Causal Relational Learning

Babak Salimi[1], Harsh Parikh[2], Moe Kayali[1],
Lise Getoor [3], Sudeepa Roy[2], Dan Suciu[1]
[1] University of Washington [2] Duke University [3] University of California, Santa Cruz

ABSTRACT

Causal inference is at the heart of empirical research in natural and social sciences and is critical for scientific discovery and informed decision making. The gold standard in causal inference is performing *randomized controlled trials*; unfortunately these are not always feasible due to ethical, legal, or cost constraints. As an alternative, methodologies for causal inference from *observational data* have been developed in statistical studies and social sciences. However, existing methods critically rely on restrictive assumptions such as the study population consisting of *homogeneous elements* that can be represented in a single flat table, where each row is referred to as a *unit*. In contrast, in many real-world settings, the study domain naturally consists of *heterogeneous elements* with complex relational structure, where the data is naturally represented in multiple related tables. In this paper, we present a formal framework for causal inference from such relational data. We propose a declarative language called CaRL for capturing causal background knowledge and assumptions, and specifying causal queries using simple Datalog-like rules. CaRL provides a foundation for inferring causality and reasoning about the effect of complex interventions in relational domains. We present an extensive experimental evaluation on real relational data to illustrate the applicability of CaRL in social sciences and healthcare.

ACM Reference Format:
Babak Salimi[1], Harsh Parikh[2], Moe Kayali[1], and Lise Getoor [3], Sudeepa Roy[2], Dan Suciu[1]. 2020. Causal Relational Learning. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD'20), June 14–19, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3318464.3389759

SIGMOD'20, June 14–19, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6735-6/20/06…$15.00
https://doi.org/10.1145/3318464.3389759

1 INTRODUCTION

The importance of causal inference for making informed policy decisions has long been recognised in health, medicine, social sciences, and other domains. However, today's decision-making systems typically do not go beyond *predictive analytics* and thus fail to answer questions such as "What would happen to revenue if the price of X is lowered?" While predictive analytics has achieved remarkable success in diverse applications, it is mostly restricted to fitting a model to observational data based on associational patterns [38]. Causal inference, on the other hand, goes beyond associational patterns to the process that generates the data, thereby enabling analysts to reason about *interventions* (e.g., "Would requiring flu shots in schools reduce the chance of a future flu epidemic?") and *counterfactuals* (e.g., "What would have happened if past flu shots were not taken?"). This adds significantly more information in data analysis compared to simple correlation or regression analysis; e.g., as the number of flu cases increases, the rate of flu shots might also increase, but that does not imply that giving flu shots increases the spread of flu. This emphasizes the common saying that "correlation is not causation", which is known to all, but is easy to overlook if one is not careful while analyzing data for insights and possible actions.

The gold standard in causal analysis is performing *randomized controlled trials*, where the *subjects* or *units* of study are assigned randomly to a treatment or a control (i.e., withheld from the treatment) group. The difference between the distribution of the outcome variable of the treated and control groups represents the *causal effect* of the treatment on outcome. However, control experiments are not always feasible due to ethical, legal, or cost constraints [2, 46]. An attractive alternative that has been used in statistics, economics, and social sciences simulates control experiments using available *observational data*. While we can no longer assume that the treatment has been randomly assigned, under appropriate assumptions we can still estimate causal effects. Rubin's Potential Outcome Framework [43] and Pearl's Causal Models [35] (reviewed in Section 2) are two well-established frameworks which have been extensively studied in the literature and used in various applications for causal inference from observational data [2, 6, 30, 34, 45].

Authors		
person	prestige	qualification (h-index)
Bob	1	50
Carlos	0	20
Eva	1	2

Submissions	
sub	score
s1	0.75
s2	0.4
s3	0.1

Authorship	
person	sub
Bob	s1
Eva	s1
Eva	s2
Eva	s3
Carlos	s3

Submitted	
sub	conf
s1	ConfDB
s2	ConfAI
s3	ConfAI

Conferences	
conf	blind
ConfDB	Single
ConfAI	Double

Figure 1: A multi-relational ReviewData instance.

Causal frameworks, however, rely on the critical assumption that the units of study are sampled from a population of homogeneous units; in other words, the data can be represented in a single flat table. This assumption is called the unit homogeneity assumption [2, 16]. In many real-world settings, however, the study domain consists of *heterogeneous units* that have a *complex relational structure*; and the data is naturally represented as multiple related tables. For instance, as presented later in our experiments with real data [14, 19], hospitals can record in several tables information about patients, medical practitioners, hospital stays, treatments performed, insurance, bills, and so on. Standard notions used in causal analysis — such as units or subjects who receive a treatment in causal analysis — no longer readily apply to relational data, prohibiting us from adopting existing causal inference frameworks to relational domains. We illustrate these challenges with the following example.

Example 1.1. Consider researchers trying to understand the impact of single-blind and double-blind reviewing policies on the review scores of submissions, in particular, understanding how the prestige of authors affects the fairness of decisions. In this setting, the outcome of interest is the review scores of a submission, and the treatment is the prestige of the authors (Figure 1 shows a simplified schema for the domain. See Section 6.1 for full details about the actual ReviewData which comes from OpenReview [33] and other datasets.)

For answering causal questions such as "Is there an effect of the prestige of authors on the review score received by the submission at a conference?", we need to control for confounders like the quality of submissions and conferences where they are submitted. This requires not only joining across multiple tables, but it also requires aggregating over authors since the authors table is related to paper submissions by a many-to-many authorship relation.

Our contributions. In this paper, we propose a declarative framework for *Causal Relational Learning*, a foundation for causal inference over relational domains. Our first contribution is a declarative language, *CaRL (Causal Relational Language)*, for representing causal background knowledge and assumptions in relational domains (**Section 3**). CaRL can represent complex causal models using just a few rules. The syntax of CaRL is designed to be intuitive for users to represent complex causal models and ask causal queries, while the details of their semantics and query answering are abstracted from the users who need not be statisticians.

Our second contribution is to define semantics for *complex causal queries* where the treatment units and outcome units might heterogeneous and controlling for confounding may require performing multiple joins and aggregates (**Section 4**). Using CaRL, we can answer complex causal queries such as:"what is the effect of not having an insurance on mortality of a patient in a critical care unit?", where we are interested in estimating the *average treatment effect* (defined later), or "what is the effect of authors' collaborators' prestige on acceptance of a paper?", where we are interested in estimating the *average relational effect*; several other types of queries are also supported.

Our third contribution consists of an algorithm for answering causal queries from the given relational data (**Section 5**). The algorithm performs a static analysis of the causal query, and it constructs a unit-table specific to the query and the relational causal model by identifying a set of attributes that are sufficient for confounding adjustment. The constructed unit-table is amenable to sound causal inference using existing techniques.

Finally, we present an end-to-end experimental evaluation of CaRL on both real and synthetic data (**Section 6**). The experiments conducted on the following real-world relational datasets: 1) ReviewData [33, 39, 40], 2) MIMIC-III (Medical Information Mart for Intensive Care Data) [19], and 3) NIS (National Inpatient Sample Data) [14]. We examine the following causal queries:

- ReviewData. What is the effect of authors' prestige on the scores given by the receivers under single-blind and double-blind review processes?
- MIMIC-III. What is the effect of not having insurance on patient's mortality and length of hospital stay?
- NIS. What is the effect of hospital size on healthcare affordability?

In each setting, we report contrasts between correlation and causation, further highlighting the need for principled causal analysis. Evaluation of CaRL on synthetic data showed that causal analysis ignoring the relational structure of data failed to recover the ground truth, but CaRL successfully recovered accurate results.

2 BACKGROUND ON CAUSALITY

This section reviews basics of causal inference. We use capital letters X to denote random variables, and use lower case letters x to denote their values. We use boldface $\mathbf{X}$, $\mathbf{x}$ to denote tuples of random variables and constants respectively;

Figure 2: A standard causal DAG for Example 3.2.

and $Dom(X)$ denotes the domain of variable X.

Probabilistic causal models. A probabilistic causal model [36] is a tuple $M = \langle \mathbf{U}, \mathbf{V}, \Pr_{\mathbf{U}}, \mathbf{F} \rangle$, where $\mathbf{U}$ is a set of *exogenous* variables that cannot be observed, $\mathbf{V}$ is a set of *observable or endogenous* variables, and $\Pr_{\mathbf{U}}$ is a joint probability distribution on the exogenous variables $\mathbf{U}$. The set $\mathbf{F} = (F_X)_{X \in \mathbf{V}}$ is a set of *non-parametric structural equations* of the form $F_X : Dom(\mathbf{Pa_V}(X)) \times Dom(\mathbf{Pa_U}(X)) \rightarrow Dom(X)$, where $\mathbf{Pa_U}(X) \subseteq \mathbf{U}$ and $\mathbf{Pa_V}(X) \subseteq \mathbf{V} - \{X\}$ are called the *exogenous parents* and *endogenous parents* of X respectively. Intuitively, the exogenous variables $\mathbf{U}$ are not known, but we know their probability distribution; the endogenous variables are completely determined by their parents (exogenous and/or endogenous).

