iteration and infer causal structures only for these three variables. In our previous work, non-causal factors can be discovered with a similar local learning approach to CCC [81]. All these algorithms can be completed in polynomial time.

In previous section of the research [81] we have shown a polynomial time algorithm for causal inference when the background knowledge variables either have no causes or are not causal factors for any others. It is a special case of the partial order where for any variables $V_a$, identified as the background knowledge in the ordered set $A(V_n, \prec)$, either $V_a \prec V_n$ or $V_n \prec V_a$, for all $V_n$s excluding $V_a$. The extended version in this research from our previous work allows variables known to be background knowledge to have both preceding and succeeding variables in any partially ordered set.

A heuristic method restricts the number of variables’ conditionings. Since in general
the computational complexity of structural learning stems from the permutations of selecting conditioned variables, any reduction to the number of conditioned variables leads to the alleviation of computational cost. Algorithms provided in this chapter reduce the number of variables that are unnecessarily to be conditioned in the process of the structural learning.

In order to clearly interpret the approach and for the convenience of the narratives, we define the following:

**Definition 4.1.** An *association graph* or a *dependency graph* is the basic structure of an undirected graph without any conditioning process, where all variables are represented by nodes and all edges connecting two nodes represent an association or dependence between these two variables.

Here we use association and dependence interchangeably, and an association graph is equivalent to a dependency graph by definition. If there is no edge directly connecting two nodes, there will be no association between two variables (i.e. they are independent from each other). The determination of association depends on specific mathematical functions.

**Definition 4.2.** A *skeleton* of the graph is an undirected version of the causal network obtained by removing directions from all edges in the graph.

The definition of a skeleton in our research is identical to that defined used for the PC algorithm in a Bayesian network[92] [122]. A skeleton is obtained from the association graph after removals of undirected edges through conditioning on all variables. For instance, if there is a graph structure of three variables (A, B, and C) where they are pairwise associated. According to Definition 4.1, if the association graph is a triangular shape that has three nodes and three edges, and suppose that after all conditioning process for all variables, variable A and variable C are conditionally independent given B, then the skeleton of such graph will be like A–B–C (i.e. the removal of edge A–C).

Obtaining the skeleton of the graph in the PC algorithm is shown in Algorithm 7 [122].
Algorithm 7 PC Algorithm: Obtaining the Skeleton

1: **Input** The association graph of all variables
2: \( n = 0; \)
3: **while** a node has at least \( i + 1 \) neighbors **do**
4: \( \text{for all nodes } A \text{ with at least } i + 1 \text{ neighbors } \text{do} \)
5: \( \text{for all neighbors } B \text{ of } A \text{ do} \)
6: \( \text{for all neighbor sets } X \text{ such that } |X| = i \text{ and } X \subseteq (\text{neighbor}(A) - \{B\}) \text{ do} \)
7: \( \text{if } A \perp B \mid X \text{ then} \)
8: \( \text{Remove the edge } A-B; \)
9: \( i \leftarrow i + 1 \)
10: **Output**: The skeleton of the network.

To obtain the skeleton of the network from an association graph, all combinations of variables will be tested for conditional independence in the original PC algorithm. Figure 4.4 shows where our contributions are in the whole process of this constraint based learning method.

**Definition 4.3.** A *background knowledge variable* is a variable in any one of the existing partially ordered sets.

**Definition 4.4.** A *component* is a subgraph of the association graph or the skeleton of the network where none of the variables in the subgraph have any path connected to other variables outside the subgraph.

**Definition 4.5.** An *triad* is a three-variable structure where all the three variables are pairwise dependent prior any conditioning process.

**Example 4.6.** Figure 4.5 shows an undirected graph where edges represent the dependency of variables, and suppose there is a known order of *background knowledge variables* \( A \) and \( Z \) (by **Definition 4.3**) that satisfies \( A \prec Z \). By **Definition 4.4**, the set containing \( X \) and \( Y \) forms one *component* since there is no other nodes of variables connecting to either of them by
path. The set containing B and C is also a component if background knowledge variable A and its incident edges are removed. By Definition 4.5, the structure consisting of variables A, B and C is a triad.

If the orders of a few variables are known, we show the following processes that learn the structure of the network through graph decomposition. The first step is to construct the association graph (Algorithm 8).

The association graphs can be constructed in $O(n^2)$ where n is the total number of variables in the dataset. The algorithm may output more than one component and each component may contain one or more variables. After finding disjoint sets, we may get multiple association graphs, and variables in each of them are connected by path.
Algorithm 8 Construction of the Association Graphs

1: **Input:** All variables (represented by nodes) in the dataset

2: if two variables are dependent (without conditioning on any third variable) then

3: Add an edge between two nodes.

4: **Output:** One or more association graphs

Finding connected components in a undirected graph, which is a reachability problem, can be solved by any method that finds disjoint sets of components such as breadth-first search, depth-first search, or union-find [26], whose time is bounded to $O(n \log n)$ for a total of $n$ elements in the dataset. Each component (i.e. the disjoint set of the entire graph) is also an association graph. Since there is no association among any inter-component variables, the structure of the network can be learned by processing each component separately. Here we focus on learning the structure from one association graph only. If there are multiple association graphs, the learning process for each one is the same.

Unlike constructing the association graph, obtaining a skeleton of the graph needs conditioning on all variables for independence test, thus the process is time consuming. In the original version of PC algorithm, the complexity is bounded by the largest degree in the graph [122]. However, even though we do not have any controls to minimize the degree of the graph, the number of computations can still be reduced if fewer variables need to be conditioned. At a result, the skeleton of the graph can be constructed faster.

Time complexity of PC algorithm is bounded to the maximum degree, but the algo-
Algorithm neither provides any method of what variables can be excluded in the conditioning process nor combine any background knowledge in the learning process. Next we provide the graph decomposition algorithms that specifically facilitate the learning process of one component with the integration of background knowledge variables.

The following algorithm (Algorithm 9) removes nodes in the association graph that are not in the path of ordered variables.

**Algorithm 9 Graph Decomposition 1**

1. **Input 1**: One association graph of all variables
2. **for** all variables that are background knowledge in the partially ordered set **do**
3. Remove these variables and their incident edges to output all components from the original association graph
4. Find components in the graph;
5. **for** each component **do**
6. Add the nodes of background knowledge variables and their incident edges that were directly connected to the component by the removed edges
7. **Output 1**: New components of the graph //where each component contains at least one background knowledge variable.

In Algorithm Figure 4.5, for example, as variable A is a background knowledge variable, removal of A forms one component containing B and C, and the other component containing the rest of variables (D, E, F, G, H, M, Z). Line 6 of the algorithm adds A to each component, so there will be a new component (A, B, and C), and another one having eight variables including A. Figure 4.6 shows the results of decompositions. We have three components in this graph. The first and the second components (1)(2) are created according to Algorithm 9. the third component (3) is inherited from the construction of the association graph.

Followed by Algorithm 9, we can continue to decompose the graph with Algorithm 10 by removing more nodes through the conditioning process.
Figure 4.6. An Example of the Association Graph after Decomposition

Algorithm 10 Graph Decomposition 2

1: **Input 2:** Output 1
2: **for** each component from the previous output graph **do**
3:   // Learn the causal structure of such component
4:     **for** all background knowledge variables **do**
5:       **if** there exists a triad containing one background knowledge variable (namely B) and two others (namely $V_i$ and $V_j$) **then**
6:         **for** each triad (B, $V_i$, and $V_j$) **do**
7:           **if** ($V_i \perp V_j$ | B) **then**
8:             Remove the edge that connected $V_i$ and $V_j$.
9:           // Continue the process until no more edges can be removed from all triads
10: Find components in the graph;
11: **for** each component **do**
12:    Add the nodes of background knowledge variables that were directly connected to the component by the removed edges
13: **Output 2:** Newly formed components // a pruned association graph

Explanations of above two algorithms are as follows: From input 1 to output 1, the algorithm tests first-order conditional independence (i.e. conditioning on one background knowledge variable only) for all existing background knowledge variables. For one background knowledge variable $v$ in a component containing $n$ variables, total number of the conditional independence tests is bounded to the number of triads containing $v$. The max-
The minimum number of triads containing \( v \) is no more than \( v_m^2 \) where \( v_m \) is the number of variables that are incident to \( v \), and clearly \( v_m < n \). If there are \( m \) background knowledge variables, time complexity for the whole conditioning process will be at most \( O(m \times v_m^2) \). From input 2 to output 2, the algorithm finds the disjoint datasets and it will take \( O(n^2) \) using adjacency matrix for \( n \) nodes in the graph. Figure 4.7 illustrates the algorithm.

