ROBUST MEASURES OF CAUSAL DEPENDENCE IN RELATIONAL DOMAINS

A Dissertation Outline Presented

by

DAVID ARBOUR

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

October 2015

College of Information and Computer Sciences

ROBUST MEASURES OF CAUSAL DEPENDENCE IN RELATIONAL DOMAINS

A Dissertation Outline Presented

by

DAVID ARBOUR

Approved as to style and content by:

David Jensen, Chair

Ben Marlin, Member

Nick Reich, Member

Dan Sheldon, Member

Bruce Croft, Dean
College of Information and Computer Sciences

ABSTRACT

ROBUST MEASURES OF CAUSAL DEPENDENCE IN RELATIONAL DOMAINS

OCTOBER 2015

DAVID ARBOUR

Directed by: Professor David Jensen

TABLE OF CONTENTS

Page

ABSTRACT . . . . **iv**

LIST OF TABLES . . . . **vii**

LIST OF FIGURES . . . . **viii**

CHAPTER

1. INTRODUCTION **1**

2. RELATIONAL BACKGROUND **3**

3. PROPOSED CONTRIBUTIONS **6**

3.1 Inferring Causal Direction of Relational Dependence 6

3.1.1 Problem Setting 7
3.1.2 Proposed Approach 9

3.1.2.1 Direction Under Linear Dependence 9
3.1.2.2 Reasoning About Confounding 12
3.1.2.3 An Extension to Non-Linear Dependence 14

3.2 Non-Parametric Testing of Relational Dependence 16

3.2.1 Proposed Approach 17
3.2.2 Existing Work Testing Causal Direction and Marginal Dependence of Relational Data 19

3.3 Measuring Conditional Dependence in Relational Data 21

3.3.1 Existing Work 22
3.3.2 Proposed Approach 23

3.4 Improved Structure Learning for Relational Domains 26

4. PROPOSED TIMELINE 28

APPENDICES

A. SYNTHETIC EXPERIMENTS 29
B. PROOFS 36

BIBLIOGRAPHY 40

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

2.1 Example relational skeleton for the Foursquare domain. This could be a small fragment of a (potentially) larger skeleton. 4

2.2 Relational model for the Foursquare domain. The underlying relational schema (ER diagram) is shown in black. The attributes on the entities are fictional. A relational dependency is shown in gray. The model shown represents the joint distribution of the domain. 5

3.1 An example of induced relational bias that occurs due to auto-dependence. 21

3.2 An example of induced relational bias that occurs *after* conditioning. 22

A.1 Scatterplots for the sum of X values of related nodes (x-axis) vs. the sum of X values of related nodes with additive Gaussian noise (y-axis). The noise coefficient (c_ϵ) varies from 0 to 2. The underlying network structure is a regular network of degree 10 with 500 nodes. 29

A.2 Orientation accuracy for regular graphs for varying degree (A.2a), size of network (A.2b), and noise coefficient (A.2c). 30

A.3 Orientation accuracy for various network types and functional forms, as the size of the graph increases. The noise coefficient is set to 0.5. 33

A.4 Orientation accuracy for various network types and functional forms, as the coefficient of the noise increases. The network size was kept constant at 1000 nodes. 35

CHAPTER 1
INTRODUCTION

Inferring causal relationships from observational data is an important task in many fields such as epidemiology, sociology and political science where experimentation is difficult or impossible. There have been great advances in field of causal inference during the past fifty years (c.f. Pearl [39], Rubin [44], Dawid [10]), but the overwhelming majority of this work assumes that the data instances are independent and identically distributed (i.i.d.). However, many real-world systems arise from systems that are structured as networks (e.g., social, technological and biological systems). Instances in these systems can be represented as interconnected nodes in a graph in which the attributes of each node are often correlated. Such *relational* data contain dependent instances, and thus violate the i.i.d. assumption. Recent work has introduced methods for learning causal structure of relational domains [30, 34]. While this work represents a significant advance, the results (both theoretical and experimental) assume an idealized setting with the presence of an independence oracle. This prevents these methods from being reliably used by practitioners.

The focus of the first two-thirds of this thesis will be to develop practical tests for causal direction and marginal and conditional dependence in the relational data setting. To date, we have studied the conditions under which the *causal direction* can be consistently estimated by comparing the test statistic in both directions. This finding implies that the set of causal directions that can be discovered from a relational observational dataset is strictly larger than in the i.i.d. setting. We have also created a consistent test of dependence where at least one variable is a relational variable (i.e., a variable that consists of a set of instances). We propose to further this work by creating consistent tests of conditional dependence for relational data. The causal direction findings have important implications for algorithms that learn the causal structure of a relational domain. Developing an algorithm that leverages this information, and characterizing the set of identifiable causal dependencies (i.e. the Markov equivalence class [46]), will be the focus of the final third of the thesis.

CHAPTER 2
RELATIONAL BACKGROUND

In this section, we introduce the basic relational concepts, following the notation and terminology of Maier, Marazopoulou, and Jensen [31] that will be used throughout the dependence testing, and structure learning sections. A *relational schema* $\mathcal{S} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, card)$ specifies the set of entity, relationship, and attribute classes of a domain. It includes a cardinality function that imposes constraints on the number of times an entity instance can participate in a relationship. A relational schema can be graphically represented with an Entity-Relationship (ER) diagram. Figure 2.2 shows the ER diagram for the Foursquare domain. Foursquare is an online social network where users "check-in" to locations using their mobile phones. In this example, there are three entity classes (*User*, *Place*, *Hometown*), and three relationship classes, (*Friends*, *ChecksIn*, *From*). The entity class *User* has three attributes: *Smokes*, *Weight*, and *Drinks*. The cardinality constraints are depicted using crow's feet notation. For example, the cardinality of the *From* relationship is one-to-many, indicating that one user has one hometown, but many users can be from the same hometown.

A *relational skeleton* is a partial instantiation of a relational schema that specifies the set of entity and relationship instances that exist in the domain. Figure 2.1 depicts an example relational skeleton for the Foursquare domain. The network consists of two *User* instances, Alice and Bob, who are friends with each other and come from the same hometown. There are two *Place* instances, Hillside Diner and Corner Cafe.

Given a relational schema, one can specify *relational paths*, which intuitively correspond to possible ways of traversing the schema (see Maier, et al.[31] for a

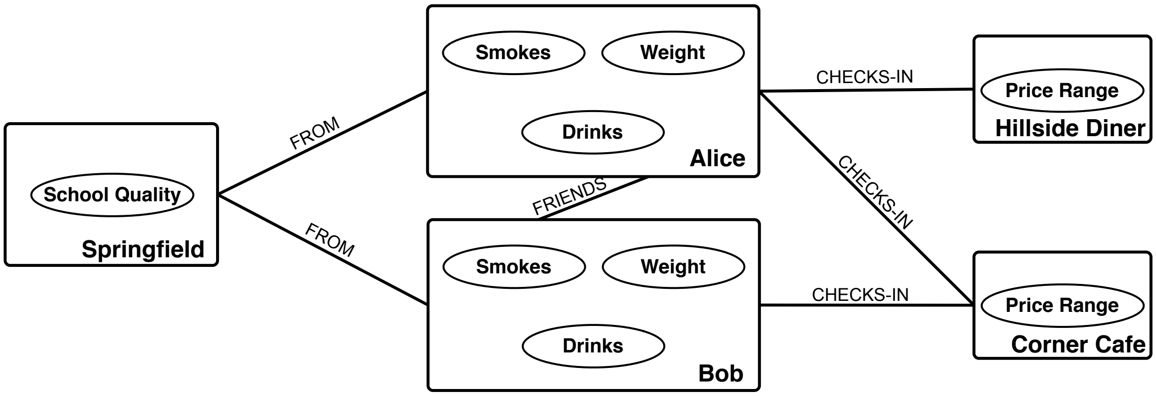


Figure 2.1: Example relational skeleton for the Foursquare domain. This could be a small fragment of a (potentially) larger skeleton.

formal definition). For the schema shown in Figure 2.2, possible paths include $[User, Friends, User]$ (a person's friends), and $[User, Friends, User, From, Hometown]$ (the hometowns of a person's friends). *Relational variables* consist of a relational path and an attribute that can be reached through that path. For example, the relational variable $[User, Friends, User].Drinks$ corresponds to the alcohol consumption of a person's friends. We briefly note that the logical predicate used to construction of this set can be defined in a number of ways. For example, we can define $[User, Friends, User]$ to be the set of friends of friends for an individual either exclusive or inclusive of that individual's friends. See Marazopoulou, Arbour, and Jensen [33] provide further details and show the impact of the choice of path predicates on effect estimation. Probabilistic dependencies can be defined between these relational variables. In this work, we consider dependencies where the path of the outcome relational variable is a single item. In this case, the path of the treatment relational variable describes how dependence is induced. For example, the *relational dependency*

$$[User, Friends, User].Drinks \rightarrow [User].Weight$$

states that the alcohol consumption of a user's friends affects that user's weight.

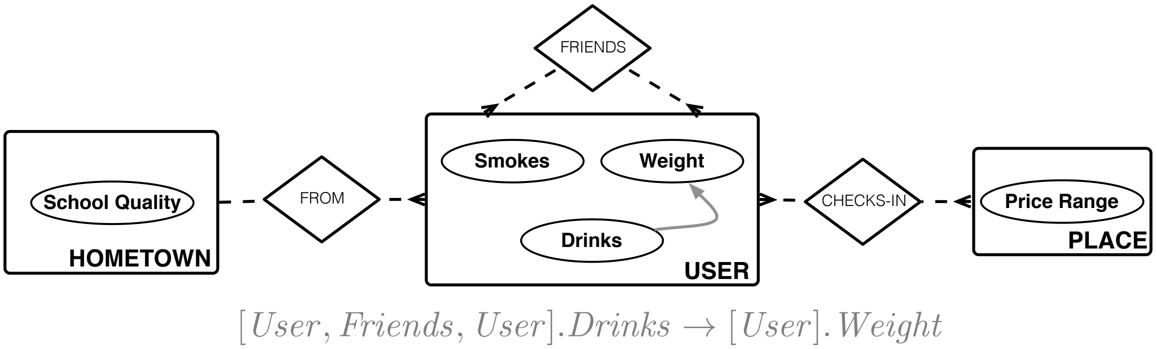


Figure 2.2: Relational model for the Foursquare domain. The underlying relational schema (ER diagram) is shown in black. The attributes on the entities are fictional. A relational dependency is shown in gray. The model shown represents the joint distribution of the domain.