Causal DAG. A probabilistic causal model is associated with a *causal DAG* (directed acyclic graph) G, whose nodes are the endogenous variables $\mathbf{V}$, and whose edges are all pairs (Z, X) such that $Z \in \mathbf{Pa_V}(X)$. The causal DAG hides exogenous variables (since we cannot observe them anyway) and instead captures their effect by defining a probability distribution $\Pr_{\mathbf{U}}$ on the endogenous variables.[1] We will only refer to endogenous variables in the rest of the paper and drop the subscript $\mathbf{V}$ from $\mathbf{Pa_V}$. Similarly, we will drop the subscript $\mathbf{U}$ from the probability distribution $\Pr_{\mathbf{U}}$ when it is clear from the context. Then the formula for $\Pr(\mathbf{V})$ is the same as that for a Bayesian network:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X | \mathbf{Pa}(X)) \tag{1}$$

Figure 2 shows a simple example of a causal graph based on Example 1.1: the Score of a paper is affected by its Quality and by the Prestige of the author (assuming the reviews are single blind), whereas both Quality and Prestige are affected by the author's Qualification. Here $\mathbf{V}$ = {Qualification, Quality, Prestige, Score} are endogenous variables, $\mathbf{U}$ endogenous variables are unknown (e.g., mood of a reviewer while reviewing the paper, the expected number of papers to be accepted, scores of other papers the reviewer reviewed, etc.) leading to a probability distribution on $\mathbf{V}$. The dependencies can be represented by three structural equations:

$$\text{Quality} \Leftarrow \text{Qualification};\ \ \text{Prestige} \Leftarrow \text{Qualification};$$
$$\text{Score} \Leftarrow \text{Quality}, \text{Prestige}. \tag{2}$$

Interventions and the do operator. Causal models give semantics to *interventions*. An intervention represents actively setting an endogenous variable to some fixed value and observing the effect denoted by the *do*-operator introduced by Pearl [37]. Formally, an intervention $\mathtt{do}(\mathbf{W} = \mathbf{w})$ consists of setting variables $\mathbf{W} \subseteq \mathbf{V}$ to some values $\mathbf{W} = \mathbf{w}$, and it defines the probability distribution $\Pr(\mathbf{V} | \mathtt{do}(\mathbf{W} = \mathbf{w}))$ given by (1), where we remove all factors $\Pr(X | \mathbf{Pa}(X))$, where $X \in \mathbf{W}$. In other words, we modify the causal DAG by removing all edges entering the variables $\mathbf{W}$ on which we intervene; this fundamentally differs from conditioning, $\Pr(\mathbf{V} | \mathbf{W} = \mathbf{w})$. Pearl has an extensive discussion of the rationale for the *do*-operator and describes several equivalent formulas for estimating the effect of $\mathtt{do}(\mathbf{W} = \mathbf{w})$ from an observed distribution [36].

Average treatment effect (ATE). The causal analysis estimates the effect treatment variable T (typically a binary variable) on some outcome variable Y. This effect is often measured by the following quantity known as the *average treatment effect (ATE)*, which is expressed as follows in our notation:

$$\text{ATE}(Y, T) = \mathbb{E}[Y | do(T = 1)] - \mathbb{E}[Y | do(T = 0)] \tag{3}$$

Much of the literature on causal inference in statistics addresses efficient estimation of ATE from observational data.

Unit of analysis and SUTVA. Both Pearl's [36] and Rubin's causality [44] rely on the assumption that the study domain consists of a set of *units*, or physical objects (e.g., authors, patients, publications, etc.) that can be subject to a treatment/intervention and exhibit a response to it. Furthermore, they rely on the assumption of *no interference between the units* or *Stable Unit Treatment Value Assumption (SUTVA)* [44]. Intuitively SUTVA states that intervening on or treating a unit does not have any consequences on the response of other units. In settings where the units of analysis are relationally connected, this assumption is typically violated. In Example 1.1, prestige of an author (treatment) influences the acceptance chance (response) of his or her co-author(s) and collaborator(s) which leads to the violation of SUTVA.

Related work. Previous work has studied causal inference in the presence of interference [4, 10, 12, 13, 29–31, 51, 54, 56]. These works address applications such as the study of infectious diseases [13, 56] or behavior and interactions in social networks [10, 17, 23, 29, 51, 53, 56]. But in these studies the units are still homogeneous (e.g., people connected by a social network), and they are unable to capture different *entities* of interests like papers, authors, reviews and their complex many-to-many relationships that we focus on in CaRL. There has been prior work on learning causality from relational data [3, 21, 22, 24]; it focuses on discovering the structure of probabilistic graphical models for this data. These models were originally proposed for Statistical Relational Learning, which aims to model a joint probability distribution over relational data amenable for

[1] This is possible under the *causal sufficiency* assumption: for any two variables $X, Y \in \mathbf{V}$, their exogenous parents are disjoint and independent $\mathbf{Pa_U}(X) \perp\!\!\!\perp \mathbf{Pa_U}(Y)$. When this assumption fails, one adds more endogenous variables to the model to expose their dependencies.

probabilistic reasoning rather than causal inference [8]. This line of work differs from our work in that our objective is to develop a declarative framework to answer complex causal queries about the effect of interventions, given the existing background knowledge. Note that *causality* has been used in various contexts [27], namely, to understand responsibility in query answering [25, 47], in database repair [26, 48], and to motivate explanations [42]. It has also inspired different applications such as hypothetical reasoning [5, 7, 20, 28]. These differ from our work in that they identify parts of the input that are correlated with the output of a transformation, which is useful but does not reflect the true causality needed for decision making.

3 CARL: DECLARATIVE FRAMEWORK

In this section, we present our declarative language called *CaRL (Causal Relational Language)* that extends causal modeling to relational data by allowing the user to (1) specify assumptions and background knowledge on the interactions among heterogenous units (Section 3.2), and (2) pose various causal queries (Section 3.3). We start with our data model, which forms the basis for our language.

3.1 Data Model: Schema and Instance

Relational causal schema (schema). The input schema for CaRL corresponds to any standard multi-relational database, e.g., ReviewData in Figure 1, but we assume the data is given in the following 'entity-relationship-attribute' form for a simpler generalization of Pearl's causal models. A *relational causal schema* is a tuple $\mathbf{S} = (\mathbf{P}, \mathbf{A})$, where $\mathbf{P} = \mathbf{E} \cup \mathbf{R}$ represents a set of *entities* $\mathbf{E}$ and their *relationships* $\mathbf{R}$, and $\mathbf{A}$ represents a set of *attribute functions* (or simply *attributes*) that encode the standard attributes of the entities and their relationships, with the only difference that some of these attributes may be '*unobserved*' with missing values in all instances. The entities and their relationships are denoted by $P(.)$, the attribute functions are denoted by $\mathtt{A}[.]$, and $\mathbf{A}_{Obs} \subseteq \mathbf{A}$ denotes the set of observed attribute functions. We illustrate the mapping from standard relational model to relational causal schema using our running example.

Example 3.1. The relational causal schema corresponding to the relational ReviewData in Figure 1 (with renames) is:

$$\mathbf{P} = \text{Person(A)}, \text{Author}(A, S), \text{Submission}(S), \text{Submitted}(S, C), \text{Conference}(C)$$
$$\mathbf{A} = \text{Prestige}[A], \text{Qualification}[A], \text{Score}[S], \text{Blind}[C], \text{Quality}[S]$$

Here $\mathbf{P}$ *consists of entities in the ReviewData:* $\mathbf{E} = \{$*People, Submission, Conference*$\}$ *and their relationships* $\mathbf{R} = \{$*Authors, Submitted*$\}$*; The attribute function* $\mathbf{A}$ *corresponds to the attributes of these entities and relationships: Prestige*$[A]$ *= the prestige of the author's institution (e.g., rankings); Qualification*$[A]$ *= the qualification of an author by h-index*[2]*); Score*$[S] \in [0, 1]$ *= the average score reviewers gave to a submission; Blind*$[C]$ *= whether a conference review policy is single or double blind; Quality*$[S]$ *= the quality of a submission. Note that Quality in* $\mathbf{A}$ *is missing in the Submissions table in Figure 1, since it is an unobserved attribute function, and will be used in causal analysis based on our background knowledge that quality of a submission may have an impact on its score.*

Observed instance and relational skeleton (instance). Similar to a standard database instance given a standard relational schema (as shown in Figure 1), an *observed relational instance* (or simply an *instance*) conforms to a given relational causal schema $\mathbf{S} = (\mathbf{P}, \mathbf{A})$ with specific values (i.e., constants), however some (unobserved) attribute functions may be missing in the instance (like 'Quality'). The set of (constant or grounded) entities and relationships in an instance (excluding the grounded attribute functions) is referred to as the *relational skeleton* of the instance and denoted by Δ.

