Inferring Causal Direction from Relational Data

David Arbour
darbour@cs.umass.edu

Katerina Marazopoulou
kmarazo@cs.umass.edu

David Jensen
jensen@cs.umass.edu

College of Information and Computer Sciences
University of Massachusetts Amherst
Amherst MA, 01003

Abstract

Inferring the direction of causal dependence from observational data is a fundamental problem in many scientific fields. Significant progress has been made in inferring causal direction from data that are independent and identically distributed (i.i.d.), but little is understood about this problem in the more general relational setting with multiple types of interacting entities. This work examines the task of inferring the *causal direction of peer dependence* in relational data. We show that, in contrast to the i.i.d. setting, the direction of peer dependence can be inferred using simple procedures, regardless of the form of the underlying distribution, and we provide a theoretical characterization on the identifiability of direction. We then examine the conditions under which the presence of confounding can be detected. Finally, we demonstrate the efficacy of the proposed methods with synthetic experiments, and we provide an application on real-world data.[1]

1 INTRODUCTION

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.) [Janzing et al., 2012, Stegle et al., 2010, Lopez-Paz et al., 2015]. For relational data, recent work has studied the problem of inferring the effects of peers via *experimentation* [Muchnik et al., 2013, Bakshy et al., 2012, Toulis and Kao, 2013]. 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 from 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 a study on 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).

This paper examines 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.
4. 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.

The rest of the paper is structured as follows. Section 2 describes the problem setting. Section 3 presents a test of causal direction under deterministic linear dependence.

[1] A full version of this paper including supplementary material can be found at `http://kdl.cs.umass.edu/papers/arbour-et-al-uai2016.pdf`

Section 4 considers a relaxation of the assumptions by allowing for latent confounding and discusses the conditions under which latent confounding can be identified. Section 5 generalizes these results to the case where the similarity is measured by embedding the data in a reproducing kernel Hilbert space (RKHS). Section 6 presents experimental evaluation of these results using synthetic data and a variety of marginal and conditional distributions, as well as networks generated from the Erdős-Rényi, Watts-Strogatz, and Barabási-Albert models. Section 7 presents a demonstration of our method on Stack Overflow, a large online community where users ask and answer computer science related questions.

2 PROBLEM SETTING

Relational domains consist of multiple types of entities that interact with each other through multiple types of relationships. Consider, for example, the domain of academic publishing: authors write papers, papers cite other papers and so on. In this work, for clarity of exposition and without loss of generality, we focus on *networks*, a specific type of relational domains with a single type of entity (e.g., people) and a single type of relationship (e.g., friendship)[2].

An instantiation of a network consists of a set of people and a set of friendships among these people. This can be represented with an undirected graph $G = \langle V, E \rangle$ with n vertices. Nodes correspond to people and an edge denotes friendship between the nodes it connects. Every node of the graph $v_i \in V$ is associated with a pair of random variables, X_i and Y_i. These correspond to attributes of a person, for example wealth and spending habits. 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 variables 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 [Koller, 1999]. 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 $X' \to Y$ or $Y' \to X$ 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 note that A is a symmetric matrix since G is an undirected graph. We can write the vector of the sum of the friends (i.e., the vector $\mathbf{x}'$) as $\mathbf{x}' = A\mathbf{x}$. Similarly, $\mathbf{y}' = A\mathbf{y}$.

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

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

2.1 UNDERLYING ASSUMPTIONS

Throughout the paper, we make the following assumptions:

A1. *G is an undirected graph.*

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

A3. *The distribution of X_i and Y_i is the same for all $v_i \in V$ (templating).*

A4. *There are no feedback cycles, i.e. $Y \to X \Rightarrow X \not\to Y$ for any two (relational or propositional) variables.*

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

A5. *There are no confounding variables, i.e., unobserved variables that are common causes of the observed attributes.*

Section 4 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 [Stegle et al., 2010, Janzing et al., 2012, Lopez-Paz et al., 2015].

[2]The extension to the more general multi-entity/multi-relationship case is straightforward. We provide the necessary details for this extension in the supplement.