A *relational model* $\mathcal{M} = (\mathcal{S}, \mathcal{D}, \Theta)$ is a collection of relational dependencies $\mathcal{D}$ defined over a relational schema along with their parameterizations Θ (a conditional probability distribution for each attribute given its parents). The structure of a relational model can be depicted by superimposing the dependencies on the ER diagram of the relational schema, as shown in Figure 2.2, and labeling each arrow with the dependency it corresponds to. If labels are omitted, the resulting graphical representation is known as a *class-dependency graph.*

Recent work by Maier, et. al [31] provides a framework that enables reasoning about *d*-separation in relational models. Toward that end, they introduce *abstract ground graphs* (AGGs), a graphical structure that captures relational dependencies and can be used to answer relational *d*-separation queries. Abstract ground graphs are defined from a given perspective, the base item of the analysis, and include nodes that correspond to relational variables.

CHAPTER 3

PROPOSED CONTRIBUTIONS

3.1 Inferring Causal Direction of Relational Dependence

Inferring the direction of causal dependence between two random variables from observational data is a fundamental problem in statistical reasoning. There have been many advances in this area for data sets that are independent and identically distributed (i.i.d.) [21, 47, 29]. For relational data, recent work has studied the problem of inferring the effects of peers via *experimentation* [36, 3, 49]. However, the problem of identifying causal direction from *observational* relational data has yet to receive the same focus. In this work, we study the problem of inferring the causal direction of peer dependence in observational relational data. We provide theoretical and experimental results to show that the causal direction of peer dependence can be robustly inferred from observational data by comparing the magnitude of two similarity measures (one for each candidate direction).

For example, consider the study of the causes of personal debt. Data consist of the net worth and the average monthly discretionary spending of a large set of individuals, along with the position of each individual within a social network. One reasonable question is whether a person's friends influence his or her spending habits. If a person's spending and wealth are correlated with the wealth and spending of their friends, what can be inferred about the *causal* dependence among these quantities? A person's spending could be caused by their friends' wealth or vice versa (direct dependence), or both quantities could be caused by an unobserved variable (confounding).

We propose to examine when and how it is possible to differentiate among these scenarios. Specifically, we:

1. Identify a set of conditions under which the causal direction of relational dependence can be consistently inferred.
2. Investigate the effect of unobserved confounding on this approach to causal inference, and provide a simple test of relational confounding.
3. Provide an extension of our method to the case of non-linear dependence via kernel embeddings.

In the appendix we also show that the proposed measures are robust to both the magnitude of the noise and the functional form of the true dependence, through a set of simulations under a variety of graph structures and functional forms.

3.1.1 Problem Setting

Let $G = \langle V, E \rangle$ be a graph with n vertices.[1] Every node of the graph $v_i \in V$ is associated with a pair of random variables, X_i and Y_i. For every node, we can define a new random variable as a function of the random variables of its neighboring nodes. Specifically, in this section, we define a new random variable $X_i{}'$ as the sum of X_j over v_i's neighbors:

$$X_i{}' = \sum_{\{v_j | \langle v_i, v_j \rangle \in E\}} X_j$$

Similarly, $Y_i' = \sum_{\{v_j | \langle v_i, v_j \rangle \in E\}} Y_j$.

For the remainder of the paper, we refer to functions of random variables of neighboring nodes, such as $X_i{}'$ and $Y_i{}'$, as *relational variables* and to random variable of the node, such as X_i and Y_i, as propositional variables. To avoid ambiguity, we refer to dependence between a relational variable and a propositional variable as *relational dependence*.

A very common assumption in relational domains is that of *templating*, i.e., random variables in different nodes follow the same distribution [24]. In our case, this would imply that the distribution of X_i is the same for all i (and the same for Y_i, X_i', and Y_i'). This allows us to reason about four random variables on a model level: X, Y, X', and Y'. The task under consideration is determining the causal direction of relational dependence. Put in another way, we wish to determine whether $X' \rightarrow Y$ or $Y' \rightarrow X$ is the true generative process.

Since we are reasoning over random variables across all nodes of the network, it is convenient to represent them as vectors. Let $\mathbf{x} = \langle X_1, \ldots, X_n \rangle$ be a vector with the random variables X_i for every node and, similarly, $\mathbf{x}' = \langle X_1', \ldots, X_n' \rangle$. Let A denote the adjacency matrix of the graph defined as:

$$A_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E. \\ 0, & \text{otherwise.} \end{cases}$$

We use D to denote the degree matrix of the graph:

$$D_{ij} = \begin{cases} d_i, & \text{if } i = j. \\ 0, & \text{otherwise.} \end{cases}$$

We note that A is a symmetric matrix as a consequence of assumption A1. We can write the vector of the sum of the friends (i.e., the vector $\mathbf{x}'$) as:

$$\mathbf{x}' = A\mathbf{x}.$$

[1] For clarity of exposition, we focus on the case of a single-entity, single-relationship network. The extension to the more general multi-entity case is straightforward, and can be applied using the relational background introduced in the background section.

Similarly, $\mathbf{y}' = A\mathbf{y}$.

Throughout the paper, we make the following assumptions:

A1. *G is undirected.*

A2. *Each node $v \in V$ has degree of at least 1.*

A3. *The distribution of X and Y is the same for all $v \in V$ (templating).*

A4. *There are no feedback cycles, i.e. $Y \rightarrow X \Rightarrow X \nrightarrow Y$ for any two (relational or propositional) variables.*

Further, we initially assume (and later relax that assumption) that:

A5. *There are no confounding variables.*

Section 3.1.2.2 is devoted to examining under which conditions this assumption can be loosened, while maintaining the ability to identify causal direction. Moreover, assumptions A4 and A5 mirror those found in the literature on determining causal direction between two propositional variables [47, 21, 29].

3.1.2 Proposed Approach

3.1.2.1 Direction Under Linear Dependence

In this section we show that, under the assumptions of linearity and absence of noise (deterministic dependence), peer dependence is asymmetric and the true causal direction can be consistently inferred. This is an inherent property of relational domains. The extension to non-linear dependencies is provided in Section 3.1.2.3.

To measure dependence between variables, we consider the square of Pearson's correlation, a common and widely employed measure of linear correlation between variables. Pearson's correlation between two variables X and Y can be computed from a sample $\mathbf{x}$, $\mathbf{y}$ as follows:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the means of $\mathbf{x}$ and $\mathbf{y}$ respectively. We consider the square of the correlation to restrict the range of the metric to [0,1], rather than [-1,1].

Given a measure of dependence, a reasonable question is whether the measure is symmetric for relational data. Surprisingly, it is not. Given this, another reasonable question is what can be inferred by examining the dependence values in both directions. Surprisingly, the causal direction of dependence can be inferred from the resulting asymmetry.

We begin by handling a simplified case: Y is a deterministic function of the X values of related nodes. Specifically, we assume that Y_i is the scaled mean of the X_j variables of the related instances:

$$Y_i = \frac{\beta}{d_i}\sum_{i=1}^{d_i} X_j$$

Or, in matrix notation: $\mathbf{y} = \beta D^{-1} A\mathbf{x}$.

Under certain assumptions about the structure of the graph and the form of the dependence, the squared correlation in the causal direction will be greater that the squared correlation in the opposite direction.

Proposition 1. *Assume that G is a d-regular graph*[2]*, the true generative process is $\mathbf{y} = \beta \cdot D^{-1} A\mathbf{x}$ for some constant β and assumptions A1-A5 hold. Then, $\rho^2(\mathbf{x}', \mathbf{y}) > \rho^2(\mathbf{y}', \mathbf{x})$.*

[2] A d-regular graph is a graph where every vertex has degree d.

Proof. The left-hand-side of the inequality, given that by definition $\mathbf{x}' = A\mathbf{x}$, can be written as:

$$\begin{aligned}\rho^2(\mathbf{x}', \mathbf{y}) &= \rho^2(A\mathbf{x}, \beta D^{-1}A\mathbf{x}) \\ &= \rho^2(A\mathbf{x}, \frac{\beta}{d}A\mathbf{x}) = 1\end{aligned}$$

It remains to show that $1 > \rho^2(\mathbf{y}', \mathbf{x})$ which holds, unless $\rho^2(\mathbf{y}', \mathbf{x}) = 1$. Equality holds only when $\mathbf{y}' = \beta AD^{-1}A\mathbf{x}$ is a linear combination of $\mathbf{x}$, or in words, when the values of a node's friends of friends are a linear combination of that node's value. For random values of X, that happens for a degenerate network structure where every node has one friend of a friend and is the exact same starting node. This would happen, for example, in the case of a regular graph with degree 1 (pairs of nodes). □

In the case where Y is a noisy function of X, a similar inequality holds.

Proposition 2. *Assume that the true generative process is* $\mathbf{y} = \beta \cdot D^{-1}A\mathbf{x} + \epsilon$ *for some constant* β*, where* ϵ *is a vector with the noise terms. Moreover, assume that assumptions A1-A5 hold and* X *and* Y *are scaled to mean 0. Then the following holds:*

$$\rho^2(\mathbf{x}', \mathbf{y}) > \rho^2(\mathbf{y}', \mathbf{x}) \Leftrightarrow$$

$$\frac{Var(AD^{-1}A\mathbf{x}) + Var(A\epsilon)}{Var(D^{-1}A\mathbf{x}) + Var(\epsilon)} > \frac{Var(A\mathbf{x})}{Var(\mathbf{x})}.$$

A full derivation can be found in the appendix. The implication of proposition 2, is that we can expect the causal direction to be accurately inferred so long as the relative influence of the noise distribution is relatively small in comparison to the relationship between $AD^{-1}\mathbf{x}$ and $\mathbf{y}$. As we show during our experimental evaluation in Section A, the method is quite robust to the effect of noise in practice.

3.1.2.2 Reasoning About Confounding

Throughout Section 3.1.1 we assumed the absence of confounding influences (assumption A5). However, in many real-world settings, this proves to be an unrealistic assumption. Within the relational setting, there are two distinct ways in which the relationship between variables can be confounded:

1. $\mathbf{x}$ and $\mathbf{y}$ may share a common relational cause, $A\mathbf{z}$, i.e., $A\mathbf{z} \to \mathbf{x}$ and $A\mathbf{z} \to \mathbf{y}$.