**Figure 4.7.** A Graph Exemplifying Decomposition Process

**Example 4.7.** In this association graph, suppose variable A and variable Z are a partially ordered pair such that \( A \prec Z \). Before any conditioning, variables R, S, and T cannot be removed from the entire graph to form a new component since all are connected to variable Z by path, even after the removal of A. In this graph, there are two triads: \( (A, R, \text{and } S) \) and \( (A, S, \text{and } M) \). If we condition on the background knowledge variable A and discover that S and M are conditionally independent given A, we can then remove the edge \( (S-M) \). After the removal of the edge \( (S-M) \), there is no path from any of the variables R, S, and T to the background knowledge variable Z (Any path through A is not considered because A is a background knowledge variable thus it is removed already). Therefore, variables R, S, and T form a new component separated from the component that contains A and Z.

Through graph decompositions, variables belonging to one component need not be conditioned to learn another component’s structure. Consequently, the fewer variables to be included in the conditioning process, the less time to take for the structural learning. However, in each component of the association graph, since edges are still undirected, process of orienting directions of edges to construct the causal structure is still a hard problem due to the nature of the complexity of conditional independence test itself. Compared with the method of learning structures by conditioning on all variables at once, the newly introduced
method decomposes graphs and reduces the size of each graph. Total time for structural learning is reduced if all different components are learned in parallel and it will largely depend on the learning complexity of the giant components in the graph.

To construct the directed acyclic graph from an undirected graph, orienting directions of edges through the discovery of v-structures is required. In addition, conditions of partial orders from background knowledge must hold. We show a method in Algorithm 11 that further reduces the learning complexity by removing even more variables while not violating conditions of existing partial orders.

Definition 4.8. Let $V_a$ and $V_b$ be two background knowledge variables in an existing component of the association graph or the skeleton of the graph and there is a partial order such that $V_a \prec V_b$. A sink structure is a set of variables that none of these variables are in any possible paths from $V_a$ to $V_b$ such that each node in the path is visited only once.

![Figure 4.8. An Example Explaining Sink Structures](image)

Example 4.9. Figure 4.8 illustrates the definition of the sink structure. Suppose A and Z are a partially ordered pair and $A \prec Z$. U itself is a sink structure since there is no path that can both come in and go out of the node, unless node J and edge J-U are visited twice, so it is not possible there is a path from A to Z through U. V, W, and X consist of another sink structure in that any possible paths from A to Z through any of these three variables must visit K twice. Moreover when such a path is oriented with directed edges, the graph will contain a cycle for variables K, V, X, and W, either clockwise or counter-clockwise. M,
J, and K are not in the sink structure.

Any node of the variable with only one edge connected to the other node must be a sink structure, since such edge has to be either incoming or outgoing. To form a path through such a variable, at least two edges should be connected to the node.

Removing sink structures is useful if we want to reduce the number of variables that are unnecessarily conditioned to learn the structure of a particularly ordered pair. However, there is a cost for the removal of sink structures. We show the following algorithm (Algorithm 11) that removes all sink structures for the purpose of learning the structure from one partially ordered pairs.

Let variables A and Z be an ordered pair, and let \( v_1, v_2, ... v_n \) be nodes that are connected with A and Z by path in the component (C) of an undirected graph.

Algorithm 11 Removal of sink structures

1: Init: \( P = \{ v_1, v_2, ... v_n \}, K = \emptyset; \)
2: \( i = 1; \)
3: while \( i \leq n \) do
4:   if \( v_i \in P \) then
5:     Remove \( v_i \) and its incident edges
6:     // Check connectivity
7:     if there is any disjoint set apart from the component C then
8:       for each nodes (e.g. \( v_x \)) in such a disjoint set do
9:         \( P = P - v_x; \)
10:        \( K = K + v_x; \)
11:        Add \( v_i \) and its incident edge back to the component C;
12:      i ++;

The algorithm maintains two sets P and K. P is a set initially having all variables in the component C and after executions of the algorithm P is the minimum set of variables for all possible paths. K is the set that contains variables for all sink structures.
The while loop executes at most n times if no variable is removed as a sink structure. Since the time of finding the disjoint set is bounded to $O(n \log n)$, the algorithm’s worst performance takes $O(n^2 \log n)$.

Given that there is an undirected graph where some nodes are removed after the executions of Algorithm 2 and Algorithm 3, and for each partially ordered pair of variables (e.g., $A_i \prec A_j$), we can further remove all sink structures from a pruned component to create a minimum set of variables for all possible paths. Such a minimum set contains each variable that may be an intermediate variable in a possible path from $A_i$ to $A_j$ but none of the variables outside the set will be an intermediate variable in any possible path from $A_i$ to $A_j$. Since no path can be formed from $A_i$ to $A_j$ through a variable in the sink structure, conditioning on any variables from the sink structure is unnecessary to orient a valid path from $A_i$ to $A_j$. As a result, we further remove unnecessary variables in the conditioning process so as to reduce the complexity of learning the structures for one ordered pair.

At least one path from $A_i$ to $A_j$ should be contained in such a minimum set. However, finding all possible paths from one node to the other is a hard problem in general since it is proven that finding the longest path in an undirected graph is hard. Therefore, the problem of recovering the complete structure in the minimum set of variables remains intractable.

Given an input of partially ordered set of variables, our research methods output several smaller sized local graphs. The rest of learning process for each local graph will be similar to already existing constraint-based methods such as PC algorithm. After the structure of each local graph is learned, we combine them into one global network structure.

Figure 4.9 illustrates all processes of the developed algorithms in this chapter.

Through graph decompositions, the structural learning problem for an entire network is divided into several sub-problems by learning locally connected structures separately. However, there are limitations of learning from a locally connected structure. For example, a structure cannot be learned by existing methods if a component has two variables only. Moreover, the entire output graph may contain erroneous information, such as wrong directions of edges or cycles in the global structure.
Figure 4.9. Processes of Graph Decomposition for Learning Causal Structures

1: Obtaining the association graph from simple associations of variables.
2: Decomposing the graph with background knowledge.
3: Decomposing the graph through conditioning with triads.
4: Pruning the graph by removing sink structures (learning causal relationships for one ordered pair only)
5.1. Measurements of Dependency

In this research, we use dependence and association interchangeably. In computer science, an expression that two variables are dependent is equivalent to the meaning in epidemiology that two variables are associated. Respectively, independence means no association. This section describes how we measure association (dependence) for the purpose of deciding if an undirected edge of two nodes of variables should be added. We use the term “Conditioning” to test conditional independence in computer science. Analogously, we use the term “Stratification” to test if two variables are associated under the stratum of the third variable in epidemiology.

To decide if variables are associated, we check two variables each iteration. For each pair of variables, there are four values in the two-by-two contingency table shown in the following form:

<table>
<thead>
<tr>
<th></th>
<th>Yes or Positive</th>
<th>No or Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes or Positive</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>No or Negative</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c</td>
<td>b+d</td>
<td>N</td>
</tr>
</tbody>
</table>

In epidemiology, one row stands for one behavior or an unexposed condition with true or false values, and the other row stands for the opposite behavior or an exposed condition with true or false values. One column stands for the presence of one disease and the other column stands for the absence of such disease (the healthy condition). The row heading is the exposure or risk factor and the column heading is the outcome or disease. In computer science, we do not know which variable is the exposure and which is the outcome. This table just tells how these values stand for. \(N\) is the sum of variables \(a, b, c, \) and \(d\): \(N=a+b+c+d\).
For the test of independence, a chi-squared ($\chi^2$) test is commonly adopted in statistics and Bayesian analysis. A chi-squared ($\chi^2$) test can be performed to evaluate the difference between the observed value and the expected value under null hypothesis ($H_0$, all variables are independent).

In epidemiology, odds ratio is used to measure the strength of associated variables in cross-sectional study and case-control study. Moreover, when each variable has only binary values (i.e. true as “presence” and false as “absence” only), odds ratio is a relatively simple and fast option. Odds ratio may be preferred in epidemiological studies [28] to test the homogeneity for two variables. A chi-squared test, particularly in measuring independence in Bayesian network, has the ability to be extended to more than two categories in one statistical test. As different measurements have their own advantages, we use both odds ratio and the chi-squared test for variables with two categories and chi-squared test only for variables with more than two categories.

The chi-squared ($\chi^2$) test is calculated as:

$$\chi^2 = \frac{1}{d} \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

where $d$ is the degree of freedom, $O$ is the Observed value, and $E$ is the Expected value.

An alpha value ($\alpha$) is set to determine the threshold either accepting or rejecting the null hypothesis (i.e. two variables are independent). If this value is set to be 0.05, we will choose to reject the hypothesis if there is only 5 percent or less probability that we make the wrong conclusion.

If our experiments only compare variables in a two-by-two table, the measurements with odds ratio will be convenient and adopted. The adoption of odds ratio to measure associations of variables in our research has the following reasons.

1) Measurement in odds ratio makes sense in epidemiology and this is commonly adopted in epidemiological research, including in case-control study, cross-sectional study, and so on.

2) Odds ratio can show the strength of association while chi-squared test is to test for independence. A simple chi-squared test used in statistics only provides the information
if two variables are independent. If we double the frequencies of samples, chi-squared values will be doubled too, but the strength of the association remains unchanged.