3 DIRECTION UNDER LINEAR DEPENDENCE

In this section we show that, under the assumptions of linearity and a small amount of noise, 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 5.

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

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

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

[3]A graph is d-regular if every vertex has degree d.

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 $\mathbf{y} = \beta D^{-1}A\mathbf{x} + \epsilon$ 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 supplement. The implication of proposition 2 is that the causal direction can be accurately inferred, as long as the relative influence of the noise distribution is small in comparison to the relationship between $AD^{-1}\mathbf{x}$ and $\mathbf{y}$. As we show during our experimental evaluation in Section 6, the method is quite robust to the effect of noise in practice.

4 REASONING ABOUT CONFOUNDING

Throughout Section 2 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., $A\mathbf{z} \rightarrow \mathbf{x}$ and $A\mathbf{z} \rightarrow \mathbf{y}$.
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., $\mathbf{z} \rightarrow \mathbf{x}$ and $A\mathbf{z} \rightarrow \mathbf{y}$.

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.*

The proof is deferred to the supplement. 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 $A\mathbf{x} \to \mathbf{y}$ 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 substituted $\mathbf{z}$ for the $\mathbf{x}$. It follows that, in the no-noise setting, an incorrect determination of direct causation will be made. □

Note that this also applies in the case of a small amount of noise, as implied 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.

5 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 and Steinwart [2010] 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:

$$K(\mu_x, \mu'_x) = e^{\frac{\|\mu_x - \mu'_x\|^2_{\mathcal{F}}}{2\theta}} \tag{1}$$

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* (KTA) is a normalized measure of dependence introduced by Cortes et al. [2012] within the context of multiple kernel learning. The measure is defined as:

$$\text{KTA}(\mathbf{x}, \mathbf{y}) = \frac{\langle K^c_{\mathbf{x}}, K^c_{\mathbf{y}} \rangle_{\mathcal{F}}}{\|K^c_{\mathbf{x}}\|_{\mathcal{F}} \|K^c_{\mathbf{y}}\|_{\mathcal{F}}} \tag{2}$$

Where $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm, $\langle K^c_{\mathbf{x}}, K^c_{\mathbf{y}} \rangle_{\mathcal{F}}$ is the Frobenius norm of the inner product between $K^c_{\mathbf{x}}$ and $K^c_{\mathbf{y}}$ which is calculated by taking the trace of the inner product. $K^c_{\mathbf{x}}$ is a centered kernel matrix, defined as:

$$K^c_{\mathbf{x}} = \left[\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right] K_{\mathbf{x}} \left[\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right]$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a column vector of ones with length m. 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 of 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.

In practice, we note that the KTA based comparison relies on a number of hyper-parameters. The difficulty in

choosing these parameters can result in poorer empirical performance. This problem has also been observed for other kernel-based approaches for causal inference [Zhang et al., 2011]. We leave the investigation of hyper-parameter selection as future work.

6 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. [4]

6.1 REGULAR NETWORKS

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:

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

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

Figure 1 shows the relationship between $D^{-1}A\mathbf{x}$ and $\mathbf{y}$ for varying values of β. In the noiseless case (Figure 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 (i) the square of Pearson's correlation, and (ii) KTA using RBF kernels with a fixed bandwidth of 1.0 for all kernel calculations. Figure 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 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* [Jensen and Neville, 2002], which will reduce the expected efficacy of both methods. The converse of this effect can be seen in Figure 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).

[4] Code is available at `https://github.com/darbour/RelationalCausalDirection.git`.