2. There is a variable $\mathbf{z}$ that is a non-relational cause of $\mathbf{x}$ and a relational cause of $\mathbf{y}$, i.e., $\mathbf{z} \to \mathbf{x}$ and $A\mathbf{z} \to \mathbf{y}$.

In what follows, we show that the first scenario is identifiable from data, while the second one is not.

Proposition 3. *If $Cov(A\mathbf{x}, A\mathbf{y}) \geq Cov(A\mathbf{x}, \mathbf{y})$ and $Cov(A\mathbf{x}, A\mathbf{y}) \geq Cov(A\mathbf{x}, \mathbf{y})$, then there exists a relational variable which is a common cause of x and y.*

Proof. Assume that the true generative structure is:

$$\mathbf{y} \sim D^{-1}A\mathbf{z} + \epsilon_{\mathbf{y}}$$
$$\mathbf{x} \sim D^{-1}A\mathbf{z} + \epsilon_{\mathbf{x}}$$

The covariance between $A\mathbf{x}$ and $A\mathbf{y}$ is then given by

$$\begin{aligned}
&Cov(Ax, Ay)\\
&\quad= Cov(AD^{-1}A\mathbf{z} + A\epsilon_{\mathbf{y}}, AD^{-1}A\mathbf{z} + A\epsilon_{\mathbf{x}})\\
&\quad= Cov(AD^{-1}A\mathbf{z} + A\epsilon_{\mathbf{y}}, AD^{-1}A\mathbf{z}) +\\
&\qquad Cov(AD^{-1}A\mathbf{z} + A\epsilon_{\mathbf{y}}, A\epsilon_{\mathbf{x}})\\
&\quad= Cov(AD^{-1}A\mathbf{z}, AD^{-1}A\mathbf{z}) + Cov(AD^{-1}A\mathbf{z}, A\epsilon_{\mathbf{x}})\\
&\quad= Cov(AD^{-1}A\mathbf{z}, AD^{-1}A\mathbf{z})
\end{aligned}$$

The covariance between $A\mathbf{x}$ and $\mathbf{y}$, is given by:

$$\begin{aligned}
&Cov(Ax, y)\\
&\quad= Cov(AD^{-1}A\mathbf{z} + A\epsilon_{\mathbf{x}}, D^{-1}A\mathbf{z} + \epsilon_{\mathbf{y}})\\
&\quad= Cov(AD^{-1}A\mathbf{z}, D^{-1}A\mathbf{z} + \epsilon_{\mathbf{y}}) +\\
&\qquad Cov(A\epsilon_{\mathbf{x}}, D^{-1}A\mathbf{z} + \epsilon_{\mathbf{y}})\\
&\quad= Cov(AD^{-1}A\mathbf{z}, D^{-1}A\mathbf{z}) + Cov(D^{-1}A\mathbf{z}, \epsilon_{\mathbf{y}})\\
&\quad= Cov(AD^{-1}A\mathbf{z}, D^{-1}A\mathbf{z})\\
&\quad\leq Cov(AD^{-1}A\mathbf{z}, AD^{-1}A\mathbf{z})
\end{aligned}$$

$AD^{-1}A\mathbf{z}$ and $D^{-1}A\mathbf{z}$ are bounded by the size of the intersection between the set of a node's immediate neighbors and the set of its two-hop neighbors, since we have assumed $\mathbf{z}$ are marginally independent by construction. Each pair of one hop and two hop neighborhoods will diverge for at least the degree of the node for each node, since the two hop walk beginning from node i will return to that node an equal number of its degree, which implies the final inequality. □

Proposition 3 implies a very simple procedure for ruling out the presence of mutual relational confounding between two variables. First, the relative dependence is measured between $A\mathbf{x}, \mathbf{y}$ and $A\mathbf{y}, \mathbf{x}$ respectively. Then, these two values are compared against the measured dependence between $A\mathbf{y}, A\mathbf{x}$. If neither are larger than the between-relational variable dependence no determination of direction is made, since observed dependence is likely due to confounding.

We now turn to scenario two, which yields the following negative result:

Corollary 1. *Under confounding scenario 2, in the absence of noise, a false conclusion of dependence* $A\mathbf{x} \rightarrow \mathbf{y}$ *will be made.*

Proof. Assume the generative structure is given by:

$$\mathbf{x} \sim \mathbf{z}$$
$$\mathbf{y} \sim D^{-1}A\mathbf{z}$$

It can be immediately seen that the form of this dependence is identical to the form of proposition 1, where we substitute $\mathbf{z}$ for the $\mathbf{x}$ used in the proposition. By simple implication, we are given that under the no-noise setting we are given that an incorrect determination of direct causation will be made. □

Note that this also applies in the case of a small amount of noise, as provided for by proposition 2. This result shows that without the assumption of no-confounding a determination of non-causation can be reliably implied, but the converse is not necessarily true.

3.1.2.3 An Extension to Non-Linear Dependence

In the previous section, we showcased the applicability of our method for detecting linear dependence in relational data using correlation. An extension to more complex variables and non-linear dependence functions can be achieved by applying the kernel trick.

Some background on kernel embeddings is useful. Let $\mathcal{X}$ be a non-empty set, $(\mathcal{X}, \mathcal{A})$ be a measurable space where $\mathcal{A}$ is a σ-algebra on $\mathcal{X}$, and let $\mathscr{P}$ be the set of all probability measures, P, on $\mathcal{X}$. $\mathcal{H}$ is the RKHS of the functions $f : \mathcal{X} \to \mathbb{R}$ with the reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The *mean map* is a function $\mu : \mathscr{P} \to \mathcal{H}$ that defines a kernel embedding of a distribution into $\mathcal{H}$:

$$\mu_P = \mu(P) = \int_{\mathcal{X}} k(x, \cdot) dP(x)$$

If a characteristic kernel is used, then this mapping is unique, i.e., there is an injective function between a distribution and its kernel mean value. In this work, the purpose of kernel mean is twofold. For propositional variables, it is used to represent the underlying distribution and, as we shall see, can be used directly in a test for dependence. For relational variables, the mean embedding serves as an aggregation function for observations. The advantage of using the kernel mean embedding is that, under the assumption that the underlying distribution belongs to the exponential family, the underlying distributions are represented completely.

To reason over the distance between distributions, we define a second kernel, K, over the kernel means. Christmann, et al. [5] showed that if the kernel inducing μ (k) is characteristic and K is the Gaussian kernel, then K is universal and thus, characteristic. This kernel is defined as:

$$K(\mu_x, \mu'_x) = e^{\frac{\|\mu_x - \mu'_x\|^2_{\mathcal{H}}}{2\theta}} \tag{3.1}$$

where $\sqrt{\theta}$ is the bandwidth of the kernel.

In addition to this measure of similarity between relational instances, we define a dependence measure. The centered kernel target alignment is a normalized measure of dependence introduced by Cortes, et al. [7] within the context of multiple kernel learning. The measure is defined as:

$$\text{KTA}(\mathbf{x}, \mathbf{y}) = \frac{\langle K^c_{\mathbf{x}}, K^c_{\mathbf{y}} \rangle_{\mathcal{F}}}{\|K^c_{\mathbf{x}}\|_{\mathcal{F}} \|K^c_{\mathbf{y}}\|_{\mathcal{F}}} \tag{3.2}$$

Where $\| \cdot \|_{\mathcal{F}}$ is the Frobenius norm. $K^c_{\mathbf{x}}$ is a centered kernel matrix, defined as:

$$K^c_{\mathbf{x}} = \left[I - \frac{11^T}{m} \right] K_{\mathbf{x}} \left[I - \frac{11^T}{m} \right]$$

If a linear kernel is used, KTA reduces to squared Pearson's correlation, which has been our measure of focus thus far. Using this connection, the following corollary provides for consistent estimation of causal direction under the deterministic case with arbitrary functional dependence.

Corollary 2. *Under assumptions A1, A2, A3, A4, A5, and further assuming that the generative structure is given by* $\mathbf{y} = D^{-1}A\phi(\mathbf{x})\beta$*, then:*

$$KTA(A\mathbf{x}, \mathbf{y}) \geq KTA(A\mathbf{y}, \mathbf{x}).$$

This follows as a straightforward extension to proposition 1. Because we are given by assumption that $KTA(A\mathbf{x}, \mathbf{y}) = 1$, and KTA is bounded from above by one the inequality holds. Equality occurs only when the values of each node's friends of friends can be expressed as a sum of (feature-space embedded) values. For random values of X, this is reduced to the degenerate case of a graph of degree 1, as in proposition 1.

Proposed work for this thesis is to examine the conditions under which the nonlinear test of relational direction will consistently return a correct conclusion under additive and non-additive noise. We will also consider the case where one or more of the variables exhibits auto-dependence that does not arise as a consequence of confounding.

3.2 Non-Parametric Testing of Relational Dependence

Detecting the presence of dependence between two random variables is fundamental to statistical and causal reasoning. Under the assumption of independent and identically distributed (i.i.d.) data, there are a number of consistent methods for detecting the presence of dependence (c.f. Gretton and Gyorfi [20]). In this paper, we focus on the *relational* setting, where data instances are not i.i.d. Specifically, our task is to detect the presence and the direction of dependence from relational (or *network*) data.

To formalize the notion of dependence between a random variable and the distribution of related instances, we address two key issues. First, we specify an aggregation function that will effectively summarize that distribution. In this work, we introduce kernel mean embeddings as an aggregation function. Next, we define a test statistic between the aggregated values and the individual instances. Our work achieves this by embedding both into a reproducing kernel Hilbert space (RKHS) and then measuring dependence with a relational derivation of the Hilbert-Schmidt independence criterion (HSIC) [18].