Example 3.2. For relational causal schema given in Example 3.1 and the instance in Figure 1, the relational skeleton comprises entities and relationships like `Person("Bob")`, `Submission("s1")`, `Author("Bob", "s1")`, *etc. The observed instance comprises the relational skeleton and the attribute functions like* `Score["s1"]`, `Blind["ConfDB"]`, *etc., but not unobserved attributes like* `Quality["s1"]`. *Note that all observed attribute functions assume a fixed value given any instance.*

3.2 Specification of Background Knowledge by Relational Causal Rules

3.2.1 Relational causal model and rules. The first step of using CaRL is encoding the user's background knowledge about potential causal dependencies among attributes in an application. This is expressed in CaRL through a set of *relational causal rules* (defined below) that capture the causal assumptions. We refer to the set of relational causal rules specified by the user as the *relational causal model.*

Definition 3.3. A relational causal rule *over a relational causal schema* $\mathbf{S} = (\mathbf{P}, \mathbf{A})$ *has the following form:*

$$\mathtt{A}[\mathbf{X}] \Leftarrow \mathtt{A}_1[\mathbf{X}_1], \ldots, \mathtt{A}_k[\mathbf{X}_k] \text{ WHERE } Q(\mathbf{Y}) \tag{4}$$

Here, $\mathtt{A}, \mathtt{A}_1, \cdots, \mathtt{A}_k \in \mathbf{A}$ *are attribute functions,* Q *is a (standard) conjunctive query over the schema* $\mathbf{P}$*, and* $\mathbf{X}$, $\mathbf{X_i}$ *(*$i = 1, \cdots, k$*),* $\mathbf{Y}$ *are sets of variables and/or constants. All variables in* $\mathbf{X} \cup \bigcup_{\mathbf{i}} \mathbf{X_i}$ *must also occur in* $\mathbf{Y}$*. We call* $\mathtt{A}[\mathbf{X}]$ *the* head *of the rule,* $\mathtt{A}_1[\mathbf{X}_1], \ldots, \mathtt{A}_k[\mathbf{X}_k]$ *the* body *of the rule, and* $Q(\mathbf{Y})$ *the* condition*. We denote by* $\phi_{\mathtt{A}}$ *the set of rules with head* $\mathtt{A}$*.*

[2] There can be other measures of qualifications as well, e.g., the number of publications or citations, or the experience in terms of years.

Example 3.4. Consider the following relational causal model Φ for ReviewData in Figure 1.

$$\text{Prestige}[A] \Leftarrow \text{Qualification}[A] \text{ WHERE Person}(A) \quad (5)$$

$$\text{Quality}[S] \Leftarrow \text{Qualification}[A], \text{Prestige}[A] \text{ WHERE Author}(A, S) \quad (6)$$

$$\text{Score}[S] \Leftarrow \text{Prestige}[A] \text{ WHERE Author}(A, S), \quad (7)$$

$$\text{Score}[S] \Leftarrow \text{Quality}[S] \text{ WHERE Submission}(S) \quad (8)$$

Rule (5) says that the qualification of a person causally affects his or her institutions' prestige; rule (6) says that the quality of a submission is affected by its authors' qualifications and prestige (authors from prestigious institutions have access to more resources); rules (7) and (8) say that reviewers' scores are based on the quality of a submission but may also be influenced by the prestige of its authors.

A major advantage of specifying background knowledge using causal rules for the users is that they simply express intuitive potential causal dependence among attributes without mentioning 'how' or associating any 'weight' to them[3], while CaRL uses them to answer different causal queries (Section 3.3).

3.2.2 Grounded rules. A *grounded rule* is a rule (4) that contains only constants from a given instance (no variables) and has no condition (i.e., $Q \equiv \texttt{true}$). A relational causal rule is a template for generating multiple grounded rules.

Definition 3.5. Let Δ be a relational skeleton. Fix a rule in the form of (4), and let $\mathbf{Z}$ denote all variables occurring in $\mathbf{X} \cup \mathbf{X}_1 \cup \ldots \cup \mathbf{X}_k$. We associate to this rule the set of grounded rules *obtained by substituting $\mathbf{Z}$ with any set of constants $\mathbf{z}$ such that $\Delta \models Q([\mathbf{Y}/\mathbf{z}])$. In other words, the query Q must be true in the database Δ after substituting the variables $\mathbf{Z}$ with the constants $\mathbf{z}$ and treating the variables $\mathbf{Y} - \mathbf{Z}$ as existentially quantified.*

3.2.3 Relational causal graphs. Given a relational causal model Φ comprising a set of relational causal rules and a relational skeleton Δ comprising the entities and relationships in an instance, Φ^{Δ} denotes the set of all grounded rules. From Φ^{Δ}, we construct the *relational causal graph* $G(\Phi^{\Delta})$. The vertices of $G(\Phi^{\Delta})$ (denoted $\mathbf{A}^{\Delta}$) comprise all grounded attributes $\mathtt{A}[\mathbf{x}]$ in Φ^{Δ} denoted $\mathtt{A}^{\Delta}$ – recall that $\mathbf{x}$ represents a tuple of constants, an attribute function $\mathtt{A}$ corresponding to an entity has a single constant parameter as in Example 3.2, but $\mathtt{A}$ corresponding to a relationship predicate will have multiple parameters. The edges of $G(\Phi^{\Delta})$ are all pairs $(\mathtt{A}[\mathbf{x}], \mathtt{A}_j[\mathbf{x}_j])$ where $\mathtt{A}[\mathbf{x}]$ and $\mathtt{A}_j[\mathbf{x}_j]$ appear in the head and body respectively of a grounded rule (4). We assume that the relational causal model is non-recursive, therefore, the causal graph is a DAG[4].

[3]This fact, along with the declarative nature of the language, makes CaRL more friendly to users who are not causal inference experts.

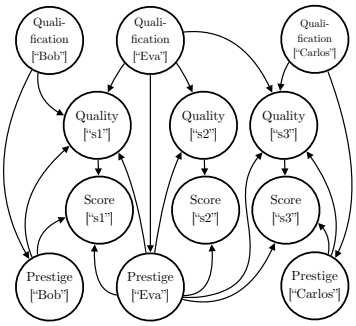

Figure 3: Relational causal graph corresponding to the grounded rules in Example 3.4.

Example 3.6. Given the skeleton Δ in Figure 1, Φ generates the following grounded rules:

$$\begin{aligned}
\text{Prestige}[\text{“}Bob\text{”}] &\Leftarrow \text{Qualification}[\text{“}Bob\text{”}] \text{ – } \textit{(also for “Carlos”, “Eva”)} \\
\text{Quality}[\text{“}s1\text{”}] &\Leftarrow \text{Qualification}[\text{“}Bob\text{”}], \text{Qualification}[\text{“}Eva\text{”}], \\
&\quad \text{Prestige}[\text{“}Bob\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \\
\text{Quality}[\text{“}s2\text{”}] &\Leftarrow \text{Qualification}[\text{“}Eva\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \\
\text{Quality}[\text{“}s3\text{”}] &\Leftarrow \text{Qualification}[\text{“}Carlos\text{”}], \text{Qualification}[\text{“}Eva\text{”}], \\
&\quad \text{Prestige}[\text{“}Carlos\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \\
\text{Score}[\text{“}s1\text{”}] &\Leftarrow \text{Quality}[\text{“}s1\text{”}], \text{Prestige}[\text{“}Bob\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \\
\text{Score}[\text{“}s2\text{”}] &\Leftarrow \text{Quality}[\text{“}s2\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \\
\text{Score}[\text{“}s3\text{”}] &\Leftarrow \text{Quality}[\text{“}s3\text{”}], \text{Prestige}[\text{“}Carlos\text{”}], \text{Prestige}[\text{“}Eva\text{”}]
\end{aligned} \quad (9)$$

These in turn lead to the causal graph shown in Figure 3.