3) Chi-squared test is based on the assumption that all conditions measured should be independent, that the data should be frequent because the sample size will impact the accuracy of measurement, and that the data are from random sampling.

Odds ratio (OR) is calculated as:

\[
\text{OR} = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}
\]

Though there are other measurements such as relative risks for different purpose of study, this research focuses on a rough determination of association (dependence) of variables only. Epidemiological measurements vary and the adoption of a particular measurement is often related to the model design [50]. We do not argue which is superior but we odds ratio and the chi-squared test for all testing datasets for the convenience of computations in general.

One important issue is to determine threshold of independence or dependence, or in other words, strong or weak relationship. No fixed criterion exists to determine when the variables are independent or dependent. We may assume, based on literature in medical fields, that an odds ratio close to 1.0 suggests independence (no association). The strength of association represented by the odds ratio between 1.1 and 1.5 is weak, between 1.6 to 3.0 is moderate, and greater than 3.0 is strong [28] [84]. Even though there is no strict quantitative standard to forcibly determine the categories, in epidemiological research, values of odds ratio could be comparable with values of the same types of studies in other research publications.

Here we point out that in epidemiology [114] more sophisticated tests such as a Cochran-Mantel-Haenszel test may be used to test for independence repeatedly and to avoid bias such as Simpson’s Paradox [121]. Due to our research focus, the determination of association (dependence) of two variables is just an estimate from calculated values of odds ratio and the chi-squared test.
5.2. Measurements of Accuracy

In this section, we describe how we evaluate the accuracy of our learned network. The term “accuracy” here only refers to the correctness of network structures and it does not indicate if a relationship is truly causal in reality. Evaluations of accuracy in the experimental results are measured using the following two-by-two table in general [67]:

<table>
<thead>
<tr>
<th>Number of Edges</th>
<th>Condition Positive</th>
<th>Condition Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Positive</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Test Negative</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

The explanations of this table are as follows.

1. For the label of “Condition” (shown in the first row), it is the gold standard from a benchmark value or the correct sample in the supervised learning that represents the reality. “Condition” represents the true classification.

2. For the label of “Test” (shown in the first column), it is the obtained value from experiments or observations. “Test” represents the imperfect classification.

3. Combined the “Condition” and the “Test”:

3.1. “True positive” (TP) is the number that the test outcome is correctly identified according to the gold standard.

3.2 “False Positive” (FP) is the number that the test outcome is incorrectly identified (known as Type I error in statistics). If the conditions are observed to be present, but in reality there should not be such kinds of conditions, then it is a Type I error.

3.3 “False Negative” (FN) is the number that the test outcome is incorrectly rejected (known as Type II error in statistics). If the conditions are not observed, but in reality there should be such kinds of conditions, then it is a Type II error.

3.4 “True Negative” (TN) is the number that the test outcome is correctly rejected (i.e. there is no identified condition both in experiments and in reality).
The accuracy is improved if more samples fall into the category of either “True Positive” (TP) or “True Negative” (TN).

In the following mathematical equations, we use abbreviations of TP as the number of “True Positive”, FP as the number of “False Positive”, FN as the number of “False Negative”, and TN as the number of “True Negative”.

Since both “False Positive” and “False Negative” are erroneous results, we need to assess different conditions and obtain a balance between wrongly included samples and missed results. The evaluations from this two-by-two table shall include: specificity, sensitivity or recall, precision, and false omission rate

Specificity is calculated as:

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

Sensitivity or recall is calculated as:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

Precision is calculated as:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

False omission rate is calculated as:

\[
\text{False Omission Rate} = \frac{FN}{TN + FN}
\]

The relations between our research and all abovementioned evaluations are shown below:

a. If there is a directed edge in our experimental network that is consistent with both the position and the direction of the edge in a gold standard network, then this edge is correctly identified as a “True Positive” (TP).

b. If a directed edge in our experiment exists but this edge does not appear in the same position in a gold standard network, then this edge is a “False Positive” (FP).

c. If there should be an edge between two nodes in the network but such an edge is not detected in the experiment, then this supposed edge is a “False Negative” (FN).
d. If there is not an edge between two nodes both in the experiment and in the gold standard network, then this non-existing edge is identified as “True Negative” (TN).

Theoretically, if there are N nodes in one network, the total number of both existing edges and non-existing edges will be N(N-1)/2. However, in a relatively sparse graph, edges do not exist between every pair of two nodes. The number of existing edges will be much smaller than N(N-1)/2. In our experiments, the resultant graph is also sparse. This situation results in a large number of “True Negative” cases. Evaluation formulas that contain “True Negative” cases such as specificity and false omission rate will not be practically useful because the measurement results will be much affected by a large number of “True Negatives”. As a result, the evaluation in our research experiments will be focused on precisions and recalls in particular.

For an overall score of accuracy, we use F-measure (F-score) [108] by adding both precision and recall into one formula. F-measure (F-score) is calculated as:

$$F\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}$$

There is an advantage of using one score to measure the accuracy of discovered network. As previously stated, only the edge with both a correct direction and a correct position in the network is considered as “True Positive”. Depending on the treatment of them, incorrect edges may be considered as non-existing, undirected, wrongly directional, or bi-directional. All of the incorrect edges may either be “False Positive” or “False Negative”. For example, if there is an association between two variables but it is not detectable on the causal direction of these two, we could decide in either way:

(1) to remove such an undirected edge because the causal relationship is unclear from the learning process, or

(2) to assign a random direction of the edge to represent a causal relationship.

No matter which decision we make to classify erroneous edges as either “False Positive” or “False Negative”, using the F-measure is appropriate because in the calculation of the F-score, the classification of “False Positive” or “False Negative” will not change the F-score.
5.3. Experiments with Local Causal Discoveries

To illustrate the heuristic methods from local causal discoveries shown in Chapter 3, we have applied our methods in a synthetic dataset, LUCAS (LUng CAncer Simple set) (http://www.causality.inf.ethz.ch/data/LUCAS.html), which was originally made up as a synthetic dataset for illustration. The dataset contains twelve variables. We have shown the algorithms of local causal discoveries in Chapter 3 to test conditional independence of variables, and the most plausible structure of this disease-related network has been learned. The dataset contains twelve variables of “Smoking,” “Yellow fingers,” “Anxiety” “Peer pressure,” “Genetics,” “Attention disorder,” “Born (on) an even day,” “Car accident,” “Fatigue,” “Allergy,” “Coughing,” and “Lung cancer”. Originally the dataset is used to predict lung cancer probability from other variables, while we use it to learn the structure of a causal network. The first experiment is used for demonstration purpose only so as to show how all of our research methods of local causal discoveries are applied.

To begin with learning, the first step is to choosing proper background knowledge variables. The background knowledge we assume is that “Genetics” has no known causes (acausal factor) and “Car accident” is a consequence which does not cause any other conditions. In this experiment, we do not argue whether such assumptions are factually correct, but we take them only as one integral step of the learning process. Hence, the structure of the layered network will be as follows:

(1) On the top of the layer, there is an acausal variable, i.e. “Genetics”.

(2) At the bottom of the layer, there is a consequence which does not cause other conditions in the dataset, i.e. “Car accident”.

(3) In the middle layer, there are other variables which are going to be determined how they are related to each other and to the variables on both top and bottom layers.

Figure 5.1 below shows the structure of the network with assumed background knowledge.

Next, we construct the undirected graph based on the strength of association. The
Figure 5.1. A Three Layered Graph Structure for Structural Learning

undirected graph is an association graph where each edge between two nodes represents an association between two variables.

We test for the strength of associations among variables based on odds ratio. We arbitrarily set the value of 2.0 or 0.50 (reciprocal of 2.0, as a negative correlation) as the threshold to determine if two variables are associated. If the odds ratio of two variables is above 2.0 or below 0.50, these two variables are associated, so we can add an undirected edge between them. No edges will be assigned between two nodes of variables if the odds ratio is between 0.50 to 2.0.

In this experiment, due to the demonstration purpose and the simplicity of this network structure, we do not include the procedure of conditional independence test for all variables’ combinations to obtain the skeleton of the network as described in PC algorithm. We assume the undirected graph is the skeleton of the network.

Followed by the construction of association graphs, we decomposing the graphs and test conditional independence of the remaining variables repeatedly. The graph is decomposed based on either one of the criteria:

(i) if nodes in set A have no edge connected to other node in set B, or
(ii) if there are connected edges, only either between top layer variables and them, or between bottom layer variables and them.
For criterion (i), we are able to exclude a variable “Born (on) an even day” since there is no other variable in the dataset associated with it. For criterion (ii), we discover that “Attention disorder” is associated with “Genetics” and “Car accident” only. Since “Genetics” and “Car accident” are from background knowledge layers (top and bottom layers), breaking the associations between these two variables and variable “Attention disorder” can make “Attention disorder” disconnected with any other variables in the dataset. Both “Born (on) an even day” and “Attention disorder” form new components. Hence, part of the network structure has been learned, as we exclude “Born (on) an even day” (a non-relevant factor), and have built a local structure as “Genetics” causes “Attention disorder”, which further causes “Car accident”. This learning process and results are shown in Figure 5.2.