3.2.1 Proposed Approach

The Hilbert-Schmidt independence criterion [18] provides a test of dependence between two random variables, X and Y. The test statistic is defined as follows:

$$HSIC(X,Y) = \|P_{XY} - P_X P_Y\|^2_{\mathcal{H}} = \|E[\phi(x) \otimes \psi(y)] - E[\phi(x)]E[\psi(y)]\|^2$$

A biased estimate of HSIC can be obtained as $T = \frac{1}{n^2}\mathbf{tr}\tilde{K_x}\tilde{K_y}$, where K_x and K_y are the kernel matrices of X and Y, $\tilde{K_x} = HK_xH, \tilde{K_y} = HK_yH$, H is a centering matrix, and n is sample size [19]. Consistency is provided by the following theorem:

Theorem 1. *[17] Let k and l be characteristic kernels for the respective RKHSs $\mathcal{F}$ on $\mathcal{X}$ and $\mathcal{G}$ on $\mathcal{Y}$, with feature maps ϕ and ψ, respectively. Define the finite signed measure $\theta := P_{XY} - P_X P_Y$. Then, $C_{YX} = \int_{\mathcal{X}\times\mathcal{Y}} \psi(y) \otimes \phi(x) d\theta(x,y) = 0$ if and only if $\theta = 0$.*

The proof is given by Gretton [17]. The implication of Theorem 1 is that given *marginal* embeddings of P_X and P_Y that are characteristic, the statistic provided by HSIC is consistent, i.e., $HSIC(P_X, P_Y) \to 0$ if and only if $P_X \perp\!\!\!\perp P_Y$ as $n \to \infty$.

Given Lemma 3, HSIC is readily extended to the relational case. We do so with the following definition:

Definition 1 (Relational HSIC)**.** *In the relational setting, the test statistic is defined between a relational variable $C.\tau.X$ and a propositional variable $C.Y$ as:*

$$HSIC(C.\tau.X, C.Y) = \|P_{C.\tau.X,C.Y} - P_{C.\tau.X}P_{C.Y}\|_{\mathcal{H}}^2 = \|E[\mu_x \otimes \psi(y)] - \mathcal{M}_x \otimes \mu_y\|^2$$

Note that the estimated statistic is now given by $T = \frac{1}{n^2}\mathbf{tr}\tilde{\mathbf{K}_x}\tilde{K_y}$, where $\mathbf{K}$ is the two-stage kernel defined in equation 3.1, with each instance defined as the neighbors of node i. While this statistic is very similar to the traditional HSIC measure, it is worth taking a moment to clarify the difference in *what* is being tested. In the relational setting, rather than testing the joint distribution of a set of pairs against a null of the product of their marginals, we assess the relationship between a set of *distributions* and the marginal of a random variable. To demonstrate this, consider the more explicit definition of relational HSIC.

$$HSIC(C.\tau.X, C.Y) = \\ \|\frac{1}{N}\sum_i^N \frac{1}{m_i}\sum_j^{m_1} \phi(x_{i,j}) \otimes \psi(y) - \frac{1}{N}\sum_i^N \frac{1}{m_i}\sum_j^{m_1} \phi(x_{i,j}) \otimes \frac{1}{N}\sum_i^N \psi(y_i)\|_{\mathcal{H}}^2 \tag{3.3}$$

Here we can see that what is being considered is the *average* of the joint embeddings of y and each x which is contained in the set constituting μ_x.

Corollary 3 (Consistency of relational HSIC)**.** *Assume that (1) the kernel $k(C.Y, \cdot)$ is characteristic, (2) the kernel $k'(C.\tau.X, \cdot)$ is characteristic, (3) each $\hat{\mu} \in C.\tau.X$ is close to its population counterpart, (4) the second-level kernel $K(\mu_{C.\tau.X}, \cdot)$ is Gaussian, and (5) the degree of the relational structure is bounded by some constant, d. Then, $HSIC(C.\tau.X, C.Y)$ provides a consistent test of dependence.*

Proof. This is a direct consequence of Theorem 1 and Lemma 3. Lemma 3 is required in order to ensure convergence to the kernel mean, and by extension injectivity. □

While convergence is guaranteed, the asymptotic approximations provided for the propositional version of HSIC are no longer appropriate. However, the null distribution can easily be simulated via permutation.

Outside of the additional considerations necessary to the relational version, HSIC thus far looks very similar to its propositional counterpart. However, as we have shown in the previous section, relational dependence is inherently asymmetric. Because HSIC is the covariance of the data in feature space, simple extension provides that it will be symmetric in the case of regular network structure. We will investigate under which conditions asymmetry occurs in non-regular network structure, and the implications for both direction and dependence testing as part of this thesis.

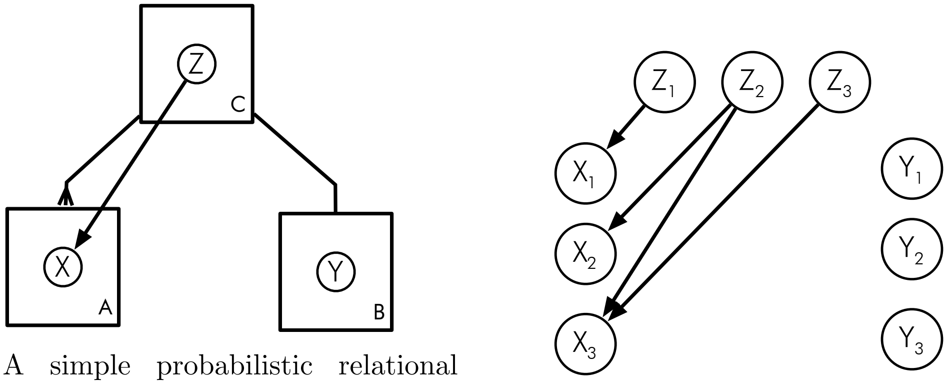
3.2.2 Existing Work Testing Causal Direction and Marginal Dependence of Relational Data

Relevant work to our investigation of methods for determining peer dependence in relational data falls into four basic categories. The most closely related work examines versions of this specific task with alternative methods. For example, Maier, et al. [30], Rattigan [42], and Poole and Crowley [41] provide scenarios in which an asymmetry may arise similar to that observed in our tests for direction. However, in contrast to prior work, we study the phenomenon of asymmetric dependence directly and provide a formal examination which provides guarantees to the circumstances under which this asymmetry can be reliably leveraged. Further, we provide extensive simulation experiments that further show conditions under which direction can be found by considering the difference in dependence in both directions.

A second category of related work focuses on measuring causal dependence in non-relational (i.i.d.) data. For example, Peters, et al. [40] examine the problem of determining the direction of dependence with i.i.d. data by either assuming non-Gaussian noise and linear dependence or non-linear dependence and Gaussian noise. The problem of identifying causal direction in the case of deterministic, i.e. non-noisy data, was studied by Daniusis, et al. [8]. The setting considered was propositional data, and the proposed solution leverages properties of information geometry in order to find asymmetries between the conditional distributions of the two variables. In contrast, the relational setting considered provides a much more direct mechanism for determining direction.

A third thread of related work aims to detect non-causal dependence in relational data. This task has attracted important attention in both statistical relational learning (SRL) community and in multiple areas of the social sciences. In SRL, Jensen and Neville [22] use a χ^2 test to detect auto-correlation in relational data and show its effect for feature selection. Angin and Neville [1] introduce a shrinkage estimator for auto-correlation in the presence of varying dependence strength. However, both of these rely on empirical evaluation as evidence of correctness. Dhurandhar and Dobra [11] and London, et al. [28, 27] provide theoretical analysis for the inductive error of classification and regression in the relational setting.

In the social sciences, relational dependence has been examined under the monikers of peer influence, spillover, and interference. In the experimental setting, Eckles, et al. [13] characterize the threat to validity arising from the bias induced by relational dependence and provide experimental designs to reduce these effects. Manksi [32], VanderWeele [50], and Aronow and Samii [2] examine methods for removing the bias associated with relational dependence, assuming discrete or linearly dependent data. Toulis and Kao [49] provide conditions for experimental design with binary treatments to identify peer influence. Ogburn and VanderWeele [38] characterize relational dependence in terms of graphical models, but do not present an explicit testing procedure. Work studying homophily and contagion (e.g., Christakis and Fowler [4],



(a) A simple probabilistic relational model (PRM) with three attributes, X, Y, Z, and their respective dependencies.

(b) A possible ground-graph arising from an instantiation of the PRM shown in Figure 3.2a. Dependence is induced between X instances via Z.

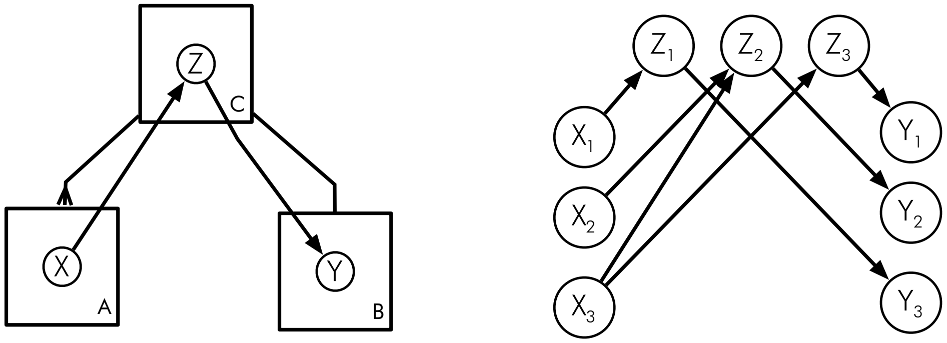
Figure 3.1: An example of induced relational bias that occurs due to auto-dependence.

Lafond and Neville [25] and Rodriguez, et al. [16]) is related but distinct in the task setup, as we do not assume the availability of temporal information.

3.3 Measuring Conditional Dependence in Relational Data

While marginal tests of dependence are fundamental to statistical inference, in most real-world settings the quantity of interest is the dependence between two variables, conditioned on an additional set of variables. Conditional independence test also play a crucial role in learning the structure of causal models, where they provide constraints on the dependence structure of a domain that can be used to find the underlying causal structure [46]. In this work we consider a test between two random variables, X and Y conditioned on a set of variables, $\mathbf{Z}$. We assume that all variables may live on separate entities.

This task is considerably more difficult than the propositional case because of *auto-dependence*—dependence amongst instances whose entity have a connecting path—which reduces the power of the test and can induce a large number of false conclusions of dependence, i.e. Type I errors. As an illustration, consider Figures 3.1a and



(a) A simple probabilistic relational model (PRM) with three attributes, X, Y, Z, and their respective dependencies.

(b) A possible ground-graph arising from an instantiation of the PRM shown in Figure 3.2a. Conditioning on Z may lead to type I errors due to the induced dependence between X values.

Figure 3.2: An example of induced relational bias that occurs *after* conditioning.