Note that the relational causal graph in Figure 3 is an extension of the standard causal DAG (by Pearl's model [35]) shown in Figure 2: the latter describes the potential causal dependence of the attributes whereas the former describes a more fine grained version based on the entities and relationships in the relational data. For example, we do not have a single node *Score*, as in Figure 2, but instead have many nodes *Score*["*s*1"], *Score*["*s*2"], etc. one for each submission in Δ in Figure 3. As in Section 2, the causal graph $G(\Phi^{\Delta})$ defines a joint probability distribution

$$\Pr(\mathtt{A}[\mathbf{x}] \mid \mathbf{Pa}(\mathtt{A}[\mathbf{x}])) \quad (10)$$

with one conditional probability for each grounded attribute $\mathtt{A}[\mathbf{x}]$; we describe these conditional probabilities in Section 4.1.

3.2.4 Aggregated rules. Using CaRL, one can extend the set of attribute functions $\mathbf{A}$ with new aggregated attribute functions using one of the *aggregate rules* of the following forms. For $\mathtt{A} \in \mathbf{A}$

$$\text{AGG\_A}[\mathbf{W}] \Leftarrow \mathtt{A}[\mathbf{X}] \text{ WHERE } Q(\mathbf{Z}) \quad (11)$$

[4]While our language allows for recursive rules which capture feed-back loops and contagion, their treatment is beyond the scope of the paper and is an interesting future work.

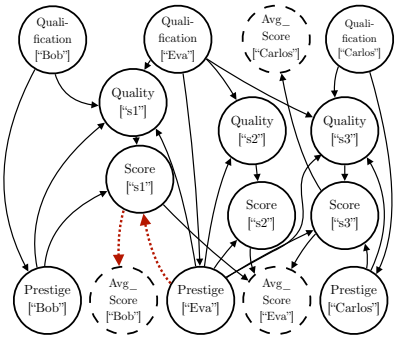

Figure 4: Extended relational causal graph from Figure 3 with aggregated attribute AVG_Score$[A]$ by (12). The directed path from relational peer Eva's prestige to average score of Bob is highlighted (Section 4.3).

Here, $\mathbf{Z} \supseteq \mathbf{X} \cup \mathbf{W}$ and AGG is an aggregate function on A, e.g., AVG (average) and VAR (variance). The new aggregated attribute functions $AGG\_\mathtt{A}$ are included in the extended attribute functions $\mathbf{A}$ (for simplicity, we use $\mathbf{A}$ for both given and extended attribute functions). Similar to relational causal rules, aggregated rules define a set of grounded rules with corresponding vertices and edges in the relational causal graph $G(\Phi^{\Delta})$. However, instead of a conditional probability distribution, a deterministic function $\text{AGG}(\mathbf{Pa}(\text{AGG\_}Y[\mathbf{w}]))$ will be associated with each $\text{AGG\_}Y[\mathbf{w}] \in \text{AGG\_}Y^{\Delta}$. For example, the following aggregate rule defines the average review score for each author.

$$\text{AVG\_Score}[A] \Leftarrow \text{Score}[S] \text{ WHERE Author}(A, S) \tag{12}$$

Figure 4 shows the extension of Figure 3 with (12).

3.3 Causal Query Language in CaRL

Once the relational causal model Φ is specified, users can start asking causal queries. CaRL supports three types of causal queries of the following form (their semantics are discussed in Section 4.4 and answering these queries is discussed in Section 5). The *ATE query* extends the notion standard ATE (discussed in Section 2) for relational data. CaRL also supports queries for aggregated response, isolated effect and relational effect.

Average treatment effect (ATE) query. An ATE query estimates the average treatment effect (see Section 2) of a *treatment attribute* $T[\mathbf{X}]$ on a response attribute $Y[\mathbf{X'}]$ and has the following form: (formally defined in Section 4.4.1)

$$Y[\mathbf{X'}] \Leftarrow T[\mathbf{X}]? \tag{13}$$

This asks "what is the effect of T on Y?". For example, the query $\text{Score}[S] \Leftarrow \text{Prestige}[A]?$ computes the ATE of Prestige of authors on Score of a paper, *i.e.*, it compares papers' scores in two hypothetical worlds in which all authors are and are not affiliated with prestigious institutions (the causal effects of 'some' authors being from prestigious institutions can be estimated from the relational effects queries described below). Following the standard assumption of binary treatments in the causality literature, we require the treatment attribute to be of binary domain, which can be enforced by using a threshold or a predicate on a non-binary domain.

Aggregated response query. An aggregated response query allows causal analysis on an aggregated form of the response variable and has the following syntax (formally defined in Section 4.4.2):

$$\text{AGG\_}Y[\mathbf{X'}] \Leftarrow T[\mathbf{X}]? \tag{14}$$

For example, $\text{AVG\_score}[A] \Leftarrow \text{Prestige}[A]?$ computes the treatment effect of the prestige of authors on the average score received by an author.

Relational, isolated, and overall effects queries. In relational domains, units that are relationally connected can have a causal influence on each other. For example, the Prestige of an author not only influences their average submission scores but also their collaborators' average submission scores. To measure such complex relational causal interactions, CaRL supports queries of the following form that output three quantities: relational, isolated, and overall causal effects (formally defined in Section 4.4.3):

$$Y[\mathbf{X'}] \Leftarrow T[\mathbf{X}] \text{ ? WHEN } \langle cnd \rangle \text{ PEERS TREATED} \tag{15}$$

where $\langle cnd \rangle$ is a condition with the following grammar:

$$\langle cnd \rangle \leftarrow \langle \text{LESS} \mid \text{MORE} \rangle \text{ THAN } k\% \mid \text{AT } \langle \text{MOST} \mid \text{LEAST} \rangle \text{ k} \mid \text{EXACTLY k} \mid \text{ALL} \mid \text{NONE} \tag{16}$$

For example, the query $Score[S] \Leftarrow Prestige[A]?$ WHEN ALL PEERS TREATED computes three values for (i) isolated (an author's prestige), (ii) relational (his/her coauthor's prestige), and (iii) overall (all authors' prestige) effect of prestige on a submission's score.

4 SEMANTICS FOR RELATIONAL CAUSAL ANALYSIS

This section defines semantics of the causal queries described in Section 3.3. We fix a relational causal schema $\mathbf{S}$, a relational skeleton $\mathbf{\Delta}$, and a relational causal model Φ with a corresponding grounded causal graph $G(\Phi^{\Delta})$. For an attribute function $\mathtt{A} \in \mathbf{A}$, denote $\mathbb{U}_{\mathbf{A}}$ to be the set of all tuples of grounded entities $\mathbf{x}$ such that $\mathtt{A}[\mathbf{x}] \in \mathbf{A}^{\Delta}$. For example, $\mathbb{U}_{Prestige}$ consists of all authors, e.g., {"Bob", "Eva", "Carlos"}, whereas $\mathbb{U}_{Score}$ consists of all submissions, e.g., {*"s1", "s2"*}. We refer to each element $\mathbf{x} \in \mathbb{U}_{\mathbf{A}}$ as a *unit* of an attribute function A.

4.1 Probability Distribution for CaRL

As discussed in Section 2, a causal DAG associates a conditional probability distribution $\Pr(X|\mathbf{Pa}(X))$ to each node

$X \in \mathbf{V}$; these conditional probability distributions are unknown and must be estimated from available data even for standard causal graphs described in Section 2, while there are additional challenges for relational causal graphs. As described in Section 3.2, in CaRL, the relational causal graph $G(\Phi^{\Delta})$ is obtained by grounding the rules w.r.t. the skeleton database, Δ, and the number of nodes is not fixed ahead of time but depends on Δ.

We introduce the following *structural homogeneity assumption*, which is critical in CaRL to estimate the conditional probability distributions from a given observed dataset, and thereby conduct causal inference. Recall that $\mathtt{A}^{\Delta} \subseteq \mathbf{A}^{\Delta}$ denotes the set of all groundings of an attribute $\mathtt{A} \in \mathbf{A}$ in $\mathbf{A}^{\Delta}$:

- **Structural homogeneity:** All grounded attributes $\mathtt{A}[\mathbf{x}] \in \mathtt{A}^{\Delta}$ of the same attribute $\mathtt{A} \in \mathbf{A}$ share the same structural equation and, hence, the same conditional probability distribution in equation (10).

For instance, in Example 3.4, we assume that all groundings of type *Prestige* have the same structural equations.