![Figure 5.2](image)

**Figure 5.2.** An Illustration of Subgraph Decomposition.

After the decompositions of the graph, we repeatedly test for conditional independence using rules of local causal discovery. According to the calculations, variables “Coughing,” “Fatigue,” and “Lung cancer” are all associated with top-layer variable “Genetics” and bottom-layer variable “Car accident”, and these three variables are also pairwise related. We can apply CCC rule and CCCx rule to test conditional independence, because there is
either a variable which has no known cause or a variable which has no known consequence. After conditional independence testing, we get “Genetics” → “Lung cancer” → “Coughing”, and “Genetics” → “Lung cancer” → “Fatigue” → “Car accident”. This conditioning process and results are shown in Figure 5.3 and Figure 5.4.

![Undirected Graph](image)

**Figure 5.3.** Illustrations of Local Causal Discoveries (the Undirected Graph).

Based on the skeleton we have previously constructed, “Allergy” is associated with “Coughing”, and the rest of four variables are all associated with “Lung cancer”. The structural learning process can then be divided into two parts, in which each part can be learned through conditional independence test independently and in parallel.

Part One: For the variable “Allergy”, since it is only associated with variable “Coughing”, and there is a learned relationship that “Lung cancer” causes “Coughing” from previous process. We have discovered that “Allergy” and “Lung cancer” are conditionally dependent given “Coughing”. CCU rule applies, so we conclude that “Allergy” causes “Coughing”.

Part Two: For four remaining variables “Smoking,” “Yellow fingers,” “Anxiety,” and “Peer pressure”, since neither is dependent with “Genetics” but all are associated with “Lung cancer”, we apply CCU rule repeatedly for all four variables and have discovered
that “Smoking” and “Genetics” are conditionally dependent given “Lung cancer”. we then conclude that “Smoking” causes “Lung cancer”.

Figure 5.5 shows two components of the graph and Figure 5.6 shows the results of causal discoveries from both components.

The subgraph consisting of “Smoking,” “Yellow fingers,” “Anxiety,” and “Peer pressure” is shown above. By applying CCU rule, we get “Anxiety” $\rightarrow$ “Smoking” and “Peer pressure” $\rightarrow$ “Smoking”, and by applying CCC rule, we get “Anxiety” $\rightarrow$ “Smoking” $\rightarrow$ “Yellow fingers”.

Till now, this disease causal network has been constructed based on all the heuristic methods of local learning, except only one edge direction remaining unclear. The complete graph is shown in Figure 5.7.

With the correct information of background knowledge we have pre-assumed in the graph, methods of our local causal discoveries applied to this specific dataset has 91.7% true positive rate, except one unclear edge remaining.
5.4. Experiments with Background Knowledge Information

For the experiments, we generate synthetic datasets with 10000 cases for both Asia network[76] and ALARM network[9]. Asia network has 8 nodes, 8 edges, 18 parameters, an average degree of 2, and the maximum in-degree of 2. ALARM network has 37 nodes,
46 edges, 509 parameters, an average degree of 2.5, and the maximum in-degree of 4. Both networks are relatively sparse graphs so that a constraint based learning is suitable.

In Asia network, there are 8 variables with only binary values:

\{ Smoking, LungCancer, VisitToAsia, Tuberculosis, Tuberculosis-or-Cancer, X-ray, Bronchitis, Dyspnea \}.

In ALARM network, there are 37 variables with categorical data:

\{ Hypovolemia, LV(Left Ventricular) Failure, LVED(Left Ventricular End-Diastolic) Volume, Stroke Volume, CVP(Central Venous Pressure), PCWP(Pulmonary Capillary Wedge Pressure), Insufficient Anest/Analgesia, Pulmonary Embolus, Intubation, Shunt, Kinked Tube, Min VolSet, VentMach, Disconnection, Vent Tube, Vent Lung, Vent Alv, FiO2, PA Sat(Pulmonary Artery Oxygen Saturation), SaO2, Anaphylaxis, TPR(Total Peripheral Resistance), Art-CO2, Catecholamine, CO(Cardiac Output), History, BP(Blood Pressure), Error Cauter, HR(Heart Rate), HR EKG, HR SAT, Error Low Output, HR BP, Exp-CO2, PAP(Pulmonary Artery Pressure), Pressure, MinVol(Tidal Volume) \}.

There is a preset structure of each network from the literature. In our experimental
design, we assume that we do not know the complete structure of all variables but know the ordering from some variables. For example, in Asia network, we have variables “Visit to Asia,” “Tuberculosis,” “Smoking,” “Lung cancer,” “Bronchitis,” “Tuberculosis or lung cancer,” “Xray result,” and “Dyspnea” and we may assume we already know from knowledge such as “Smoking” causes “Lung cancer” (“Smoking”→“Lung cancer”, or “Smoking”≺“Lung cancer”) or ”Visit to Asia” precedes “Xray result” (“Visit to Asia”≺“Xray result”), but we have no additional information on the relationships of other variables.

The input of the experiments contains a complete record of all variables and a few predetermined sets of partial orders. All predetermined partial orders are faithful to the real network structures. In other words, if a gold standard structure shows that A precedes B, then in our predetermined partially ordered set, A has to precede B, too. We repeatedly change our “knowledge” set of partial orders and learn the structure from the partially ordered set. The output of the experiments is either the complete network structure or a specific component related to the partially ordered set only. We analyze the efficiency and accuracy by comparing the experimental results to the complete gold standard structure.

In our experiment, we set alpha level of significance as 0.05 for the chi-squared test, and odds ratio 2.0 and 0.5 (either above 2.0 or below 0.5 which is the reciprocal of 2.0) as the threshold to determine dependency.

Regarding the learning through graph decomposition, since this paper does not address issues of parallel methods at this moment, we record the maximum time taken for learning each component for the convenience of comparing the performance between learning from each one component separately and learning from the entire graph at once. The program is written with R.

5.5. Analysis of Experiments

In our own experiment, the learned network from the synthetic dataset using the chi-squared ($\chi^2$) test is much different from the gold standard Asia network. There may be several reasons such as the difference of probability distribution between the dataset we use and the one in gold standard, the threshold for determining associations, or just
the limitations of chi-squared test. Nevertheless, we instead use odds ratio to measure independence in Asia network. Figure 5.8 shows the gold standard Asia network.

Figure 5.8. A Standard Asia network

The first step in our algorithms is to obtain an association graph from data. For Asia network, it is simple and only the node of “Tuberculosis(TB) or Cancer” is detected to be in the wrong place (Figure 5.9). The reason why this node is not consistent with the correct network structure may be due to our chosen type of statistical testing or the ambiguity of this variable itself. We get 87.5% true positive rate of discoveries if we ignore this variable and continue the learning process. The F-score in this experiment is 0.471.

By pre-assigning the ordering of some variables in the graph such as “Tuberculosis(TB)” $\prec$ “X-ray”, we execute the program for structural learning. We record the time of learning without any orderings (number of background knowledge variables is 0) and the time of learning with partially ordered information(Table 5.3).

Figure 5.10 shows a “standard graph” of ALARM network generated by TETRAD 5.1 (http://www.phil.cmu.edu/tetrad), using chi-squared ($\chi^2$) test with alpha value (\(\alpha\)) set to 0.05. This network is used to compare the graph generated from multiple subgraphs in our experiments. Note that the standard graph contains errors too. Some edges are not
directed or bi-directional. Those errors are either due to the deviations of the true probability distributions from the synthetic dataset, or systematic errors in the original sample itself (in other words, even if the probability distributions in our samples follow strictly to the true distributions in the original literature, some erroneous edges will still appear because of the limitation of constraint based learning itself).
Figure 5.10. The ALARM network

The association graph of the ALARM network contains 315 undirected edges. Performance of efficiency varies from the selection of background knowledge variables.

We at first randomly choose all variables to be the background knowledge where the precedence is not in a chain order, such that $V_a \prec V_b, V_c \prec V_d$, but the ordering information for $V_a$ and $V_c$ is unknown. In this type of selection, we do not observe any significant reduction of learning time mainly because the strongly connected graph cannot be decomposed into several smaller components. Then we choose an ordering that is a chain order such as $V_a \prec V_b \prec V_c \prec V_d$, or in a way that multiple variables precede a common variable such as $V_a \prec V_d, V_b \prec V_d$, and $V_c \prec V_d$. For example, if we select variables with known ordering “TPR,” “Catechol,” “HR,” and “HRBP” as background knowledge, the largest component after decomposition contains 10 variables only.
Process of learning causal structures is efficient when the original association graph can be effectively decomposed. The structure of small components needs fewer conditional independence tests and less learning time.

We only observe very limited effect as there are not as many variables in the triad to be removed upon conditioning on background knowledge variables. The phenomena are probably due to the specific structure of the ALARM network. Since the number of variables that are removed from the triad structure is limited, all remaining variables are still necessary to placed in the conditional independence test in the learning process.