3.2a. Figure 3.1a represents marginal auto-dependence as described by Jensen and Neville [23]. Even though X and Y are marginally dependent, because Z induces dependent between X instances the *effective samples size* of the test is decreased, resulting in an increased number of Type I errors. This can be addressed by explicitly conditioning on the values of Z to remove the auto-dependence. Figure 3.2a is a less obvious case. Because all X and Y instances are marginally independent the direct test of dependence, $X \perp\!\!\!\perp Y$ will return the correct answer. However, because there is a collider between Z and instances of X, after conditioning on Z auto-dependence is induced. This case is especially troublesome given that a marginal test of auto-dependence amongst X values, as proposed by Jensen and Neville [23] will fail to detect this threat to validity. Further, it is not immediate that it is possible to remove the auto-dependence by conditioning on additional variables.

3.3.1 Existing Work

In contrast to the case of testing for marginal dependence, there is very little existing work formally addressing testing conditional dependence in relational domains. Prior work on constraint based causal structure learning of relational domains, a task that necessarily involves conditional independence testing, Rattigan, Maier and Jensen [43] developed *relational blocking*, a conditional operator which controls for the effect of variables on a connected entity by using a fixed-effect—the entity identifier. While relational blocking is a promising approach, it is restricted to the setting where X and Y reside on the same entity, $[A]$, and all members of $\mathbf{Z}$ reside on an entity that is connected to $[A]$ via a one-to-many relationship.

Our approach is based on an embedding of the data into an RKHS. Fukumizu, et al. [15] and Zhang, et al.[51] proposed kernel-based conditional independence testing based on a generalized definition of partial correlation (discussed in the next section). Both approaches rely on the i.i.d. assumption. Doran, et al.[12] developed an alternative test of conditional dependence based on a conditional permutation procedure. That work also assumes i.i.d. data. Further, it is not clear that the approach can be extended to non-i.i.d. data because of the difficulty in finding exchangable pairs for permutation that respect both the conditional similarity of $P(X|Z), P(Y|Z)$ and the auto-dependence similarity, i.e. $P(x|\text{neighbors}(x))$.

3.3.2 Proposed Approach

We propose an approach based on the embedding of the data into an RKHS. To do this, we seek to non-trivially extend the work of Zhang, et al. [51] to the relational setting, where each variable instance may consist of a set of observations and there may be auto-dependence amongst independence. To operationalize this into a test, we first must decide on a definition of conditional independence. Here, we use the following characterization, due to Daudin [9]:

Lemma 1. *[9] Let $\mathcal{E}_{YZ}, \mathcal{E}_{XZ}$ be the space of all functions of X, Z and Y, Z respectively. The following conditions are equivalent:*

1. $X \perp\!\!\!\perp Y | Z$

2. $\mathbb{E}(\tilde{f}\tilde{g}) = 0, \forall \tilde{f} \in \mathcal{E}_{XZ}$ *and* $\tilde{g} \in \mathcal{E}_{YZ}$

3. $\mathbb{E}(\tilde{(f)}g) = 0, \forall f \in \mathcal{E}_{XZ}$ *and* $g \in L^2_{XZ}$

4. $\mathbb{E}(\tilde{f}\tilde{g}') = 0, \forall \tilde{f} \in \mathcal{E}_{XZ}$ *and* $\tilde{g}' \in \mathcal{E}'_{YZ}$

5. $(\tilde{f}g') = 0, \forall \tilde{f} \in \mathcal{E}_{XZ}$ *and* $g' \in L^2_Y$

Intuitively, the second condition of Lemma 1 says that independence can be asserted if any function of the residuals of (X, Z) given Z is uncorrelated with (Y, Z) given Z. While this provides a more general condition for partial correlation, it is infeasible in practice, since it requires the ability to reason over *all* functions of X, Z and Y, Z, in L^2. However, if this requirement is relaxed and the space of functions are restricted to be those residing in Hilbert spaces, $\mathcal{H}_{\ddot{\mathcal{X}}}$, $\mathcal{H}_{\mathcal{Y}}$, where $\ddot{X} = (X, Z)$, then the following characterization, due to Fukumizu, et al. [15] provides a definition which can be practically realized:

Lemma 2. *[15] Let* $k_{\ddot{\mathcal{X}}} \triangleq k_{\mathcal{X}\mathcal{Z}}$*. Assuming* $k_{\ddot{\mathcal{X}}}k_{\mathcal{Y}}$ *is characteristic w.r.t.* $(X \times Y) \times Z$*,* $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}},$ *and* $\mathcal{H}_{\mathcal{Z}}$ *are contained in* L^2 *and* $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ *is dense in* $L^2(P_Z)$*, then*

$$\Sigma_{\ddot{X}Y|Z} = 0 \iff X \perp\!\!\!\perp Y|Z \tag{3.4}$$

This implies that conditional dependence can be determined by constructing the spaces resulting from the residuals of a nonparametric regression and testing dependence between the kernel matrices the same way that one determines marginal dependence, i.e. $HSIC_{X,Y|Z} = \frac{1}{n}Tr(\tilde{K}_{\ddot{X}|Z}\tilde{K}_{Y|Z})$.

The method presented thus far has two constraints that prevents its application to relational data: (1) the assumption of individual instances, and (2) the assumption of i.i.d. data. To remedy this we must address the two aspects of the i.i.d. method that are compromised. The first is the consistent estimation of the residual function, the second is a consistent simulation of the null distribution. We will now address both of these individually.

Szabo, et al. [48] studied the problem of distribution to distribution and distribution to real regression in kernel spaces, and showed consistency under the assumption of i.i.d. samples. We will extend these results to the relational setting where instances may exhibit auto-dependence. This involves developing theory that examines both consistency and the statistical efficiency of the estimators with rates that explicitly take into account the auto-dependence between instances.

Effectively simulating from the null distribution in the presence of auto-dependence is significantly more challenging. In the i.i.d. case, the null distribution can be approximated by explicitly simulating the null distribution by repeatedly permuting the values of the variables and calculating the test statistic on the permuted values. However, permutation testing in the presence of auto-dependence will provide an unacceptable number of type I errors, since the null distribution which is estimated represents the null of hypothesis of no dependence between X and Y *and* no auto-dependence amongst either X or Y.

Recently, the dependent wild-bootstrap [45], a non-parametric bootstrap procedure for time series data that operates by explicitly augmenting, or corrupting the data, has been extended to degenerate U and V-statistics [26] and kernel dependence measures (which themselves are degenerate U and V-statistics) [6]. We will extend the dependent wild-bootstrap to the more general non-i.i.d. setting of relational data. To achieve this, we will revisit the results of Shao [45], and consider the conditions under which the dependent wild-bootstrap provides a consistent estimate under alternate notions of dependence between instances. The current definition relies on a notion of weak dependence which is specific to time-series data and the resulting analysis relies heavily on this assumption. Finally, we note that while the explicit aim of this work is to define a test of conditional dependence for relational data, the tests defined here will be trivially extensible to other non-i.i.d. data such as time-series, spatial and spatial-temporal data.

3.4 Improved Structure Learning for Relational Domains

Building on recent advances in defining *d*-separation semantics for relational data [31], Maier, et al. [30] and Marazopoulou, et al. [34] define a constraint based algorithms for learning the causal structure of relational domains and temporal-relational domains, respectively. Both algorithms are extensions to the PC algorithm [46], a constraint-based method for learning the causal structure of i.i.d. data. PC, and all constraint-based algorithms follow the following template for learning the causal structure of an observational dataset, $\mathcal{X}$:

1. Identify marginal independencies between variables.
2. For separating set size $1 \dots |\mathcal{X}|$, find the minimum separating set Z for each $x, y \in \mathcal{X}$ such that $x \perp\!\!\!\perp y | Z$, if it exists.
3. Orient the dependencies according to some set of pre-defined semantics.

Steps 1 and 2 are referred to as the "skeleton discovery" phase, step three is commonly referred to as the "orientation" phase.

PC has been proven to be sound, i.e. the algorithm will never conclude the presence of a causal dependence that does not exist, and complete, i.e. the algorithm will orient the maximum number of causal dependencies correctly, in the case of i.i.d.-data [35]. Both Maier, et al. [30] and Marazopoulou, et al. [34] claim a sound and complete procedure for learning the causal structure of relational data. Given our results showing the inherent directionality of relational dependence, the algorithms as presented are neither sound or complete. Both follow the constraint-based learning procedure outlined earlier, and conclude independence between variables if a test, with randomly chosen perspective, determines independence. However, as we have seen, this approach is insufficient. If both directions are not *explicitly* tested then it is likely that independence will be falsely concluded, i.e., a type II error. The result is an algorithm that may regularly exclude dependencies that exist in the true causal structure, i.e. an unsound algorithm. We propose to remedy this by defining an algorithm that explicitly takes into account the direction of dependence in the skeleton discovery phase. In addition to defining a sound and complete algorithm, this is also likely to result in efficiency gains. Given an orientation of dependence, it may be possible to exclude certain conditional dependence tests.

Finally, to date, real-world experiments of causal-structure learning for relational domains have been mostly a demonstration of runtime tractability since none of the employed marginal or conditional tests provided theoretical guarantees on correctness. As the final contribution of this thesis, we will perform a comprehensive evaluation of relational causal discovery on real-world domains and compare the resulting causal inferrences to those learned using simpler procedures such as relational dependency networks [37] and non-causal learning of probabilistic relational models which is performed by searching over the space of structures with respect to likelihood [14]. We believe that this will help both to confirm the efficacy of relational causal learning algorithms, as well as help to identify directions for future work.

CHAPTER 4
PROPOSED TIMELINE

Below is a proposed timeline. Dates given indicate the month when the work is expected to be completed. It refers to the end of the month, unless otherwise indicated.

- Finalize theoretical contributions for conditional dependence testing and run experiments (April 2016).
- Write marginal and conditional dependence section, including experimental and theoretical results (May 2016).
- Finalized definition joint structure learning algorithm that incorporates findings of relational dependence testing, with experimental results (July 2016).
- Finish writing (August 2016).
- First draft of thesis given to committee (early September 2016).
- Incorporate initial committee edits, distribute to committee, and defend (October 2016).

APPENDIX A

SYNTHETIC EXPERIMENTS

Our theoretical results focus on regular graphs, linear dependence, and absence of noise. In this section, we examine the effect that the network structure, the functional form of the dependence, and the presence of noise have on the efficacy of the linear and kernel based methods.