The structural homogeneity assumption, however, is not easily captured because different groundings of the same attribute can have different number of parents. For instance, consider the atoms *Score["s1"]* and *Score["s2"]* from equation (9). *Score["s1"]* has two *Prestige* parents (since it has two authors, *"Eva"* and *"Bob"*), whereas *Score["s2"]* has one *Prestige* parent (*"Eva."*). We address this issue by using another layer of aggregate functions, that we call *embeddings*, ψ, and change Equation (10) to

$$\mathtt{Pr}\Big(\mathtt{A}[\mathbf{x}] \mid \Psi^{\mathtt{A}}(\mathbf{Pa}(\mathtt{A}[\mathbf{x}]))\Big) \tag{17}$$

where $\Psi^{\mathtt{A}}$ is a collection of mappings that projects the parents of $\mathtt{A}[\mathbf{x}]$ into a low-dimensional vector with fixed dimensionality for all $\mathtt{A}[\mathbf{x}] \in \mathtt{A}^{\Delta}$. Intuitively, we assume that the mappings provide sufficient statistics for evaluating the underling structural questions corresponding to all $\mathtt{A}[\mathbf{x}] \in \mathtt{A}^{\Delta}$. More formally, we assume that $\Psi^{\mathtt{A}}$ is known and consists of a set of mappings $\{\psi^{\mathtt{A}}_{\mathtt{A}_1}, \psi^{\mathtt{A}}_{\mathtt{A}_2}, \ldots\}$, one for each type of attribute $\mathtt{A}_j$ occurring on the RHS of a rule (4), where each $\psi^{\mathtt{A}}_{\mathtt{A}_j}$ is an *embedding function* that maps the set of values of all parents of type $\mathtt{A}_j$ into a low-dimensional *embedding space* with fixed dimensionality. The embedding function can be a simple aggregate like average; other types of embeddings are discussed in Section 5.2.2.

Example 4.1. Consider the three nodes of type Score for "s_1", "s_2", "s_3" in Figure 3, and consider their parents of type Prestige. The number of their parents is 2 (for "s_1" – "Bob" and "Eva" with vector $\langle 1, 1\rangle$ for prestige), 1 (for "s_2" – "Eva" with vector $\langle 1\rangle$), and 2 (for "s_3" – "Eva" and "Carlos" with vector $\langle 1, 0\rangle$) respectively (the prestige values of the authors are in Figure 1), but under the homogeneity assumption, the conditional probability of scores given prestige of authors would be computed by the same function by using a mapping $\psi^{Score}_{Prestige}$ to aggregate the vectors of Prestige *values; we discuss choices for this aggregate function in Section 5.2.2.*

To summarize, the grounded causal graph $G(\Phi^{\Delta})$ defined by a relational causal model defines a joint probability distribution given by:

$$\mathtt{Pr}(\mathbf{A}^{\Delta}) = \prod_{\mathtt{A}[\mathbf{x}] \in \mathbf{A}^{\Delta}} \mathtt{Pr}\Big(\mathtt{A}[\mathbf{x}] \mid \Psi^{\mathtt{A}}(\mathbf{Pa}(\mathtt{A}[\mathbf{x}]))\Big) \tag{18}$$

In some scenarios, the structural homogeneity assumption may not hold, for instance, the structural equations for single-blind and double-blind conferences can be different. Such situations can be expressed in CaRL by adding multiple rules at *different granularities* in which the structural homogeneity assumption is perceived to hold, e.g.,

$$\text{SBlind\_Score}[S] \Leftarrow \text{Quality}[S] \text{ WHERE Submission}(S)$$
$$\text{DBlind\_Score}[S] \Leftarrow \text{Quality}[S] \text{ WHERE Submission}(S)$$

4.2 Treated and Response Units

In standard causal analysis, the units can be considered tuples in a single unit table, with one attribute corresponding to the treatment and another attribute corresponding to the response. For instance, in the schema given in Figure 1 and relational causal graph in Figure 3, one could analyze the causal effect of qualification of authors on their prestige, and then the 'authors' form both the treated and response units. In contrast, for multi-relational causal analysis in CaRL, when one analyzes the causal effect of prestige of authors on scores of submissions, then intuitively the authors form the treated units and the submissions form the response units. Even when authors (or submissions) form both the treated and response units, CaRL allows inclusion of additional attributes from other relations that are covariates and required for answering causal queries (see Section 5.1). Next we formally define these concepts.

In relational causal analysis, we are given a *treatment attribute function* $T[\mathbf{X}] \in \mathbf{A}$ and a *response attribute function* $Y[\mathbf{X}'] \in \mathbf{A}$; The set of *units* $\mathbb{U}_T$ (resp. $\mathbb{U}_Y$) denotes the entities or relationships corresponding to the treatment (resp. response) attribute function T (resp. Y). For example, to study the effect of authors' prestige on submission scores, $Prestige[A]$ is the treatment attribute function and $Score[S]$ is the response attribute function, $\mathbb{U}_{Prestige}$ denotes all authors as treated units and $\mathbb{U}_{Score}$ denotes all submissions as response units (we assume without loss of generality that the attribute function names are unique and correspond to a single entity or relationship). We assume the treatment attribute has binary values whereas the response can be any real number.

Given a set of treated units $\mathbb{U}_T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ and a binary vector $\vec{t} = (t_1, t_2, \ldots)$, we are interested in the effect of a set of interventions $\mathtt{do}(T(\mathbf{x}_i) = t_i)$ for all treated units x_i, where

each intervention replaces the NSE associated with $T(\mathbf{x}_i)$ with a constant t_i. In our example of the effect of prestige on score, the vector $\vec{t}$ corresponds to a particular assignment of prestige to all authors, e.g., the vector $\vec{1}$ identifies an intervention that *hypothetically changes 'all' authors' affiliations to prestigious ones*. By abuse of notation, we denote with $\mathsf{do}(T[\mathbb{S}] = \vec{t}_{\mathbb{S}})$ a set of interventions in which an arbitrary subset of treated units $\mathbb{S} \subseteq \mathbb{U}_T$ receive $\vec{t}_{\mathbb{S}}$ (with an implicit assumption on the order of elements in the set $\mathbb{S}$). Having treated/response units and the treatment vectors allows us to have (1) *non-uniform units* that may be different entities or relationships, and (2) *different types of treatments*, e.g., forcing all authors to be of prestigious institutions as $\vec{1}$ vs. one or some of the authors from prestigious institutions as $(1, 0, 0, \cdots)$. CaRL aims to answer causal queries that compare the *average response* of the response units $\mathbb{U}_Y$ to two alternative intervention strategies $\vec{t}$ and $\vec{t}'$ applied to the treated units $\mathbb{U}_T$, which we discuss next.

4.3 Relational Paths and Peers

Before we can formalize the semantics of causal queries described in Section 3.3, especially for the isolated and relational effects, we need to establish a one-to-one correspondence between treated and response units by using aggregations carefully. To this end, we first define relational paths.

Definition 4.2. A relational path *is a sequence of entities and relationships of the following form:*

$$\mathcal{P} : E_1(X_1) \xleftrightarrow{R_1(X_1,X_2)} E_1(X_2) \cdots E_{\ell-1}(X_{\ell-1}) \xleftrightarrow{R_{\ell-1}(X_{\ell-1},X_\ell)} E_\ell(X_\ell) \tag{19}$$

where $E_i(X_i) \in \mathbf{E}$ *and* $R_{i-1}(X_{i-1}, X_i) \in \mathbf{R}$*, for* $i = 1, \cdots, \ell$.

For instance, $\text{Conference}(C) \xleftrightarrow{\text{Submitted}(S,C)} \text{Submission}(S)$ is a relational path in our example. The treated and response units corresponding to treatment and response attribute functions T and Y are said to be *relationally connected* if there exists a relational path $\mathcal{P}$ that includes the entities or relationships for T and Y either as the endpoints in the path or as the labels of the edges at the ends of the path. For example, for $T[\mathbf{X}]$ = Prestige[A] and $Y[\mathbf{X}']$ = Score[S], the treatment is an attribute function of the entity $\text{Author}(A)$, the response is an attribute function of the relationship Author(A,S), and the treated and response units are relationally connected by the following relational path:

$$\text{Author}(A) \xleftrightarrow{\text{Author}(A,S)} \text{Submission}(S) \tag{20}$$

In this paper, we make the natural assumption that the treated and response units are relationally connected by at least one relational path as otherwise the effect of treatment on the response is not meaningful. These units can then be unified using the aggregated response $\text{AGG\_}Y[\mathbf{X}]$ defined with the following aggregate rule (see Section 3.2.4) that maps attribute Y of response units $\mathbb{U}_Y$ to treatment units $\mathbb{U}_T$, where the units can be either entities or relationships.

$$\text{AGG\_}Y[\mathbf{X}] \Leftarrow Y[\mathbf{X}'] \text{ WHERE } R_1(X_1, X_2), \ldots, R_{\ell-1}(X_{\ell-1}, X_\ell) \tag{21}$$

For example, to unify the treated and response units associated to $T[\mathbf{X}]$ = Prestige[A] and $Y[\mathbf{X}']$ = Score[S], the aggregate rule[5] associated with the relational path in (20) coincides with (12): $\text{AVG\_}Score[A] \Leftarrow Score[S]$ WHERE $\text{Author}(A, S)$.