In our experiment, we assume the total learning time depends on the component with largest size when the structure of each component can be learned in parallel. Table 5.4 shows the performance of learning the causal structure from the largest component in the ALARM network.

**Table 5.4. Performance of Learning ALARM Network**

<table>
<thead>
<tr>
<th>Number of Background Knowledge Variables</th>
<th>Learning Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Best (worst)</td>
</tr>
<tr>
<td>0</td>
<td>398.024 (420.057)</td>
</tr>
<tr>
<td>2</td>
<td>268.019 (464.602)</td>
</tr>
<tr>
<td>4</td>
<td>107.587 (491.123)</td>
</tr>
<tr>
<td>6</td>
<td>81.165 (516.607)</td>
</tr>
<tr>
<td>8</td>
<td>55.667 (512.103)</td>
</tr>
<tr>
<td>10</td>
<td>40.340 (449.405)</td>
</tr>
</tbody>
</table>

We do not include results from the algorithm of removing sink structures for the comparison of performance efficiency because this algorithm is used to learn a local causal structure for a specific ordered pair of variables. We will discuss this issue in more detail in next chapter.

Depending on the selection of background knowledge and the input of a particular ordering of variables, a true positive rate of discoveries is between 64.9% and 78.4%. Er-
roneous information includes a non-detected edge as false negatives, a redundant edge as false positives, or a wrongly directed edge including a bi-directional edge in the network. F-scores in this experiment are between 0.632 and 0.773, depending on the specific input of background knowledge.

5.6. Experiments on Real Epidemiological Datasets

In this section, we an experimental result when our methods are applied into real epidemiological datasets. The purpose of the experiment is to show the applications in practice as well as the limitations of our computational methods. All calculations are just the estimates for deciding variables’ associations.

In this experiment, we show an example using 2010 Behavioral Risk Factor Surveillance System (BRFSS) data to test potential causes and effects for certain diseases. According to the descriptions of the Centers for Disease control and Prevention (CDC), “The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.” (http://www.cdc.gov/brfss/)

In the experiment we consider three different variables in each testing iteration. Among the chosen three variables, one variable is assumed to be an acausal factor, which cannot be caused by any of the other two. By giving this assumption, we apply the method of local causal discoveries.

First, we choose pairwise correlated variables in a dataset. We test the relationships among “Obesity,” “Diabetes,” and “Alcoholic beverages consumption” in the dataset. Since this research does not focus on any improvements of data quality problems such as any non-response values, any potential statistical deviations, and any weighting issues, we use the data directly only for an estimation of variables’ relationships. We assume that “Obesity” is the acausal factor among these three variables so that “Obesity” can only be the potential cause of the other two variables. The goal is to test which role “Alcoholic beverages consumption” plays in the given set. “Alcoholic beverages consumption” can be a risk factor
that contributes to the development of diabetes, or it can just be an associated factor in the group of people who have diabetes because of their living behaviors.

Since we only take boolean data and ignore polytomous data for simplicity, we choose a set of two groups for comparisons respectively: 1) the group of “Obese” (BMI $\geq 30$) and the group of ”Neither overweight nor obese”; 2) the group of “Drink any alcoholic beverages in past 30 days” and the group of ”No alcoholic beverages consumption”; 3) the group of “Currently having diabetes” and the group of “no diabetes”.

In a preliminary experiment, the raw odds ratio of the group “Obese” and “Diabetes” is approximately 3.4, indicating that there is a strong relationship of obesity and diabetes. The experiment also shows that the odds ratio of group “Diabetes” and “Alcoholic beverages consumption” is 0.37. The result indicates that people who currently have diabetes have less alcohol consumption, compared with group of the people without diabetes.

We then use the measuring form of “No alcoholic beverages consumption”, since “Alcoholic beverages consumption” does not suggest either obesity or diabetes according to raw calculations. The relationship is shown in Figure 5.11. Three variables are pairwise correlated. In this figure, we use “Ob” as “Obesity”, “Da” as “Diabetes”, and “¬Ac” as “No alcoholic beverages consumption”.

![Figure 5.11. Three Pairwise Correlated Variables](image)

(1) Stratify-conditioning on the variable of “Diabetes”.

The experiment shows that the odds ratio of “Obesity” and “No alcoholic beverages consumption” is 1.3, indicating some degree of dependence though it is not considered as a strong association. The odds ratio of “No alcoholic beverages consumption” and “Diabetes” is as high as 2.7. Since “Obesity” and “Diabetes” are correlated as stated previously, the
CCC rule is applied by conditioning on variable “Diabetes” or on “No alcoholic beverages consumption”.

By stratify/conditioning on the groups of “Diabetes”, the result shows that the odds ratio of “Obesity” and “No alcoholic beverages consumption” in the group of “Diabetes” is 0.96, which is very close to 1.0. An Odds ratio close to 1.0 suggests that two variables are independent. In this situation, conditioning on the third variable “Diabetes” makes two previously dependent variables “Obese” and “No alcoholic beverages consumption” independent.

For “No diabetes” cases, odds ratio of “Obesity” and “No alcoholic beverages consumption” is 0.86, as such a statistical testing may indicate people with obesity may consume alcoholic beverages more frequently.

In this experiment, it is very clear only in the groups of “Diabetes”, there is an independence (no association) between “Diabetes” and “No alcoholic beverages consumption”. In conclusion, “Obesity” causes “Diabetes”, and “Diabetes” causes “No alcoholic beverages consumption”.

This learning process is shown in Figure 5.12, where the dashed edge between “Obesity” and “No alcoholic beverages consumption” means previously dependent variables become independent. In other words, there is no association between “Obesity” and “No alcoholic beverages consumption” under the stratum of “Diabetes”.

![Figure 5.12. Conditioning on “Diabetes”](image)

(2) Stratify/Conditioning on the variable of “No alcoholic beverages consumption”.

By stratifying/conditioning on the variable “No alcoholic beverages consumption” to test if the association between “Obesity” and “Diabetes” exist, we still find these two
variables are associated. In the group of “alcoholic beverages consumption”, the odds ratio of “Obesity” and “Diabetes” is 2.0. In the group of “No alcoholic beverages consumption”, the odds ratio of “Obesity” and “Diabetes” is 2.4. Neither of the odds ratios makes these two variables “Obesity” and “Diabetes” independent (unassociated) from each other.

From the test result we conclude that it is nearly impossible for “No alcoholic beverages consumption” to be a cause of “Diabetes”. This situation is shown in Figure 5.13. After this conditioning process (stratification analysis), three variables are still pairwise correlated.

From the test result we conclude that it is nearly impossible for “No alcoholic beverages consumption” to be a cause of “Diabetes”. This situation is shown in Figure 5.13. After this conditioning process (stratification analysis), three variables are still pairwise correlated.

From this experiment, a possible causal structure of three variables has been learned given that there is a known acausal factor as cause. In summary, “Diabetes” is caused by “Obesity”. “No alcoholic beverages consumption” is statistically independent from “Obesity” under the condition of “Diabetes”. “No alcoholic beverages consumption” is a reasonable explanation for “Diabetes”.

The causal structure discovered from this experiment reflects a property of context-specific independence [14]. Context-specific independence shows that variables are conditionally independent given a specific context of the third variable. In our experimental results, conditions of “Obesity” and “No alcoholic beverages consumption” are conditionally independent given the context of “Diabetes” being TRUE only:

“Obesity” $\rightarrow$ “Diabetes” (TRUE) $\rightarrow$ “No alcoholic beverages consumption”.

However, when the context of “Diabetes” is FALSE, there is no causal relationship between “No diabetes” and “Alcoholic beverages consumption”. In other words, there is no indication if people who do not have diabetes consume more alcohol or less alcohol in this observational result.
Related issues and computational methods of the discovering relationships from context-specific independence will be left as our future work.
In this chapter, we discuss issues related to the fields of epidemiology, computer science, medical experiments, and health science research:

(1) the reliability of causal discoveries,
(2) Big data problems in this research,
(3) performance improvements in our algorithms, and
(4) the implications on health science research.

Causal discoveries have a long history but methods are evolving. We discuss potential problems and limitations along with research assumptions. The big data problem becomes a focusing topic in computer science and we discuss the problems and their possible solutions. The applications in graph theory research have been extended to causal analysis, and we show our methods are empirical of efficiently solving the problems in health science research.

6.1. Reliability of Possible Causal Discoveries

Regarding reliability of possible causal discoveries, this research hinges on the criteria of database completeness and data integrity, Bayesian causal network model, no selection bias, causal faithfulness, and valid statistical testing. No matter which method we choose and how precise the calculation can be made, deviations and erroneous results are still inevitable. Reasons of implausible results are several [53]. Variables’ relationships could be any conditions of bias, chance, confounding, or causation.