6.2 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 equal to 0.2. The Watts-Strogatz model generates "small-world networks". It begins with a lattice with a given neighborhood size and randomly rewires edges according to a fixed probability. 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:

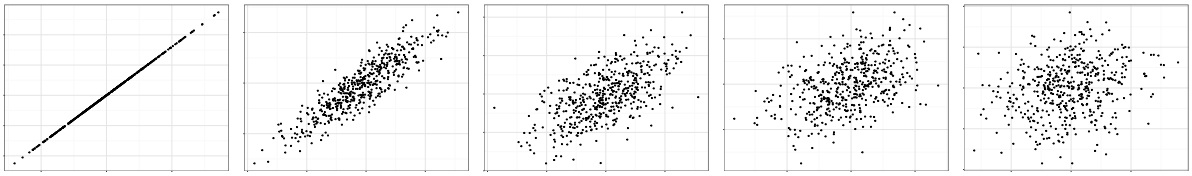
$$\begin{aligned}\mathbf{x} &\sim \mathcal{N}(0,1)\\ \epsilon &\sim \mathcal{N}(0,1)\\ \mathbf{y} &\sim f(D^{-1}A\mathbf{x}) + \beta\epsilon\end{aligned}$$

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)

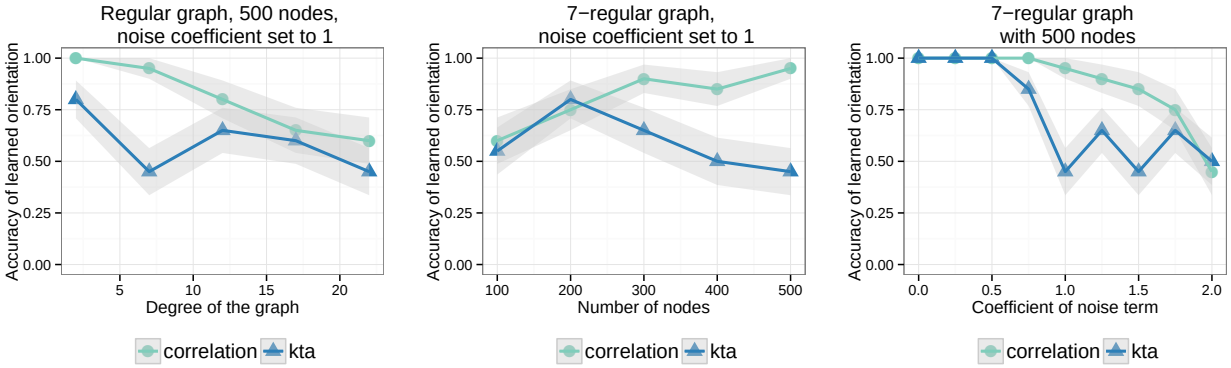
For each setting, β was varied between 0 and 2 by increments of 0.25.

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 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



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

Figure 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.



(a) (b) (c)

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

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 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 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.

6.3 A COMPARISON TO RELATIONAL BIVARIATE EDGE ORIENTATION

We also compared our results to the relational bivariate edge orientation (RBO) [Maier et al., 2013], the only other known method for testing causal direction in relational data. Maier et al. [2013] 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. [2013] 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. [2013] 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 $D^{-1}A\mathbf{x} \rightarrow \mathbf{y}$,