Therefore, we assume from here on that the response units $\mathbb{U}_Y$ are the same as the treated unit $\mathbb{U}_T$. Henceforth, we simply refer to elements of $\mathbb{U}_Y$ and $\mathbb{U}_T$ as *units* and denote them with $\mathbb{U}_{(T,Y)} = \mathbb{U}_T = \mathbb{U}_Y$. In our example, after the unification, the $AVG\_Score[A]$ can be considered as a new attribute function of authors (as in a 'view' in relational databases), and the authors form $\mathbb{U}_{(T,Y)}$.

Relational peers. Next, we define the notion of relational peers of a unit, which is central to the notion of relational and isolated effects. Recall that the grounded causal graph $G(\Phi^\Delta)$ is extended with vertices and edges corresponding to aggregated attributes as discussed in Section 3.2.4.

Definition 4.3. Given a treated attribute function $T[\mathbf{X}]$*, and a (possibly aggregated) response attribute function* $Y[\mathbf{X}]$*, we define the* relational peers *of a unit* $\mathbf{x} \in \mathbb{U}_{(T,Y)}$ *as a set of units* $\mathbb{P}(\mathbf{x}) \subseteq \mathbb{U}_{(T,Y)} - \{\mathbf{x}\}$ *such that for each* $\mathbf{p} \in \mathbb{P}(\mathbf{x})$*, there exists a directed path from* $T[\mathbf{p}]$ *to* $Y[\mathbf{x}]$ *in* $G(\Phi^\Delta)$.

For example, in Figure 4, treatment and aggregated response functions $Prestige[A]$ and $\text{AVG\_}Score[A]$, $\mathbb{P}(\text{“}Bob\text{”}) = \{\text{“}Eva\text{”}\}$ and $\mathbb{P}(\text{“}Eva\text{”}) = \{\text{“}Bob\text{”}, \text{“}Carlos\text{”}\}$. In practice, the relational causal model is expected to form relational peers $\mathbb{P}(\mathbf{x})$ that consist only of units that are in some *relational proximity* of $\mathbf{x}$, e.g., authors from the same institution, same research interests, etc.[6]

The following quantity measures the expected response of a unit $\mathbf{x} \in \mathbb{U}_{(T,Y)}$ when it receives the treatment t, and its relational peers receive the vector of treatments $\vec{t} = (t_1, t_2 \ldots)$.

$$Y_{\mathbf{x}}(t, \vec{t}) \stackrel{\text{def}}{=} \mathbb{E}[Y[\mathbf{x}] \mid \mathsf{do}(T[\mathbf{x}] = t), \mathsf{do}(T[\mathbb{P}(\mathbf{x})] = \vec{t})] \tag{22}$$

In this paper, we assume $\mathsf{do}\big(T[\mathbb{P}(\mathbf{x})] = \vec{t}\big)$ is a *well-defined intervention* for all units $\mathbf{x}$, *i.e.*, it uniquely determines which relational peers of a unit would receive which treatment. For instance, this holds if $\mathbb{P}(\mathbf{x})$ is of the same size for all $\mathbf{x}$, and it either has a natural ordering or is ordering-invariant. However, we allow several relaxations on the size and type on $\vec{t}$ in our framework as discussed later.

[5] We aggregate the response and not the treatment since aggregating treatments may lead to interventions that are not well defined.

[6] This assumption is far less strict than than the assumption of partial interference, which is standard in statistics, to extend Rubin's causality to handle interference [54]. Also note that the assumption of no interference (or SUTVA) [43] translates to the statement $\mathbb{P}(\mathbf{x}) = \emptyset$ for all $\mathbf{x} \in \mathbb{U}_{(T,Y)}$ in relational causal models.

4.4 Semantics of Causal Queries

In this section, we define the semantics of causal queries outlined in Section 3.3 in terms of intervention; how these causal queries are answered in CaRL using unification of treated and response units, embeddings, and selection of covariates is discussed in Section 5.

4.4.1 Average treatment effect queries. The primary causal query in CaRL is average treatment effect (ATE) query of the form $Y[\mathbf{X}'] \Leftarrow T[\mathbf{X}]$? as given in (13). Given treatment and response attribute functions T, Y, ATE is defined as follows:

$$\text{ATE}(T,Y) \stackrel{\text{def}}{=} \sum_{\mathbf{x}' \in \mathbb{U}_Y} \frac{1}{m} (\mathbb{E}[Y[\mathbf{x}'] \mid \mathsf{do}(T[\mathbb{U}_T] = \vec{0})] - \mathbb{E}[Y[\mathbf{x}'] \mid \mathsf{do}(T[\mathbb{U}_T] = \vec{1})]) \tag{23}$$

Intuitively, ATE compares the expected response of the response units in two regimes of intervention: one in which all units receive treatment and another where none do. For example, ATE(Prestige, Score) compares scores of submissions under two interventions in which all authors are and are not affiliated with prestigious institutions.

4.4.2 Aggregated response queries. Aggregate response queries of the form $\text{AGG\_}Y[\mathbf{X}'] \Leftarrow T[\mathbf{X}]$? as given in (14) is defined similar to ATE above, where we replace Y with AGG_Y everywhere. Note that in the extended relational causal graphs, there are nodes corresponding to AGG_Y as shown in Figure 4.

4.4.3 Relational and isolated effects queries. The CaRL query (15) computes the following three quantities, which compare the average isolated (AIE), relational (ARE), and overall (AOE) effects of two alternative intervention strategies $(t, \vec{t})$ and $(t', \vec{t}')$ over n response units[7].

$$\text{AIE}(t; t' \mid \vec{t}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t}) \tag{24}$$

$$\text{ARE}(\vec{t}; \vec{t}' \mid t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t, \vec{t}') \tag{25}$$

$$\text{AOE}(t, \vec{t}; t', \vec{t}') \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t}') \tag{26}$$

Intuitively, the isolated causal effect of a treatment fixes the treatment of the relational peers of a unit and compares its expected response under two treatment strategies assigned to the unit. On the other hand, the relational causal effect of a treatment fixes the treatment of a unit $\mathbf{x}$ and compares its expected response under two treatment strategies assigned to its relational peers. For example, the relational effect of $\textit{Prestige}[A]$ on $\text{AVG\_}\textit{Score}[A]$ fixes the prestige of an author such as "*Bob*" and compares the counterfactual response $\text{AVG\_}\textit{Score}[\text{"}\textit{Bob}\text{"}]$ under two regimes of interventions in which the relational peers of "*Bob*", e.g., "*Eva*", receive two different treatment strategies, e.g., all of them have prestigious affiliations versus none of them having such affiliations. Note that the overall causal effect is an extension of ATE (23) for two arbitrary treatment strategies. Indeed, ATE coincides with $\text{AOE}(1, \vec{1} \mid 0, \vec{0})$ when the treated and response units are unified. The following proposition shows the connection between relational, isolated and overall effects (we omit the proof due to lack of space).

Proposition 4.4. *The average overall effect can be decomposed into the average isolated and average relational effects, as follows:*

$$\begin{aligned} AOE(t, \vec{t}; t', \vec{t}') &= AIE(t, t' \mid \vec{t}) + ARE(\vec{t}, \vec{t}' \mid t') \\ &= AIE(t, t' \mid \vec{t}') + ARE(\vec{t}, \vec{t}' \mid t) \end{aligned} \tag{27}$$

5 ANSWERING CAUSAL QUERIES

Given the syntax of different causal queries in Section 3.3 and their semantics in Section 4.4, now we describe how we answer these queries in CaRL. The query answering component of CaRL consists of *covariate detection* (Section 5.1) and *covariate adjustment* (Section 5.2). The goal of covariate detection is to identify a sufficient set of covariates that should be adjusted for to remove confounding effects. Then, in the process of covariate adjustment, the data is transformed into a flat, single-table format so that causal inference can be performed using standard methods.