A.0.1 Regular Networks and Linear Dependence

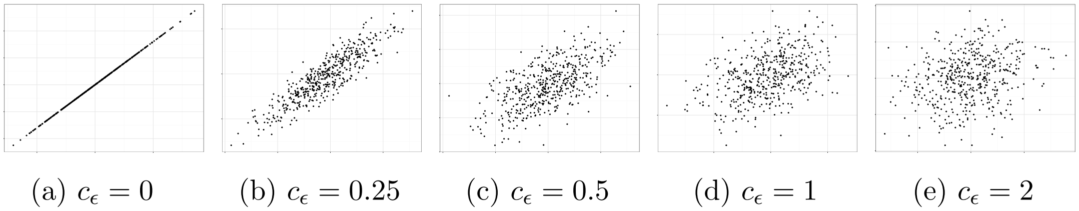
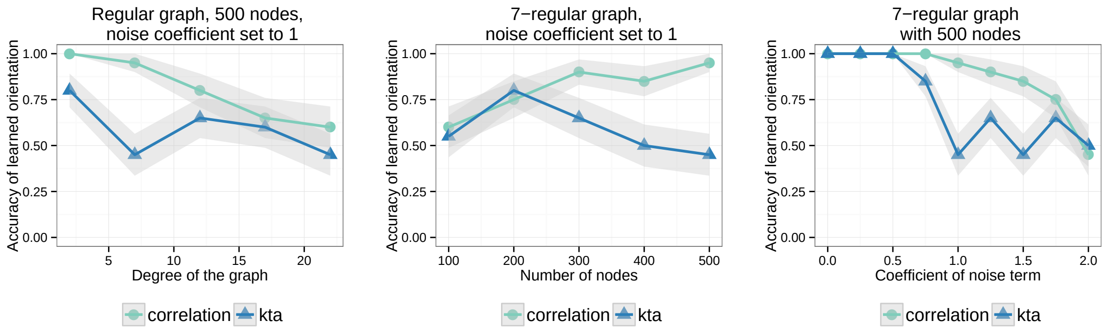

(a) $c_\epsilon = 0$ (b) $c_\epsilon = 0.25$ (c) $c_\epsilon = 0.5$ (d) $c_\epsilon = 1$ (e) $c_\epsilon = 2$

Figure A.1: Scatterplots for the sum of X values of related nodes (x-axis) vs. the sum of X values of related nodes with additive Gaussian noise (y-axis). The noise coefficient (c_ϵ) varies from 0 to 2. The underlying network structure is a regular network of degree 10 with 500 nodes.

We first considered regular graphs with linear dependence—a setting that matches our theoretical analysis—and we examined the effect of noise. We considered networks with the total number of nodes ranging from 100 to 500 and varied the degree between 2 and 22 by increments of 5. For every graph structure, we generated data as follows:



(a) (b) (c)

Figure A.2: Orientation accuracy for regular graphs for varying degree (A.2a), size of network (A.2b), and noise coefficient (A.2c).

$$\mathbf{x} \sim \mathcal{N}(0,1)$$

$$\epsilon \sim \mathcal{N}(0,1)$$

$$\mathbf{y} \sim D^{-1}A\mathbf{x} + \beta\epsilon$$

where β is the coefficient of the noise and was varied between 0 and 2.

Figure A.1 shows the relationship between $D^{-1}A\mathbf{x}$ and $\mathbf{y}$ for varying values of β. In the noiseless case (Figure A.1a), $D^{-1}A\mathbf{x}$ and $\mathbf{y}$ are perfectly linearly correlated, as expected from the generating process. However, as the noise increases, the correlation between $D^{-1}A\mathbf{x}$ and $\mathbf{y}$ decays very quickly, approaching an adversarial case by the time the noise coefficient is $\beta = 1.0$.

We then measured dependence in each direction ($\mathbf{x}$ and $A\mathbf{y}$, $\mathbf{y}$ and $A\mathbf{x}$). The direction that produced the higher value for dependence was recorded as the inferred causal direction. To measure dependence, we used

1. the square of Pearson's correlation, and

2. KTA using RBF kernels with a fixed bandwidth of 1.0 for all kernel calculations.

Figure A.2c shows the accuracy of both methods for a graph with 500 nodes and degree 7, while varying β. As expected from the our earlier theoretical results, both methods perform perfectly in the noise-less case, and continue to do so through $\beta = 0.5$. The linear method is significantly more robust to noise, remaining nearly perfect until $\beta = 1.0$.

We also examined the interplay between the graph structure (degree and number of nodes) and and the performance of each method. Figure A.2a shows the performance for the case of a 500-node graph with noise coefficient of 1.0 with the degree varied between 2 and 22. Both methods become systematically worse as the degree (and thus the density of the network) increases. This is expected behaviour since an increase in the degree results in a lower *effective sample size* [22], which will reduce the expected efficacy of both methods. The converse of this effect can be seen in Figure A.2b, where the accuracy of the linear based approach improves significantly as the size of the network increases while the degree is kept constant (and thus the density of the network decreases).

A.0.2 Non-Regular Networks

We next compared the performance of both methods to a departure from the assumption of network regularity. We considered the three most common generative models of graphs. The Erdős-Rényi model creates networks where two nodes are connected with a given probability. Throughout the experiments, we considered a fixed connection probability 0.2. The Watts-Strogatz model generates “small-world networks”. It begins with a lattice with a given neighborhood size and randomly rewiring edges according to a probability fixed across edges. For our experiments, we used neighborhood size 5 and rewiring probability equal to 0.2. The final generative model we considered was the Barabási-Albert model. This model generates graphs that display preferential attachment. For our experiments the power of preferential attachment was set to 1.0. For each network we considered sizes between 100 and 1000, by increments of 100, with 20 graphs being drawn for each size.

We then considered the following data generation scenarios for all graph types:

$$\begin{aligned} \mathbf{x} &\sim \mathcal{N}(0,1) \\ \epsilon &\sim \mathcal{N}(0,1) \\ \mathbf{y} &\sim f(D^{-1}A\mathbf{x}) + \beta\epsilon \end{aligned}$$

where $f(\cdot)$ is a function of $D^{-1}A\mathbf{x}$. We considered three functional forms:

- $f(\cdot)$ is a simple linear function (linear)
- $f(D^{-1}A\mathbf{x}) = \tan(D^{-1}A\mathbf{x})$ (nonlinear)
- $f(D^{-1}A\mathbf{x}) = \left(D^{-1}A\mathbf{x}\right)^4$ (quad)

For each setting, β was varied between 0 and 2 by increments of 0.25.

The performance of both the linear and KTA method for fixed network size of 1000 nodes with the magnitude of noise varied is shown in Figure A.4. For the Barabási model under linear dependence, both the linear and kernel methods appear to be very robust up until a noise coefficient of 2.0. The KTA based method generally outperforms the linear dependence method for non-linear dependencies. This is to be expected, as Pearson's correlation is a measure of linear dependence.

The performance in the case where β is held to 0.5 and the size of the network is varied from 100 to 1000 can be seen in Figure A.3. Here we can see that in both the Barabási-Albert and Watts-Strogatz graph models, Pearson's correlation and KTA achieve better performance under linear dependence as the size of the network increases. However, for in the case of the Erdős-Rényi models both methods perform poorly consistently as the size of the network increases. This is due to the nature of the graph-generation process. Both the Barabási-Albert and Watts-Strogatz

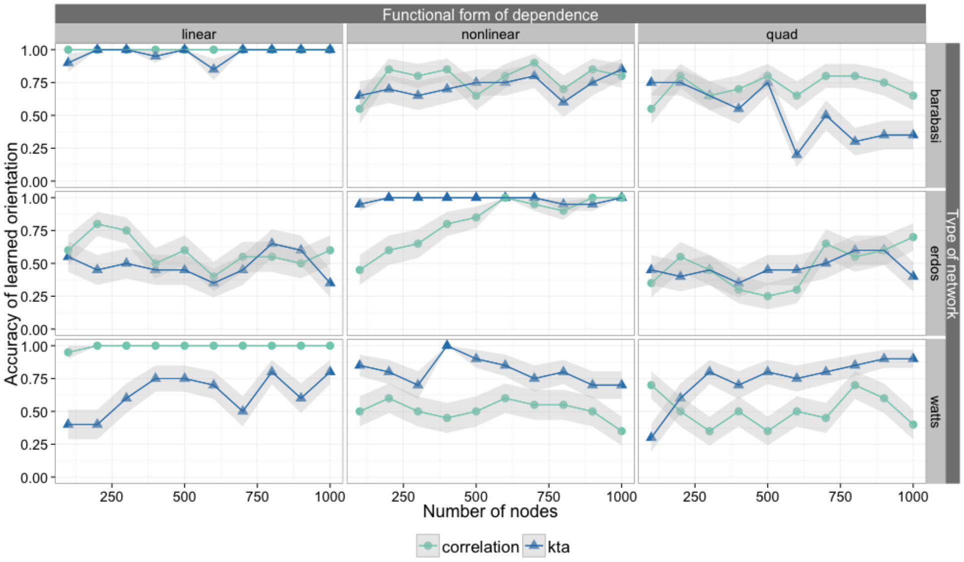


Figure A.3: Orientation accuracy for various network types and functional forms, as the size of the graph increases. The noise coefficient is set to 0.5.

models become increasingly sparse as the size of the network is increased. However, in the case of Erdős-Rényi, the probability connection is constant. As a result, the effective sample size remains low when the number of nodes increases. This likely accounts for the poor performance of the linear estimator. The opposite effect is seen in the case of the Barabási-Albert model. In nearly all cases the performance of the estimators is highest in the case of the Barabási-Albert networks.