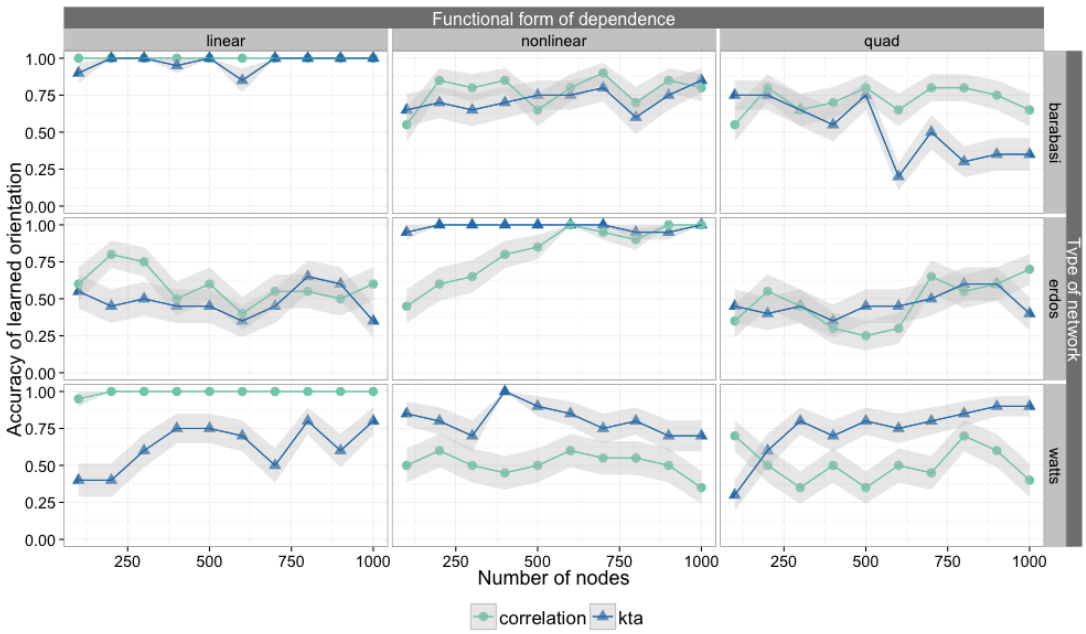


Figure 3: Orientation accuracy for various network types and functional forms, as the size of the graph increases. The noise coefficient is set to 0.5.

otherwise return $D^{-1}A\mathbf{y} \to \mathbf{x}$

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. [2013]. As currently implemented, the RBO rule would have produced approximately $\%50$ error rate, since it does not explicitly check both directions. Using our more conservative method, RBO would fire less frequently. 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.

7 REAL WORLD DEMONSTRATION

In contrast to the propositional setting, where there is a number of labeled ground-truth data-sets for testing novel methods of causal inference (e.g. [Lichman, 2013]), to our knowledge, there are no known publicly available datasets which contain ground-truth relational causal relationships. In the absence of the ability to verify the relative efficacy of our findings on real-world datasets, we provide a demonstration of our method on a real-world dataset. Specifically, we considered Stack Overflow, an online community where users pose and answer questions regarding software development. A user can post a question, which can be answered by anyone else within the community. Other users can then up/down vote questions and the given answers. These votes are tracked and the accrual of achieved points is displayed as the "reputation" of a user on the site. Moreover, users can comment on a question. Comments receive votes as well, but do not affect the reputation of a user. The dataset consists of all users, questions, answers, comments, and votes from the inception of the site to 2014.

We tested three questions about user behavior on Stack Overflow. For every question we consider 100 sub-samples of 1000 data points. We computed KTA and Pearson's correlation in each direction. Significance of dependence was determined by performing permutation tests with 1000 permutations. For all tests we set the significance threshold to be 0.01. When dependence was determined to be statistically significant, we also recorded how many times each direction was chosen by comparing test statistics in both directions.

The first question was: "Is there a relationship between the quality of a question and the quality of its subsequent answers?" To answer this, we used the scores of the questions and answers as proxies for their quality. All methods determined significance in both directions across all trials. However, the normalized statistics consistently determined the direction of dependence to be Question Quality $\to$ Answer Quality, while both of the un-normalized statistics consistently determined the direction of dependence to be