5.1 Covariate Detection

Given treatment and response attribute functions $T[\mathbf{X}]$ and $Y[\mathbf{X}']$, to answer all types of causal queries defined in Section 4.4, we need to estimate the effect of interventions of the form $\mathsf{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)$ on a set of treated units $\mathbb{S} \subseteq \mathbb{U}_T$, on a response unit $\mathbf{x}' \in \mathbb{U}_Y$. This section proves a graphical criterion to select a sufficient set of covariates from a relational causal graph $G(\Phi^{\Delta})$ that enable the estimation of quantities of the form $\mathbb{E}\left[Y[\mathbf{x}'] \mid \mathsf{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\right]$ and thereby the queries in Section 4.4. For this purpose we use the extended relational causal graph as shown in Figure 4 to map possibly varying number of parent nodes to a fixed and smaller dimension by adopting the idea of embedding functions introduced in Section 4.1. We illustrate this with an example below.

Example 5.1. In Example 4.1, $\psi^{\text{Score}}_{\text{Prestige}}(S)$ now corresponds to a new attribute of a submission that maps the Prestige attribute of *Author*s of that submission. Figure 5 shows the relational causal graph with augmented attributes computed using the mapping functions or embeddings represented by the triangles.

[7] We do not need the treatment vectors $\vec{t}, \vec{t}'$ applied to the peers to have the same size although they are applied to all units $\mathbf{x}$. We also do not need all units $\mathbf{x}$ to have the same number of peers in $\mathbb{P}(\mathbf{x})$. As the grammar defined in (16) describes, we can assign treatments to "at least/most k or k%" neighbors, and that is well-defined for all units $\mathbf{x}$ even if they do not have the exact same number of peers in $\mathbb{P}(\mathbf{x})$. On the other hand, for such conditions, we do need to assume that the effects of interventions to the peers are *ordering-invariant*, e.g., the intervention can be applied to any of the k peers (with possible truncations for smaller peer sets) in $\mathbb{P}(\mathbf{x})$.

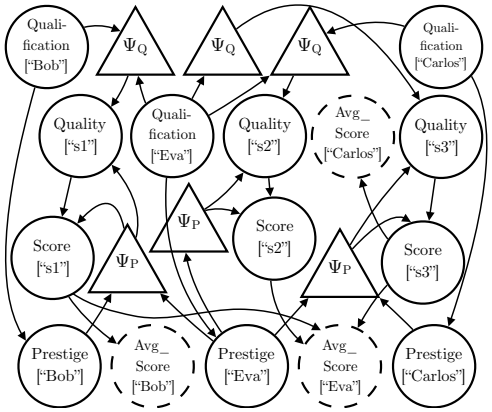

Figure 5: Final relational causal graph obtained by (further) augmenting the graph in Figure 4 with embedding functions. For clarity, $\psi^{\textbf{Quality}}_{\textbf{Qualifications}}[S]$ is represented as ψ_Q, and $\psi^{\textbf{Quality}}_{\textbf{Prestige}}[S]$, $\psi^{\textbf{Score}}_{\textbf{Prestige}}[S]$ as ψ_P.

.

The following theorem formalizes how the do-operator for relational causal graph can be estimated from observed data. This theorem uses the concept of *d-separation* from conditional independence in graphical models [35], denoted by $X \perp\!\!\!\perp Y|_G Z$. The review of these concepts and the proof of the theorem is deferred to the full version [49] of the paper due to lack of space.

Theorem 5.2 (Relational Adjustment Formula). *Given an augmented relational causal graph $G(\Phi^{\Delta})$, treatment and response attribute functions T, Y, and a set of treatment units (entities or relationships) $\mathbb{S}$ and their treatment assignment vector $\vec{t}_{\mathbb{S}}$, we have the following* relational adjustment formula*:*

$$\mathbb{E}\left[Y[\mathbf{x}'] \mid \mathsf{do}(T[\mathbb{S}] = \vec{t}_{\mathbb{S}})\right] = \sum_{\mathbf{z} \in Dom(\mathbf{Z})} \mathbb{E}\left[Y[\mathbf{x}'] \mid \mathbf{Z} = \mathbf{z}, T[\mathbb{S}'] = \vec{t}_{\mathbb{S}'}\right] \Pr(\mathbf{Z} = \mathbf{z}) \quad (28)$$

where $\mathbb{S}' \subseteq \mathbb{S}$ is such that, for each unit $\mathbf{x} \in \mathbb{S}'$, there exists a directed path from the node $T[\mathbf{x}]$ to the node $Y[\mathbf{x}']$ in $G(\Phi^{\Delta})$, and $\mathbf{Z}$ is set of nodes in $G(\Phi^{\Delta})$ corresponding to the groundings of a subset of observed attribute functions $\mathbf{A}_{Obs}$ such that:

$$Y[\mathbf{x}'] \perp\!\!\!\perp \left(\bigcup_{\mathbf{x} \in \mathbb{S}} \mathbf{Pa}(T[\mathbf{x}])\right) |_{G(\Phi^{\Delta})} \left(\bigcup_{\mathbf{x} \in \mathbb{S}} T[\mathbf{x}],\ \mathbf{Z}\right) \quad (29)$$

Further, choosing Z to be the parent nodes of $\mathbb{S}'$ in $G(\Phi^{\Delta})$ always satisfies (29) as a sufficient condition.

(Intuitively, it is always sufficient to condition for the 'parents' of treated units as they separate them from the rest of the graph ensuring independence.) Here we illustrate with an example.

Example 5.3. To compute ATE(Prestige, Score) in our example, we need to compute expectations of the form

$$\mathbb{E}\left[Score[s] \mid \mathsf{do}\left(Prestige[\{\text{“}Bob\text{”}, \text{“}Eva\text{”}, \text{“}Carlos\text{”}\}] = \vec{t}\right)\right] \text{ for } \vec{t} \in \{\vec{0}, \vec{1}\} \quad (30)$$

Algorithm 1: Constructing a unit table.

Input: An augmented relational causal graph $G(\Phi^{\Delta})$, treated and outcome attribute functions $T[\mathbf{X}]$ and $Y[\mathbf{X}']$.
Output: The unit table $\mathbf{D}(Y, \psi_T, \psi_{\mathbf{Z}})$.

1 **for** $\mathbf{x}' \in \mathbb{U}_Y$ **do**
2 $\quad \mathbb{U}'_T \leftarrow$ A minimal subset of $\mathbb{U}_T$ such that there exits a directed path in $G(\Phi^{\Delta})$ from $T[\mathbf{x}]$ to $Y[\mathbf{x}']$ for all $\mathbf{x} \in \mathbb{U}'_T$
3 $\quad \mathbf{Z} \leftarrow$ A minimal set of vertices in $G(\Phi^{\Delta})$ that satisfies the d-separation statement in Eq (29)
4 $\quad \psi_T \leftarrow \psi^Y_T(\langle T[\mathbf{x}_1], \ldots, T[\mathbf{x}_{|\mathbb{U}'_T|}]\rangle)$
5 $\quad \Psi_{\mathbf{Z}} \leftarrow \Psi^{\mathbf{Y}}_Z(\mathbf{Z})$
6 $\quad$ Insert the tuple $(Y[\mathbf{x}], \psi_T[\mathbf{x}], \psi_{\mathbf{Z}}[\mathbf{x}]))$ to unit table $\mathbf{D}$

where we intervene on all three authors in the example. By applying Theorem 5.2 for submission $s =$ "s_1", note that directed paths to Score["s_1"] exists only from Prestige["Bob"] and Prestige["Eva"], which form the subset $\mathbb{S}'$. Further, it is sufficient to condition on the parents of these two Prestige nodes, i.e., $\mathbf{Z}$ =\{Qualifications["Bob"], Qualifications["Eva"]\}. Therefore, (30) reduces to:

$$\sum_{\mathbf{z} \in Dom(\mathbf{Z})} \mathbb{E}\left[Score[\text{“}s_1\text{”}] \mid \mathbf{Z} = \mathbf{z}, Prestige[\{\text{“}Bob\text{”}, \text{“}Eva\text{”}\}] = \vec{t}\right] \Pr(\mathbf{Z} = \mathbf{z}) \quad (31)$$

Similarly, for $s =$ "s_2" and $\mathbf{Z} = \{Qualifications[\text{“}Eva\text{”}]\}$, we obtain

$$\sum_{\mathbf{z} \in Dom(\mathbf{Z})} \mathbb{E}\left[Score[\text{“}s_2\text{”}] \mid \mathbf{Z} = \mathbf{z}, Prestige[\{\text{“}Eva\text{”}\}] = t\right] \Pr(\mathbf{Z} = \mathbf{z}) \quad (32)$$

Note that the relational adjustment formula in (28) controls *for an adequate set of covariates Z that* confound *the causal effect of a treatment on an outcome (Z is called the set of confounding covariates or covariates). For example, the causal effect of Prestige on Score is confounded by Qualifications. This is because, qualified researchers are likely to belong to prestigious universities and qualified researchers are more likely to submit high quality papers. Therefore, to compute the ATE of Prestige on Score we need to control for author's qualifications as in (31). For estimating ATE(Quality, Score) (assuming quality is observed) by applying (29) we find that for each submission s, $\mathbb{E}\left[Score[s] \mid \mathsf{do}(Prestige[\mathbb{U}_T] = \vec{1})\right]$ can be estimated by adjusting for the embedded attribute functions $\mathbf{Z}$ = $\{\psi^{Score}_{Prestige}[s], \psi^{Score}_{Qualifications}[s]\}$.*

To estimate ATE(Quality, AVG_Score) (the effect on average acceptance rate of an author), on the other hand, we need to estimate $\Pr\left(\text{AVG\_Score}[A] = y \mid \mathsf{do}(\text{Prestige}[\mathbb{U}_T] = \vec{1})\right)$, for each author. According to Equation (29), this can be done by adjusting for the joint distribution of the qualifications of *all* their coauthors, which is potentially very high-dimensional, and therefore we need another round of embeddings to aggregate that information as discussed next.