(1) Bias. There are many types of biases such as information bias as misclassification, selection bias, inference bias, and reporting bias. A collider-stratification may also be prevalent in causal models [49]. We may determine that a strong correlation exists between two variables if the observed odds ratio of these two variables has a large value. However, the observed odds ratio may not represent accurate odds ratio because of the bias or measurement deviation. When sample size is small, there is a high possibility that the bias occurs.
Measurement errors are suggested to be addressed at the the stage of planning a study and collecting the data [67].

(2) Chance. Correlation between two factors just happens by chance. A frequent gambler can always summarize some patterns of stakes but those patterns will change over time and they provide no help to win. Analogously, it is common that the correlation between two factors disappears when the investigators change to another dataset with the same experiment but discover no correlation or a completely different outcome. Moreover, different settings of confidence intervals, alpha values in chi-squared test and other statistical measurements may make the results unreliable to support previous hypothesis and claims.

(3) Confounding. Confounding is an extraneous pathway we want to exclude as it owes to another effect. For example, the fact that smokers carry matches does not lead to any conclusions of causal relationship between carrying matches and lung cancer, since carrying matches is a confounder in the study of smoking and lung cancer. There is no statistical test to distinguish a confounding factor and a real cause, since confounding factors are integral to causal links [102]. Experiments outside statistical tests are required to differentiate confounding and the cause. We have shown an real example that justifies all confounders are not detectable only through statistical tests in our experiments.

(4) Causation. While a causation may have been identified, researchers still have to be cautious because a real cause should be thoroughly examined through medical experiments or through long-term observations such as randomized controlled trials. The controlled experiments ensure such discoveries of causation are based on temporality, proved mechanisms of experiments and both internal and external validations, but not due to statistical correctness.

The accuracy from learning results will not be achieved without database completeness or data integrity. In order to archive reliable conclusions, additional work may be made with missing data problems [29]. In a constraint based structural learning method especially, complete record of data is necessary to ensure validity. Hence, we incorporate one step in the general framework (Figure 3.1) to complete data pre-processing for all mentioned data
quality issues.

Without the assumption of Bayesian causal network model, our probabilistic models and inferred causal relationships cannot be established. For instance, in a cyclic causal situation or bi-directional case, such that A influences B, and B will have feedback to A, causal inference is impossible because a Bayesian network itself does not entail cycles.

The assumption of no selection bias ensures the representation of conditional dependence is only owing to the discovery of v-structures (colliders) within collected samples. Without this assumption, a selection bias may exist in a dataset that misrepresents the probability distributions of the entire population. For example, a falsely concluded association between bone fractures and fever can be due to a biased selection from samples of hospital visitors only.

In our assumption of causal faithfulness, if two variables are causally related, they must be associated/dependent with each other. The necessary condition of causation is correlation. Without this assumption, we cannot discover possible causal structures and analyze them from the correlations of variables.

The correctness of identification in causal relationships depends on the valid statistical testing. A valid statistical testing is often supported by a large number of cases in the dataset. Although there is no standard to determine the threshold of the number of samples to ascertain the identified relationships are truly causal since causality must be researched in experimental study, a valid statistical testing still remains valuable in the stage of determining associations and conditional independence in the structural learning process.

6.2. Computational Complexity on Big Data

The concept of big data has not emerged until twenty-first century, whereas computational time complexity has long been an issue on an NP-hard problem. When actual inputs are small, algorithms with exponential running time may still be satisfactory [26]. Whereas most epidemiologists, to their expertise, may focus on a limited number of pre-selected variables which they believe to be potentially important or causally related, computer scientists address the big data problem to encompass almost all available variables and solve it by giv-
ing approximate or near-optimal solutions. Causal discoveries on a big dataset has been an issue and it is still challenging with the increasing number of surveyed variables. Throughout the research, probability theory is applied for processing large-scale data and heuristics play an essential role for modeling diseases and analyzing multivariate causal relationships.

Since the term of big data is new, there may be misconceptions on big data, such that untrained people only believe that big data is only pertaining to the datasets themselves that cannot be stored in one single database. The determination of big data however in this research is related to both the number of computations and the space used for search and storage, not just to the storage of such raw datasets.

Suppose there is a table of data that contains n elements and m variables shown below (Table 6.1):

**Table 6.1. An Example of Row-Oriented Table of Data**

<table>
<thead>
<tr>
<th>ID</th>
<th>Variable-1</th>
<th>Variable-2</th>
<th>Variable-3</th>
<th>Variable-4</th>
<th>...</th>
<th>Variable-m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Record-1.1</td>
<td>Record-1.2</td>
<td>Record-1.3</td>
<td>Record-1.4</td>
<td>...</td>
<td>Record-1.m</td>
</tr>
<tr>
<td>2</td>
<td>Record-2.1</td>
<td>Record-2.2</td>
<td>Record-2.3</td>
<td>Record-2.4</td>
<td>...</td>
<td>Record-2.m</td>
</tr>
<tr>
<td>3</td>
<td>Record-3.1</td>
<td>Record-3.2</td>
<td>Record-3.3</td>
<td>Record-3.4</td>
<td>...</td>
<td>Record-3.m</td>
</tr>
<tr>
<td>4</td>
<td>Record-4.1</td>
<td>Record-4.2</td>
<td>Record-4.3</td>
<td>Record-4.4</td>
<td>...</td>
<td>Record-4.m</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>Record-n.1</td>
<td>Record-n.2</td>
<td>Record-n.3</td>
<td>Record-n.4</td>
<td>...</td>
<td>Record-n.m</td>
</tr>
</tbody>
</table>

There are two different big data problems: (1) when the number of n (rows) is large, and (2) the number of m (columns) is large. In a disease related dataset, each row in the table represents an individual and each column represents a variable. This research emphasizes on the problems and solutions of learning causal structures among observed variables in the dataset. Heuristic algorithms are given to solve the second problem. Algorithms for discovering causal and non-causal factors take three variables (m=3) in each iteration and their executions are in polynomial time. Algorithms with graph decompositions reduce m to a smaller number in the variables’ conditioning process.
Another important issue is the storage of a dataset. Most raw datasets in public health domain such as those from Behavioral Risk Factor Surveillance System (BRFSS) and our experimental datasets are originally designed and recorded in the style shown in Table 6.1. Data are stored row by row. Row-oriented storage has several advantages in that data are easily recorded, viewed, and edited. Especially for computations such agent-based modeling, data can be extracted row by row through reading each record sequentially from the database.

Row-oriented data storage may not be suitable for structural learning in this research. Column-oriented database management systems [1] have advantages in reading variables rather than records of each individual. Table 6.2 shows how the same data of n elements and m variables are stored column by column.

Table 6.2. An Example of Column-Oriented Table of Data

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable-1</td>
<td>Record-1.1</td>
<td>Record-2.1</td>
<td>Record-3.1</td>
<td>Record-4.1</td>
<td>...</td>
<td>Record-n.1</td>
</tr>
<tr>
<td>Variable-2</td>
<td>Record-1.2</td>
<td>Record-2.2</td>
<td>Record-3.2</td>
<td>Record-4.2</td>
<td>...</td>
<td>Record-n.2</td>
</tr>
<tr>
<td>Variable-3</td>
<td>Record-1.3</td>
<td>Record-2.3</td>
<td>Record-3.3</td>
<td>Record-4.3</td>
<td>...</td>
<td>Record-n.3</td>
</tr>
<tr>
<td>Variable-4</td>
<td>Record-1.4</td>
<td>Record-2.4</td>
<td>Record-3.4</td>
<td>Record-4.4</td>
<td>...</td>
<td>Record-n.4</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Variable-5</td>
<td>Record-1.m</td>
<td>Record-2.m</td>
<td>Record-3.m</td>
<td>Record-4.m</td>
<td>...</td>
<td>Record-n.m</td>
</tr>
</tbody>
</table>

In this research, we need to calculate the number of instances for each variable. For instance, when we want to research the relationships between “Obesity” and “Diabetes”, we need to know how many individuals have “Diabetes” with the “True” value and how many with have the “False” value, and then know the number of individuals having the “True” value of “Obesity” and the “False” value. For each comparison of two variables, we only need to read two rows from a column-oriented database. Suppose the database system reads data sequentially row by row, a column-oriented design of data storage will save much more time than a traditional row-oriented one when the learning task is calculating numbers of
variables and analyzing relationships among variables.

6.3. Performance Improvements

Our novel methods show an improvement of learning causal structures to balance between accuracy and computational cost. These methods reduce the total number of conditional independence tests for learning causal structures of a partially ordered pair or a global network. With the inputs of background knowledge variables, our approach decomposes the graph and performs the learning process for each subgraph independently. When the causal structure of each component is learned independently, our methods can be implemented in parallel.

We discuss the performance improvements of local causal discoveries and methods of learning from partially known background knowledge respectively. Followed by the discussion of learning efficiency, we discuss the limitations pertaining to the problems in accuracy.

(1) Achievements of learning local causal structures.