A.0.3 A Comparison to Relational Bivariate Edge Orientation

We also compared our results to the relational bivariate edge orientation (RBO)[30], the only other known method for testing causal direction in relational data. Maier, et al. [30] introduced the relational bivariate edge orientation (RBO) as an edge-orientation procedure within the context of learning causal models of relational domains. RBO is defined with respect to conditional independence properties of relational models. Specifically, rephrasing the definition of Maier, et al. [30] for single-entity single-relationship networks, for a relational dependence between Y' and X, RBO checks if Y' is in the separating set of X and X'. If not, then Y' is effectively a "relational" collider and is oriented as such: $Y' \leftarrow X$. Otherwise, the only alternative model is $Y' \rightarrow X$, given that dependencies that induce feedback cycles (such as $X \rightarrow X'$) are excluded by assumption. The correctness of RBO is defined with respect to a conditional dependence oracle. In practice, Maier, et al. [30] follow the following procedure to infer causal direction between two relational variables:

1. Learn a linear model $\mathbf{x} \sim D^{-1}A\mathbf{x} + D^{-1}A\mathbf{y}$ to determine if $\mathbf{x} \perp\!\!\!\perp D^{-1}A\mathbf{x} \mid D^{-1}A\mathbf{y}$
2. If $\mathbf{x} \not\perp\!\!\!\perp D^{-1}A\mathbf{x} \mid D^{-1}A\mathbf{y}$, then return $D^{-1}A\mathbf{x} \rightarrow \mathbf{y}$, otherwise return $D^{-1}A\mathbf{y} \rightarrow \mathbf{x}$

We applied this procedure to the linear data-generating scenarios used in the previous two subsections, with one modification. Rather than testing a single perspective, we explicitly tested the conditional independence facts from the perspective of both $\mathbf{x}$ and $\mathbf{y}$. We found that between all scenarios, RBO failed to induce dependence in 80-90% of cases. This has important ramifications for the RCD algorithm of Maier, et al. [30]. As currently implemented, the RBO rule would have produced $\sim$ %50 error rate, since it does not explicitly check both directions. Using our more conservative method, RBO would fire a fraction as often. In contrast, by incorporating the findings of the more direct marginal comparison presented here, vast numbers of edges would be accurately oriented. We plan on examining further integration of our findings into joint causal structure learning algorithms in future work.

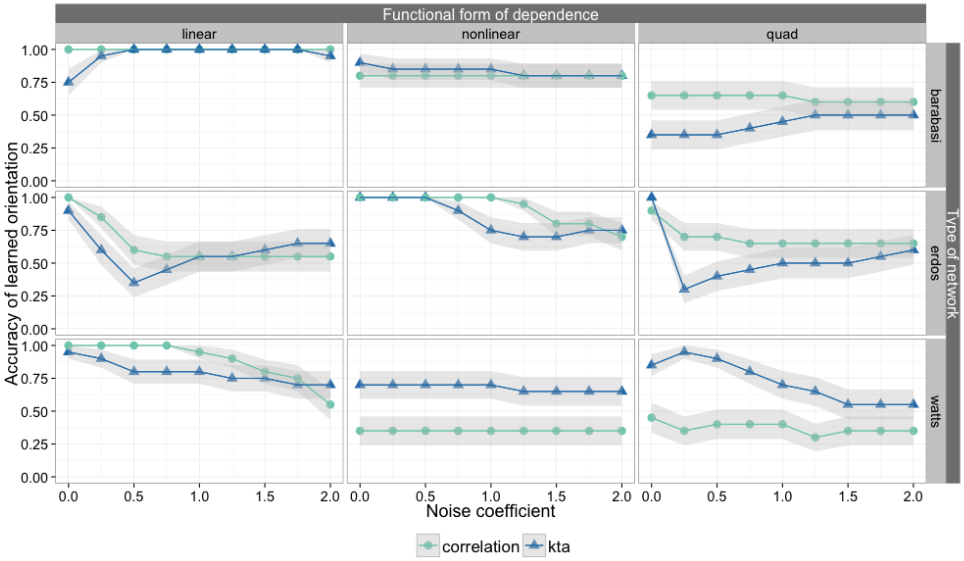


Figure A.4: Orientation accuracy for various network types and functional forms, as the coefficient of the noise increases. The network size was kept constant at 1000 nodes.

APPENDIX B

PROOFS

B.1 Proof of Proposition 2

Proposition 2. *Assume that the true generative process is* $\mathbf{y} = \beta \cdot D^{-1}A\mathbf{x} + \epsilon$ *for some constant* β*, where* ϵ *is a vector with the noise terms. Moreover, assume that assumptions A1-A5 hold and* X *and* Y *are scaled to mean 0. Then the following holds:*

$$\rho^2(\mathbf{x}', \mathbf{y}) > \rho^2(\mathbf{y}', \mathbf{x}) \Leftrightarrow$$
$$\frac{Var(AD^{-1}A\mathbf{x}) + Var(A\epsilon)}{Var(D^{-1}A\mathbf{x}) + Var(\epsilon)} > \frac{Var(A\mathbf{x})}{Var(\mathbf{x})}.$$

Proof.

$$\begin{aligned}
\rho(\mathbf{x}', \mathbf{y}) =& \rho(A\mathbf{x}, D^{-1}A\mathbf{x} + \epsilon) && \text{(B.1)} \\
=& \frac{Cov(A\mathbf{x}, D^{-1}A\mathbf{x}) + Cov(A\mathbf{x}, \epsilon)}{Var(A\mathbf{x})\big(Var(D^{-1}A\mathbf{x}) + Var(\epsilon)\big)} \\
=& \frac{Cov(A\mathbf{x}, D^{-1}A\mathbf{x})}{Var(A\mathbf{x})\big(Var(D^{-1}A\mathbf{x}) + Var(\epsilon)\big)} && \text{(B.2)}
\end{aligned}$$

$$\begin{aligned}
\rho(\mathbf{y}', \mathbf{x}) =& \rho(AD^{-1}A\mathbf{x} + D^{-1}A\epsilon, \mathbf{x}) && \text{(B.3)} \\
=& \frac{Cov(AD^{-1}A\mathbf{x}, \mathbf{x}) + Cov(\mathbf{x}, D^{-1}A\epsilon)}{Var(\mathbf{x})\big(Var(AD^{-1}A\mathbf{x}) + Var(D^{-1}A\epsilon)\big)} \\
=& \frac{Cov(AD^{-1}A\mathbf{x}, \mathbf{x})}{Var(\mathbf{x})\big(Var(AD^{-1}A\mathbf{x}) + Var(D^{-1}A\epsilon)\big)} && \text{(B.4)}
\end{aligned}$$

The covariance, given that the mean of X and Y is 0, is equal to the inner product of the variables.

$$\begin{aligned} Cov(A\mathbf{x}, D^{-1}A\mathbf{x}) &= \langle A\mathbf{x}, D^{-1}A\mathbf{x}\rangle \tag{B.5}\\ &= \mathbf{x}^{\top}A^{\top}D^{-1}A\mathbf{x} \tag{B.6}\\ &= \mathbf{x}^{\top}AD^{-1}A\mathbf{x} \tag{B.7}\\ Cov(AD^{-1}A\mathbf{x}, \mathbf{x}) &= \langle AD^{-1}A\mathbf{x}, \mathbf{x}\rangle \tag{B.8}\\ &= \mathbf{x}^{\top}AD^{-1}A\mathbf{x} \tag{B.9} \end{aligned}$$

□

We will make use of the following definitions in order to provide guarantees regarding the convergence of the kernel mean.

Definition 2. *(Total Variation Distance) The total variation distance for two probability distributions, P and Q, on σ-algebra Σ over sample space ω is defined as:*

$$\|P - Q\|_{TV} \equiv \sup_{A\in\Sigma} |P(A) - Q(A)|.$$

Definition 3. *(Dependency Matrix) Assuming a fixed ordering π of $\mathcal{X} \equiv \{\mathcal{X}_i\}_{i=1}^{n}$, we can define a measure of dependence with respect to this ordering. Specifically, we define the upper triangular dependency matrix, $\Theta_n^{\pi} \in \mathbb{R}^{n\times n}$, where*

$$\theta_{i,j} = \begin{cases} 1 & \textit{if } i = j \\ \eta_{i,j} & \textit{if } i < j \\ 0 & \textit{otherwise} \end{cases}$$

$\eta_{i,j} \equiv \sup \|P(\mathcal{X}_{j:n}|\mathbf{x}_1 : 1, x_i) - P(\mathcal{X}_{j:n}|\mathbf{x}_{1:i-1}, x_i')\|_{TV}$.

Note that this ordering is *not* necessarily a temporal ordering. The dependency matrix Θ can be viewed as a covariance matrix between each variable in the network. It will be used in order to bound the bias associated with our test.

Theorem 2. *(Generalized McDiarmid's Inequality)[27] Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a measurable function with a constant c such that $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$ that differ only at a single coordinate, $|f(\mathbf{x}) - f(\mathbf{x}')| \leq \frac{c}{n}$. Then for all $\epsilon > 0$:*

$$P\{f(\mathcal{X}) - \mathbb{E}[f(\mathcal{X})] \geq \epsilon\} \leq \exp\left(\frac{-2n\epsilon^2}{c^2 \|\Theta_n^\pi\|_\infty^2}\right)$$

We refer the reader to London et al. [27] for the proof.

B.2 Proof of the convergence rate of the empirical mean embedding under weak dependence

Lemma 3. *Under the assumptions that each kernel mean, $\hat{\mu}$, is close to their population values, and the degree of the network is bounded by some constant, d, and the random variable that gives rise to μ is independently distributed, the estimate $\hat{\mathcal{M}} = \frac{1}{N}\sum_i^N \hat{\mu}_i$ is a consistent estimator of the true embedded mean, $\mathcal{M}$.*

Let the mean of second-level mean embedding of $\mu \in \mathcal{M}(\Omega)$ into the RKHS provided by the kernel, $k\cdot$). We will assume $k(\cdot)$ is bounded by $k(\mu, \mu) \leq B_k (\forall \mu \in \Omega)$. We are given N samples, $mu_1, \ldots, \mu_N$, from a weakly dependent process, whose covariance matrix is given by Θ. Further, define the empirical mean embedding as $\mu_{\hat{\mu}} = \frac{1}{N}\sum_{n=1}^N k(\cdot, \mu_n)$. Then $\mathbb{P}(\|\mathcal{M}_{\hat{\mu}} - \mathcal{M}_\mu\|_H \geq \epsilon) \leq e^{-\frac{\epsilon^2 N}{2B_k\|\Theta\|_\infty}}$