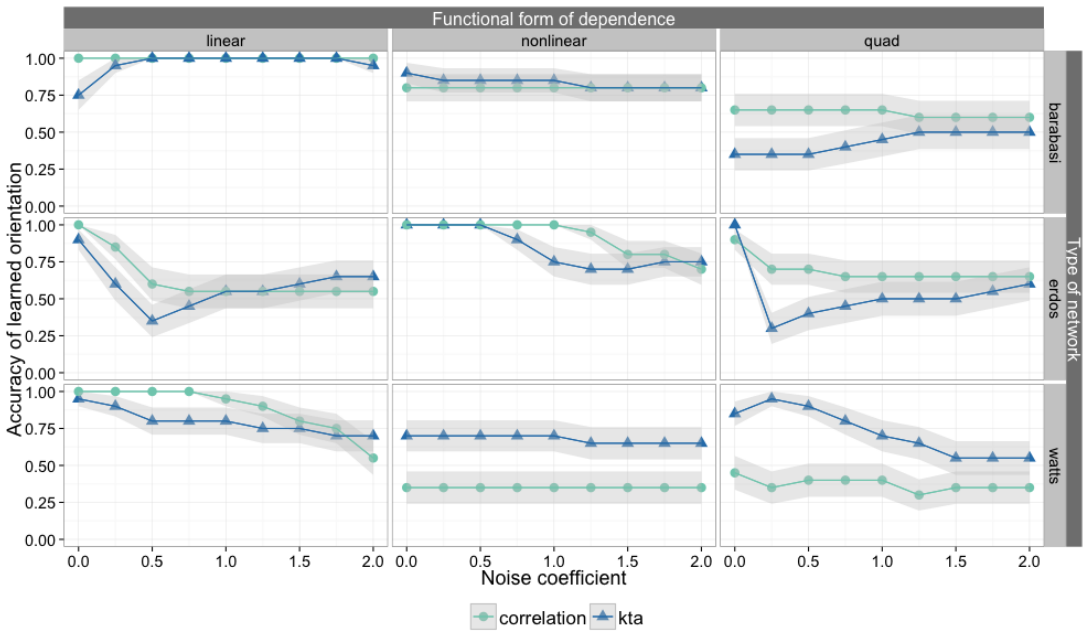


Figure 4: Orientation accuracy for various network types and functional forms, as the coefficient of the noise increases. The network size was kept constant at 1000 nodes.

Question Quality ← Answer Quality. Clearly, the former conclusion matches intuition and temporal ordering far better than the latter.

The second question we considered was whether users with high reputation receive higher quality answers. This was quantified by using the reputation of a user and the score of the answers as a proxy for quality. In this case, we found that KTA and Pearson both detected significance for both directions. For direction, we found that both KTA and Pearson determined direction to be Reputation → Answer Quality for over 90% of the cases. This indicates that there may be bias in the Stack Overflow community towards questions asked by high reputation users. We caution that this does not take into account the possible latent confounder of question quality, i.e., higher reputation users may simply ask higher quality questions.

Finally, we looked at the efficacy of comments as a quality improvement mechanism, i.e., whether allowing users to comment on a question causes the poster to improve or clarify her post. We constructed this test with the comments posted for a question and whether revisions were subsequently made to the question. In this case we found that all of the methods inferred that there was *not* a significant relationship between the score of the comments and subsequent revisions to posts. This negative result indicates that the commenting system provided by Stack Overflow is not an effective mechanism for improving the quality of questions on the site.

8 RELATED WORK

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. [2013], Rattigan [2012], and Poole and Crowley [2013] 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. [2014] 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. [2010]. 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 attention in both statistical relational learning (SRL) community and in multiple areas of the social sciences. In SRL, Jensen and Neville [2002] use a χ^2 test to detect auto-correlation in relational data and show its effect for feature selection. Angin and Neville [2008] 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 [2012] and London et al. [2013] 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. [2014] characterize the threat to validity arising from the bias induced by relational dependence and provide experimental designs to reduce these effects. Manski [2013], VanderWeele [2008], and Aronow and Samii [2013] examine methods for removing the bias associated with relational dependence, assuming discrete or linearly dependent data. Toulis and Kao [2013] provide conditions for experimental design with binary treatments to identify peer influence. Ogburn and VanderWeele [2014] 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 [2009], La Fond and Neville [2010]) is related but distinct in the task setup, as we do not assume the availability of temporal information.

Finally, our work is strongly connected and can serve as a complement to existing work on causal learning of relational domains. Maier et al. [2013] and Marazopoulou et al. [2015] present constraint-based algorithms to learn the structure of relational models from data. However, for their experiments they either rely on a d-separation oracle (without actual data), or use linear regression with mean-aggregation on synthetically generated data. As we showed in our synthetic experiments, these choices can lead to a large number of type II errors. This is especially troublesome for constraint-based structure learning algorithms where type II errors can lead to large deviations from the true causal model [Cornia and Mooij, 2014]. Such algorithms could leverage our test in order to improve results reported on data. Additionally, the directionality results presented in this paper have implications for future work in constraint-based structure learning algorithms, since they imply a smaller Markov-equivalence class than what is commonly assumed.