Currently existing methods for local causal discoveries can identify the simple causal structure from three variables where they are causally related. However, they lack of identifying variables that are not causally related. This research provides the algorithms that are able to identify non-causal factors to fill in the shortage of locally examining causal structures. The heuristic methods are executable in polynomial time. All non-causally related variables are excluded in the set of conditioned variables.

(2) Achievements of learning causal structures from background knowledge.

Learning the structure of a complete causal network is computationally difficult. An assumed variables’ ordering can significantly reduce the time complexity and searching space though such a process of learning with a large number of variables is still intractable. A complete ordering may not exist in real health science dataset but the precedence (the order) of a few variables may be partially known. This research takes advantage of utilizing partially known background knowledge to construct the causal network. Our methods have shown the efficiency over general learning methods. The experimental results show that we have achieved almost ten percent of time as compared to exact learning for discovering causal
structures within one decomposed graph. The experimental results suggest the effectiveness of a graph decomposition is the key to the reduction of learning time.

There are several limitations with our methods. First, this research is a constraint based learning method and in contrast to score-based methods, performance of this type of learning is affected by the sparseness of graph, completeness of data, and sample size of data [89]. Second, the entire network integrated from all learned subgraphs may contain more erroneous information. Third, the implementation of these algorithms itself has a cost with respect to learning time and searching space.

To settle the first issue with the constraint based learning method itself, score-based learning methods may be considered and the comparisons of performance will be evaluated in our future work. To overcome the second issue, an additional learning process for a combined graph is needed. Wrongly directed edges and undiscovered edges may be detected with higher-order conditional independence test by compromising learning efficiency and accuracy. To solve the third issue, we aim to find a balance of time between obtaining decomposed graphs and learning from a complete structure through repeated tests on conditioned variables.

6.4. Implications on Health Science Research

Methods of local causal discoveries and algorithms for learning causal structures through graph decompositions are presented. They have been proven to be theoretically effective, compared with the exact learning algorithms such as PC algorithm. Local learning algorithms advance in efficiently discovering a causal structure among three variables in linear time and graph decomposition algorithms have advantages in identifying causal relationships for variables with partially known background knowledge.

Computational methods have been shown to be able to identify both causal and non-causal factors efficiently and perform structural learning process with partially known background knowledge. Next we discuss the following issues how this research facilitates and differs from health science study.

1) Analysis of one causal relationship.
In an observational dataset, not all variables need to be considered if they are neither causally related nor relevant to one specific aim. In an experimental dataset, all variables that are collected in the survey are believed to be relevant nevertheless. It may be unnecessary to get a complete causal network from all variables from an observational dataset because there is a specific goal such as to identify one certain causal relationship only.

In spite of a global causal network, discovering and analyzing one particular causal relationship is also interested by epidemiologists. For example, if the research focus is just to find out all causally related factors for one particular pair of variables such as “Smoking” and “Lung cancer”, but we have no extra interests at this moment in any other variables’ relationships, we will focus on learning from a smaller set of variables. Local learning algorithms explicitly determines the target variable for research and are processed with a smaller set of variables. Graph decompositions remove unrelated variables based on conditional independence test and they are advantageous in learning the structure for a local graph rather than for a global one.

2) Variables’ selection in health science study.

In health science study, researchers need to prioritize the experiments for analyzing causal relationships with limited resources and a limited time frame. Not all experiments, clinical trials or the controlled study can be carried on for causal analysis. Particularly, a clinical experiment has only one objective with at most two variables and it is designed for a specific purpose to prove the existence of causal relationship between an exposure and an outcome. The decision of which variables to be selected in the experimental design remains a question unless researchers have prior knowledge which exposure is observed and what an outcome is expected.

In a cross-sectional study, researchers are able to get an estimate of the relationships between potential exposure and its possibly related diseases. Evidently, not all variables will be considered to be in the stage of clinical experiments because we have no reason to believe that all surveyed variables must be related to a disease. In addition, we cannot claim that variables that do not exist in the survey must be unrelated to the disease. Moreover, the
cost of examining all pairs of variables in an experimental setting is unaffordable in reality.

If researchers want to obtain as many causal relationships as possible when there is a limited number of controlled experiments they need to conduct, novel methods presented here may provide assistance in the decision making process. A computational approach through statistical test of conditional independence can efficiently and effective tell the structure of correlated variables so that these correlated variables will be examined further in medical research but uncorrelated variables may possibly be excluded in future research. Examples of such decision problems are: how to improve the experimental plan aided by the graphical analysis of a disease network, how to use existing knowledge to infer previously unknown causal structures, and what kinds of irrelevant variables should be excluded during the design of experiment.

3) Relationships between computational models and causality in health science research.

In addition to the discussions on theoretical aspects of research improvement, research scope and decision problems in health science study, we need to point out that the discovery of causal relationships in this research is not a substitute for any experimental study in health science. This research is motivated by the need to efficiently discover causal structures in the dataset and causal structures are depicted as directed acyclic graphs. A directed edge connecting two nodes “suggests” a directional relationship and causality between two variables. Causal models [44] and directed acyclic graphs are different. For instance, in observational epidemiology, a directed acyclic graph may contain many latent nodes to explain real causal relationships among variables. Causal relationships may not be correctly inferred unless the mechanisms that generate the data are known [37]. The abstraction of causal models by directed acyclic graphs only reflects the existing knowledge of relationships on observed variables.

Causal modeling itself [12] is an abstract quantitative method to represent causal relationships of the real world, but it does not give researchers any estimates and tests of actual target effects. Researchers are able to warrant causal inferences with external
validations in real medical or clinical datasets or predict disease progressions in controlled experiments with the help of causal modeling in statistics and efficient analytical methods in computing.

6.5. Summary

Motivated by chronic disease modeling in epidemiology, inspired by the interests of causal reasoning, and driven by the need of solving big data problems in computer science, we have described the necessity of computational approaches for causal discovery in epidemiology and in health science study. Big data problems exist in epidemiological study due to a large number of variables in the dataset. Learning the causal network from data requires the tests of conditional independence from the all permutations of variables and thus the computational cost of learning the optimal structure is not negligible.

As a remedy from computer science, we have designed a background knowledge-based modeling framework, emphasized on the structural learning of causally related disease network, provided methods for local causal discoveries as well as several extended versions of current algorithms, and showed a solution for subgraph decomposition. Moreover, we have presented three graph decomposition algorithms to learn causal structures with background knowledge for the entire network. We have provided a novel strategy by incorporating background knowledge and precedence of variables supplementary to the existing learning approaches. We focus the contributions in constraint based structural learning methods. We summarize our contributions as a review of these algorithms in this dissertation.

1) Discovering non-causal factors and learning locally non-causal relationships.

Our algorithm (CCCx) for discovering non-causal factors is the extension of the CCC rule [120]. In the CCC rule, we assume one acausal variable. In the CCCx rule, we assume that one variable is fixed to be a consequence that does not possibly cause other variables in the dataset. These algorithms rely on the same research assumptions such as d-separation properties [99] in particular. However, there is a significant difference with their use in practice. The CCC rule presented in the original literature is motivated to discover causally related variables in market data. The CCCx rule on the other hand is used to discover
non-causally related variables in epidemiological data. In reality, any non-causal factors to a disease will be excluded in future experimental design. Complexity of learning is reduced because fewer variables need to be conditioned. CCCx rule takes three variables in each iteration. When \( n \) variables are considered, time complexity for the extension of the CCCx rule to discover all locally non-causal factors from a single factor is \( O(n^3) \) in the worst case.

(2) Graph decompositions from background knowledge.

There is an assumption with the methods for local causal discoveries that one variable has to be either an acausal factor (i.e. it is not caused by others) or a factor of consequence (i.e. it does not cause any others). When such an assumption does not hold in reality, we realize that background knowledge variables can be anywhere in the causal network. Algorithms for graph decompositions learn a smaller component of the network through decomposing graphs from background knowledge. Each component, as a disjoint data structure, can be learned independently. The running time for finding disjoint datasets is satisfactory in practice. The advantage of implementing graph decompositions to learn each component independently is the possibility of learning smaller structures in parallel.

(3) The algorithm of removing variables from the triad structure.

The first step of our graph decomposition algorithms decomposes the entire association graph into multiple components. The graph is decomposed by the removal of background knowledge variables. This algorithm can correctly remove variables that are not in any causal paths from one background knowledge variable to the other (if these two variables are an ordered pair). However, it does not make an effort to remove all variables that are unnecessarily to be conditioned for the structural learning. The method of removing variables from the triad structure enhances the decomposition process. In this algorithm, we perform conditional independence tests. In a three-variable structure, these tests check if two variables are conditionally independent given the background knowledge variable. Similar to the CCC rule and CCCx rule, if one variable is conditionally independent from the other given a third variable, we can exclude such an variable from any variables’ set containing the other. The running time is polynomial because this algorithm takes exactly three variables in each
iteration for the test of conditional independence.

(4) Removal of sink structures.