Proof. Let $\phi(\mu) = k(\cdot, \mu)$, and by extension $k(\mu, \mu) = \|\phi(\mu)\|_H^2$. Let $g(S) = \|\mathcal{M}_{\hat{\mu}} - \mathcal{M}_\mu\|_H = \|\frac{1}{N}\sum_{n=1}^N \phi(\mu_n) - \mathcal{M}_x)\|_H$, with S being the set of samples, i.e., $S = \{\mu_1 \ldots mu_N\}$. Also let $S' = \{\mu_1, \ldots, \mu_{j-1}, \mu_j', x_{j+1}, \ldots, \mu_N\}$. We have

$$\begin{aligned}
|g(S) - g(S')| = \Big|\|\frac{1}{N}\sum_{i=1}^N \phi(mu_n) - \mathcal{M}_\mu\|_H - \|\frac{1}{N}\sum_{i=1}^N \phi(\mu_n) + \frac{1}{N}\phi(\mu_j') - \mathcal{M}_\mu\|_H\Big| \\
\leq \frac{1}{N}\|\phi(\mu_j) - \phi(\mu_j')\|_H \leq \frac{1}{N}(\|\phi(\mu_j)\|_H) + \|\phi(\mu_j')\|_H) \\
\leq \frac{1}{N}\Big[\sqrt{k(\mu_j, \mu_j)} + \sqrt{k(\mu_j', \mu_j')}\ \Big] \leq \frac{2\sqrt{B_k}}{N}
\end{aligned}$$

We can now use the generalized version of McDiarmid's inequality, yielding

$$\begin{aligned}
\mathbb{P}(g(S) - \mathbb{E}[g(S)] \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{n=1}^N \left(\frac{2\sqrt{B_k}}{N}\right)\|\Theta\|_\infty}\right) \\
= \exp\left(-\frac{2\epsilon^2}{N\frac{4B_k}{N^2}\|\Theta\|_\infty}\right) \\
= \exp\left(-\frac{\epsilon^2 N}{2B_k\|\Theta\|_\infty}\right)
\end{aligned}$$

Where $\|\Theta\|_\infty$ is the L_∞ norm of the covariance matrix between nodes as described earlier. We see now that convergence is governed not only by the maximum value of the kernel, but also the dependence amongst instances. Because we have assumed that the data from which each μ has arisen is i.i.d. the only source of bias is due to the overlap in nodes. This implies that if the degree is bounded by some finite constant, d as the network grows to infinity, there will be a maximum value of $\|\Theta\|_\infty < \infty$, which implies convergence. □

BIBLIOGRAPHY

[1] Angin, Pelin, and Neville, Jennifer. A shrinkage approach for modeling nonstationary relational autocorrelation. In *Eighth IEEE International Conference on Data Mining* (2008), IEEE, pp. 707–712.

[2] Aronow, Peter M., and Samii, Cyrus. Estimating average causal effects under interference between units. *arXiv preprint arXiv:1305.6156* (2013).

[3] Bakshy, Eytan, Eckles, Dean, Yan, Rong, and Rosenn, Itamar. Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (2012), ACM, pp. 146–161.

[4] Christakis, Nicholas, and Fowler, James. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives.* Hachette Digital, Inc., 2009.

[5] Christmann, Andreas, and Steinwart, Ingo. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems* (2010), pp. 406–414.

[6] Chwialkowski, Kacper P., Sejdinovic, Dino, and Gretton, Arthur. A Wild Bootstrap for Degenerate Kernel Tests. In *Advances in Neural Information Processing Systems* (2014), pp. 3608–3616.

[7] Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research 13*, 1 (2012), 795–828.

[8] Daniusis, P, Janzing, D, Mooij, J, Zscheischler, J, Steudel, B, Zhang, K, Schölkopf, B, Spirtes, Grünwald P, et al. Inferring deterministic causal relations. In *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)* (2010), AUAI Press, pp. 143–150.

[9] Daudin, JJ. Partial association measures and an application to qualitative regression. *Biometrika 67*, 3 (1980), 581–590.

[10] Dawid, A Philip. Causal inference without counterfactuals. *Journal of the American Statistical Association 95*, 450 (2000), 407–424.

[11] Dhurandhar, Amit, and Dobra, Alin. Distribution-free bounds for relational classification. *Knowledge and Information Systems 31*, 1 (2012), 55–78.

[12] Doran, G, Muandet, K, Zhang, K, and Schölkopf, B. A permutation-based kernel conditional independence test. In *30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)* (2014), AUAI Press, pp. 132–141.

[13] Eckles, Dean, Karrer, Brian, and Ugander, Johan. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530* (2014).

[14] Friedman, Nir, Getoor, Lise, Koller, Daphne, and Pfeffer, Avi. Learning probabilistic relational models.

[15] Fukumizu, Kenji, Gretton, Arthur, Sun, Xiaohai, and Schölkopf, Bernhard. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* (2007), pp. 489–496.

[16] Gomez-rodriguez, Manuel, Leskovec, Jure, et al. Modeling Information Propagation with Survival Theory. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013), pp. 666–674.

[17] Gretton, A. A simpler condition for consistency of a kernel independence test. *ArXiv e-prints* (Jan. 2015).

[18] Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., and Smola, A.J. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* (2008), pp. 585–592.

[19] Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schoelkopf, B. Kernel methods for measuring independence. *Journal of Machine Learning Research 6* (2005), 2075–2129.

[20] Gretton, Arthur, and Györfi, László. Consistent nonparametric tests of independence. *The Journal of Machine Learning Research 11* (2010), 1391–1423.

[21] Janzing, Dominik, Mooij, Joris, Zhang, Kun, Lemeire, Jan, Zscheischler, Jakob, Daniušis, Povilas, Steudel, Bastian, and Schölkopf, Bernhard. Information-geometric approach to inferring causal directions. *Artificial Intelligence 182* (2012), 1–31.

[22] Jensen, David, and Neville, Jennifer. Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning. In *Machine Learning, Proceedings of the Nineteenth International Conference* (2002), pp. 259–266.

[23] Jensen, David, and Neville, Jennifer. Autocorrelation and linkage cause bias in evaluation of relational learners. In *Inductive Logic Programming*. Springer, 2003, pp. 101–116.

[24] Koller, Daphne. Probabilistic relational models. In *Inductive logic programming*. Springer, 1999, pp. 3–13.

[25] La Fond, Timothy, and Neville, Jennifer. Randomization tests for distinguishing Social Influence and Homophily Effects. In *Proceedings of the 19th international conference on World wide web* (2010), ACM, pp. 601–610.

[26] Leucht, Anne, and Neumann, Michael H. Dependent Wild Bootstrap for Degenerate U-and V-statistics. *Journal of Multivariate Analysis 117* (2013), 257–280.

[27] London, Ben, Huang, Bert, Taskar, Benjamin, and Getoor, Lise. Collective Stability in Structured Prediction: Generalization from One Example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (2013).

[28] London, Ben, Huang, Bert, Taskar, Benjamin, and Getoor, Lise. PAC-Bayesian Collective Stability. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics* (2014).

[29] Lopez-Paz, D., Muandet, K., and Recht, B. The randomized causation coefficient. *Journal of Machine Learning* (2015).

[30] Maier, Marc, Marazopoulou, Katerina, Arbour, David, and Jensen, David. A Sound and Complete Algorithm for Learning Causal Models from Relational Data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence* (2013), pp. 371–380.

[31] Maier, Marc, Marazopoulou, Katerina, and Jensen, David. Reasoning About Independence in Probabilistic Models of Relational Data. *arXiv preprint arXiv:1302.4381* (2013).

[32] Manski, Charles F. Identification of Treatment Response With Social Interactions. *The Econometrics Journal 16*, 1 (2013), S1–S23.

[33] Marazopoulou, Katerina, Arbour, David, and Jensen, David. Refining the Semantics of Social Influence. *Networks: From Graphs to Rich Data. NIPS Workshop* (2014).

[34] Marazopoulou, Katerina, Maier, Marc, and Jensen, David. Learning the Structure of Causal Models with Relational and Temporal Dependence. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence* (2015).

[35] Meek, Christopher. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (1995), Morgan Kaufmann Publishers Inc., pp. 403–410.

[36] Muchnik, Lev, Aral, Sinan, and Taylor, Sean J. Social influence bias: A randomized experiment. *Science 341*, 6146 (2013), 647–651.

[37] Neville, Jennifer, and Jensen, David. Relational Dependency Networks. *The Journal of Machine Learning Research 8* (2007), 653–692.

[38] Ogburn, Elizabeth L., VanderWeele, Tyler J., et al. Causal Diagrams for Interference. *Statistical Science 29*, 4 (2014), 559–578.

[39] Pearl, Judea. *Causality.* Cambridge University Press, 2009.

[40] Peters, Jonas, Mooij, Joris M, Janzing, Dominik, and Schölkopf, Bernhard. Causal Discovery with Continuous Additive Noise Models. *The Journal of Machine Learning Research 15*, 1 (2014), 2009–2053.

[41] Poole, David, and Crowley, Mark. Cyclic Causal Models with Discrete Variables: Markov Chain Equilibrium Semantics and Sample Ordering. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (2013), AAAI Press, pp. 1060–1068.

[42] Rattigan, Matthew JH. *Leveraging Relational Representations for Causal Discovery.* PhD thesis, University OF Massachusetts Amherst, 2012.

[43] Rattigan, Matthew J.H., Maier, Marc, and Jensen, David. Relational Blocking for Causal Discovery. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (2011), pp. 145–151.

[44] Rubin, Donald B. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association* (2011).

[45] Shao, Xiaofeng. The Dependent Wild Bootstrap. *Journal of the American Statistical Association 105*, 489 (2010), 218–235.

[46] Spirtes, Peter, Glymour, Clark N, and Scheines, Richard. *Causation, Prediction, and Search*, vol. 81. MIT press, 2000.

[47] Stegle, Oliver, Janzing, Dominik, Zhang, Kun, Mooij, Joris M, and Schölkopf, Bernhard. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems* (2010), pp. 1687–1695.

[48] Szabo, Zoltan, Gretton, Arthur, Poczos, Barnabas, and Sriperumbudur, Bharath. Two-Stage Sampled Learning Theory on Distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (2015), pp. 948–957.

[49] Toulis, Panos, and Kao, Edward. Estimation of Causal Peer Influence Effects. In *Proceedings of The 30th International Conference on Machine Learning* (2013), pp. 1489–1497.

[50] VanderWeele, Tyler J. Ignorability and Stability Assumptions in Neighborhood Effects Research. *Statistics in Medicine 27*, 11 (2008), 1934–1943.

[51] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-Based Conditional Independence Test and Application in Causal Discovery. AUAI Press, pp. 804–813.