9 CONCLUSIONS AND FUTURE WORK

Inferring relational dependence is a task of general interest in a wide number of fields, from statistical relational learning to the social sciences. In this work, we have studied the problem of inferring causal direction in relational data. We have shown that, in contrast to the propositional setting, causal direction can be accurately inferred in relational data under the simplest functional forms such as linear deterministic dependence, without additional assumptions on the distribution of the underlying data. We then studied the problem of identifying confounding, showing the conditions when the presence of a relational confounding variable can be identified. Our experimental evaluation shows that these measures are robust, providing accurate inference under model and network mis-specification.

There are several promising avenues for future research. For causal learning, the ability to detect the direction of dependence in relational data implies that a different Markov equivalence class [Spirtes et al., 2000] holds for the relational setting than what is commonly assumed. Integration of the findings of this work into a causal learning algorithm could substantially improve the efficacy of existing methods such as RCD [Maier et al., 2013]. Further analysis of the interaction between the network structure and inference may further strengthen the robustness of the methods discussed here. Finally, the asymmetries shown to be inherent to relational data here may result in significant bias of conditional independence testing procedures. Incorporating this additional information is a first step in developing robust measures of conditional dependence in relational data to help determine causation, a problem which has broad application in both the statistical learning and social science communities.

Acknowledgements

Funding was provided by the U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA) under Contract Number W911NF-11-C-0088. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

P. Angin and J. Neville. A shrinkage approach for modeling non-stationary relational autocorrelation. In *8th International Conference on Data Mining*, pages 707–712, 2008.

P. M. Aronow and C. Samii. Estimating average causal

effects under interference between units. *arXiv preprint arXiv:1305.6156*, 2013.

E. Bakshy, D. Eckles, R. Yan, and I. Rosenn. Social influence in social advertising: evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 146–161. ACM, 2012.

N. Christakis and J. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Hachette Digital, Inc., 2009.

A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *Advances in Neural Information Processing Systems*, pages 406–414, 2010.

N. Cornia and J. M. Mooij. Type-II errors of independence tests can lead to arbitrarily large errors in estimated causal effects: An illustrative example. In *Proceedings of the UAI 2014 Workshop Causal Inference: Learning and Prediction*, pages 35–42, 2014.

C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828, 2012.

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

A. Dhurandhar and A. Dobra. Distribution-free bounds for relational classification. *Knowledge and information systems*, 31(1):55–78, 2012.

D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *arXiv preprint arXiv:1404.7530*, 2014.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 259–266, 2002.

D. Koller. Probabilistic relational models. In *Inductive logic programming*, pages 3–13. Springer, 1999.

T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th international conference on World wide web*, pages 601–610, 2010.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

B. London, B. Huang, B. Taskar, and L. Getoor. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013.

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

M. Maier, K. Marazopoulou, D. Arbour, and D. Jensen. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 371–380, 2013.

C. F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

K. Marazopoulou, M. Maier, and D. Jensen. Learning the structure of causal models with relational and temporal dependence. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015.

L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

E. L. Ogburn and T. J. VanderWeele. Causal diagrams for interference. *Statistical Science*, 29(4):559–578, 2014.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

D. Poole and M. Crowley. Cyclic causal models with discrete variables: Markov chain equilibrium semantics and sample ordering. In *Proceedings of the 23rd international joint conference on Artificial Intelligence*, pages 1060–1068, 2013.

M. J. Rattigan. *Leveraging Relational Representations for Causal Discovery*. Ph.D. thesis, University of Massachusetts Amherst, 2012.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695, 2010.

P. Toulis and E. Kao. Estimation of causal peer influence effects. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1489–1497, 2013.

T. J. VanderWeele. Ignorability and stability assumptions in neighborhood effects research. *Statistics in medicine*, 27(11):1934–1943, 2008.

K. Zhang, J. Peters, and D. Janzing. Kernel-based conditional independence test and application in causal discovery. In *In Uncertainty in Artificial Intelligence*. Citeseer, 2011.