Variables in a sink structure can be removed from the association graph if we want to learn the causal path for a particular pair of variables. Since variables in a sink structure are not in the causal path between two background knowledge variables, examining the relationships of variables in the sink structure is unnecessary. However, a sink structure should not be removed for learning a complete network structure. A sink structure itself may not contain any background knowledge variable and the learned network from a sink structure may thus violate the pre-set conditions of variables’ ordering.

The experiments from the synthetic dataset have shown the benefits with respect to the reduction of learning time and space. For the methods of learning causal structures by incorporating background knowledge, the total learning time is reduced both in theory and in experiments, albeit dependent on specific inputs of data.

There are mature methods such as stratified analysis to obtain variables’ relationships in epidemiological research. However, the lack of solutions to identify and analyzing all possible combinations of variables make the stratified analysis computationally infeasible in reality. Such an issue provokes us to develop computational models including local causal discoveries and learning causal structures with background knowledge to make the problem more tractable.

In summary, we provide efficient ways for computer scientists and public health professionals being able to discover much useful information from observational data through heuristics, even though we cannot promise to discover all causal or non-causal relationships, nor can we find a way to solve all exact causal inference problems in a particular polynomial time. Partially identified findings of potential causal relationships may still facilitate future research, by not having to mix causes and effects associated with a certain disease and by not having to wrongly include non-causal factors in any analytical models for disease prevention and disease prediction.
6.6. Future Work

For the initial research on chronic disease modeling, several publicly available datasets from Behavioral Risk Factor Surveillance System (BRFSS), National Health Interview Survey, National Health and Nutrition Examination Survey, Healthcare Cost and Utilization Project (HCUP), Hospital inpatient discharge data, National Survey of Children’s Health, and Vital Statistics Data have been collected. However, due to the nature of these available datasets that are not intended to be used for causal discoveries in practice by medical professionals, we herein use them partially for the demonstration purpose only.

Experimental results from synthetic datasets, however, have shown the performance improvements as indicated in theory with respective to the reduction of learning time and searching space.

Several open questions and challenges are left unexplored and we leave these for future research endeavors.

(1) Thresholds to determine associations. Determinations of dependence and independence are based on calculated numbers. For the chi-squared test, an alpha value of 0.05 is set. For odds ratio, we assume 2.0 is the threshold to determine association. When such kinds of thresholds change, the output association graph will be different. Whether the value of a threshold is set properly to best represent the true relationships is an unsolved issue in our experiments.

(2) Possible extensions from current causal models. Bayesian networks as well as causal networks cannot handle bi-directional relationships of variables due to the nature of directed acyclic graphs. However in reality, a “cyclic feedback” structure may exist. For instance, “low income” may lead to “poor health condition”, which in return leads to “low income”. In our experiments, there are certain relationships classified as either no explicit direction of variables or bi-directional. Whether such kinds of undirected relationships are due to their possible cyclic structure remains to be another unsolved issue. Temporal analysis with external datasets may be required to ascertain this issue.

(3) Ability to discover hidden variables in data. We assume there is no unobserved
variables in the entire structure, but in reality there may be many variables hidden in the causal path. The challenge of this problem is that there are possibly too many variables from environment or individual’s behaviours we may want to include but only few of them can be identified to be possibly associated. Furthermore, for those few variables that are associated, only a small portion of them can be proven to be possibly causal.

(4) A wider acceptance by public health professionals. Computational epidemiology absorbs the merits of fast processing data in computer science to feed epidemiological research. The credibility of discovered causal relationships by computational methods is a challenge because these methods may have taken heuristic approaches to trade-off accuracy for speed albeit computationally feasible. How to balance computational feasibility and the credibility of the discovered results will be a major problem in the future research.

Particular in computer science and for the future work, our developed methods will be extended in a parallel learning setting. Parallel learning algorithms for Bayesian networks exist [72] [125] for score based learning, but there is no available algorithm for constraint based learning when background knowledge variables are incorporated in the network. Issues such as fault tolerance, load balancing problems, and inter-processor communication cost regarding to the parallel implementation will be weighed in. In the subsequent study we will improve our methods by considering processing data more efficiently with context-specific independence properties [14] and by overcoming problems of order-dependent issues in the constraint-based learning [22]. For analysis with context-specific independence, it examines variables’ conditional independence given a specific context of the third variable. For example, two variables are conditionally independent given the status of the third variable being “True” but not conditionally independent given the status of the third variable being “False”. An order-independent constraint-based learning may solve the issue of getting different results due to the learning order of conditional independence test and the improvement of accuracy of learned structure is expected.
A.1. Proof of the CCCx Rule

The properties of conditional independence in Bayesian Networks and their corresponding proofs have been mentioned in previous literature [11] [31] [74] [102] [115]. We use them as the prerequisite for the proofs of our theorems.

Given a local structure with three variables where all Bayesian network conditions are met, and for each variable that is connected by two other variables, there are three structures: 1) head-to-tail, 2) tail-to-tail, and 3) head-to-head, shown in Figure A.1.

For the first two structures, conditioning on the variable that is either head-to-tail or tail-to-tail linked to the other two variables can make these variables conditionally independent, if such two variables were previously dependent without observation of the conditioned variable.

For the third structure, conditioning on the head-to-head linked variable (i.e. collider) can make two previously independent variables which are both linked to it become conditionally dependent.

Both CCU rule and CCC rule mentioned in the section of research methods follow such prerequisite. As previously pointed out, we use “correlated,” “dependent,” and “associated” interchangeably. Correlation here does not represent correlation coefficient, and all of these terms are non-directional. For CCCx rule, as it is derived directly from CCC rule, we show the theorem Theorem A.1 below:

**Theorem A.1.** For any three variables A, B, and C in a given dataset, where all previously stated assumptions are met, and these three variables are pairwise correlated, knowing that C...
has no known consequence, if after conditioning on variable B, A and C become independent from each other, then A is not the cause of C.

Proof. Before conditioning on a variable A or B, A and C were correlated, and B and C were correlated. In our causal model, if there is a causal relationship between two variables, there must be a direct link between them. Since we know that C is neither the cause of A nor B, and there is a link between A and C, and a link between B and C, we can conclude that A and/or B can only be the cause of C, if there exists any causal relationship.

Since A and B are correlated, which means there is a link between A and B, A can cause B or B can cause A. A and B are equivalent variables in the pre-assumed case. If we select B for studying, the basic structure can only be either from A to B to C, or from B to both A and C. Figure A.2 shows the structures with three variables after conditioned on variable B.

![Figure A.2. Structures with Three Variables after Conditioning.](image)

The first condition for variable B is head-to-tail and the second condition is tail-to-tail. From the lemma shown previously, when one of the structures (head-to-tail or tail-to-tail) is met, conditioning on variable B will make A and C conditionally independent. Since A and C become independent given B, the previous direct link between A and C is “broken”, so there is no direct link between A and C. Hence, we conclude that A will not be the cause of C.

A.2. Proof of the Algorithm for Discovering Non-Causal Factors

From the proof of the CCCx algorithm, we can extend the theorem by showing the corollary (Corollary A.2) next.

**Corollary A.2.** For all variables, with each (A) that was previously correlated with a variable (C) that is a non-causal factor for any variables in the dataset, if conditioning
on any one of the third variables (B), where such third variable (B) and the two correlated variables (A and C) were pairwise correlated, can make these two variables (A and C) become conditionally independent, then such variable (A) will not be the cause of the non-causal variable (C).

**Proof.** We show a basic local structure of variable A, B, and C. Suppose A just stands for one of the variables other than B and C in the entire dataset, A, B, and C are pairwise correlated, and C is a non-causal variable. We can temporarily establish an undirected link from B to C denoting correlation only, or a directed link from B to C denoting B causes C. No matter which way we choose, such a link cannot be from C to B because we already state that C does not cause any other variable. Then for variable B, there will not be a head-to-head condition. In other words, variable B is not a collider. On the other side of variable B that is connected with A, no matter if A causes B or B causes A, conditioning on such non-collider B will make A and C independent from each other. Since A and C become independent, given any conditioned variable in any one of the situations, A will not directly cause C, so that A is not a cause of C.

Variable A may still be dependent on C, given any other variable B’, but this will simply indicate that selection of variable B’ does not block the causal link between A and C. However, if under one particular selection of a variable for conditioning, A and C become conditionally independent, A will not be the cause of C no matter what the possible relation between A and B’ is. **Figure A.3** shows the structures with three variables B’, A, and C, when conditioning on B’ does not make A and C independent.

![Figure A.3](image)

**Figure A.3.** A structure with Three Variables with No Conditional Independence.


[54] Jiawei Han, Micheline Kamber, and Jian Pei, *Data mining: concepts and techniques: concepts and techniques*, Elsevier, 2011.


[80] Jiuyong Li, Ada Wai-chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay,


[129] Guillaume Wunsch, Federica Russo, and Michel Mouchart, *Do we necessarily need
longitudinal data to infer causal relations?, Bulletin de Méthodologie Sociologique 106 (2010), no. 1, 5–18.