Unit	Outcome (Y)	Embedded coauthors' treatments (ψ_T^Y)	Embedded Collaborators' Covariates(Ψ_Z^Y)	
Author ID	AVG_Score	Prestige (AVG)	Centrality (COUNT)	H-index (AVG)
Bob	0.75	1	1	2
Carlos	0.1	1	1	2
Eva	0.41	0.5	2	35

Table 1: The unit table for $T[\mathbf{X}]$ = Prestige$[A]$ and $Y[\mathbf{X}']$ = AVG_Score$[A]$ based on Figure 1.

5.2 Covariate Adjustment

There are two challenges in estimating the causal queries in Section 4.4 using the relational adjustment formula (28): (1) when the set of confounding covariates $\mathbf{Z}$ has high dimensionality, estimating the conditional expectation in (28) from data is challenging, and (2) the causal queries need to compute *averages* across all response units. Hence, we need to estimate the formula (28) separately for each response unit that is not feasible. For instance, in Example 5.3, (31) and (32) need to be estimated separately.

To address these issues, similar to Section 4.1, we use a set of embedding functions ψ_T^Y and $\Psi_\mathbf{Z}^Y$ to project the treatment and covariate vectors, respectively, into a low-dimensional embedding space with *fixed dimensionality* and for all response units. This enables us to transform a (multi-) relational instance to a single low-dimensional flat table.

5.2.1 Unit table. In the classical causal inference framework model discussed in Section 2, the units of interest are stored in a single unit table with attributes corresponding to the treatment, response, and confounding covariates as the columns. Here we generalize this concept to capture units in relational causal analysis.

Given a relational causal graph $G(\Phi^\Delta)$ and treatment and response attribute functions $T[\mathbf{X}]$ and $Y[\mathbf{X}']$, we use Algorithm 1 to construct a *unit table*, which is a standard relation (table) with schema $\mathbf{D}(Y, \psi_T^Y, \Psi_\mathbf{Z}^Y)$ (note that $\Psi_\mathbf{Z}^Y$ denotes a vector of values for possible multiple covariates $\mathbf{Z}$). It consists of tuples $(Y[\mathbf{x}'], \psi_T^Y[\mathbf{x}'], \Psi_\mathbf{Z}^Y[\mathbf{x}'])$ for each response unit $\mathbf{x}' \in \mathbb{U}_Y$, where $\psi_T^Y[\mathbf{X}']$ and $\Psi_\mathbf{Z}^Y[\mathbf{X}']$ (with abuse of notation) are relational embedded attribute functions that correspond to the result of applying ψ_T^Y and $\Psi_\mathbf{Z}^Y$ to the treatment and covariate vectors respectively.

Example 5.4. Table 1 shows the unit table corresponding to $T[\mathbf{X}]$ = Prestige$[A]$ and $Y[\mathbf{X}']$ = AVG_Score$[A]$. Here Authors *constitute the response units and the aggregated response is an attribute of authors. In this table simple mappings such as average and count are used for embedding. Note that Table 1 also serves as the unit table for $T[\mathbf{X}]$ = Prestige$[A]$ and $Y[\mathbf{X}']$ = Score$[A]$. In this case since the treated and response units are different CaRL uses the aggregated response AVG_Score$[A]$ for unification (see Section 4.3).*

By rewriting the RHS of the relational adjustment formula (28) in terms of the attributes of the unit table and $\vec{t}^e_{\mathbb{S}'}$ the embedded representation of the treatment assignment $\vec{t}_{\mathbb{S}'}$, (*i.e.*, $\vec{t}^e_{\mathbb{S}'} = \psi_T^Y(\vec{t}_{\mathbb{S}'})$), we obtain

$$\sum_{\mathbf{z} \in Dom(\Psi_\mathbf{Z}^Y)} \mathbb{E}\left[Y \mid \Psi_\mathbf{Z}^Y = \mathbf{z}, \psi_T^Y = \vec{t}^e_{\mathbb{S}'}\right] \Pr(\Psi_\mathbf{Z}^Y = \mathbf{z}) \tag{33}$$

Once we have a flat unit table with columns for treatment, response, and covariates as in Section 2, the causal queries in Section 4.4 can be estimated using (33) by applying the standard approaches to causal analysis like regression [2] (the conditional expectation in (33) is a regression function) or matching methods [11, 15, 18] (matching treatment and control units with the same/similar values). The validity of treatment effect estimates is conditional on the assumption that the background knowledge is accurate.

5.2.2 Choice of embedding functions. Embedding as a technique addresses both the issues of the high dimensionality and the variable size of the treatments and covariates that correspond to the response units, thereby making the estimation of causal queries more convenient. However, (33) only approximates (28), hence the quality of the answers depends on whether the embeddings preserve sufficient statistics. In this work, we use the following natural choices of embeddings, a formal study of the choices of embedding functions in multi-relational causal analysis is an interesting direction of future work. (1) *Mean* and *median:* Uses basic aggregation functions, such as mean and median, together with the cardinality of the vectors (to account for the underlying topology of the relational skeleton, e.g, number of authors or collaborators). (2) *Padding:* Pads each variable size vector with out-of-band "empty markers" to make create same-sized vectors to use directly as the embedding. (3) *Moments:* Uses a vector consists of k moments (*i.e.*, mean, variance, skewness, etc.), where k is chosen to minimize response prediction loss.

6 EXPERIMENTS

In this section, we conduct an experimental evaluation of CaRL, addressing three questions. **End-to-end performance**: is CaRL effective in answering causal queries on relational data? Can it avoid simply discovering correlations instead of true causation? Can it distinguish isolated effects from relational effects? **Quality of estimates**: when ground truth is available, can CaRL recover the true treatment effects? And is the relational structure necessary for recovering the correct treatment effect? **Sensitivity to embeddings**: how sensitive is CaRL's performance to the choice of the type of embedding strategy?

Experimental setup. The experiments were performed locally on a 64-bit Linux server with 1TB RAM and 4 Intel Xeon processors with 15 cores @ 2.8GHz each.

Dataset	Tables [#]	Att. [#]	Rows [#]	Unit Table Cons.	Query Ans.
MIMIC-III	26	324	400M	6h	4.5h
NIS	4	280	8M	4m	30s
ReviewData	3	7	6K	10.6s	1.2s
Synthetic ReviewData	3	7	300K	17.2s	1.3s

Table 2: Data description and query runtime.

6.1 Datasets

We used three real datasets, two from the medical domain, and one about conferences, summarized in Table 2. All datasets contain interesting relationships that inform CaRL's causal analysis. In addition, we generated a synthetic dataset in order to have control over the ground truth.

MIMIC-III. The Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database is a large-volume, multi-parameter dataset collected from the ICUs of Beth Israel Deaconess Medical Center from 2008 to 2014 representing 38,597 adult patients, 58,976 hospital admissions [19]. There are 26 tables with 400M rows and 324 attributes (see Table 2), which include patients' information like demographics, length of stay, medications, laboratory test results and health insurance data. We specified the following causal model in CaRL, where Eth = ethnicity, Pa = patient, Diag = diagnosis, Doc = doctor, and Len = length of stay:

$$\begin{aligned}
\text{SelfPay}[P] &\Leftarrow \text{Eth}[P], \text{Religion}[P], \text{Sex}[P] \text{ WHERE Pa}(P)\\
\text{Diag}[P] &\Leftarrow \text{Eth}[P], \text{Religion}[P], \text{Sex}[P] \text{ WHERE Pa}(P)\\
\text{Dose}[D] &\Leftarrow \text{Diag}[P], \text{Severe}[P], \text{Doc}[C] \text{ WHERE Drug}(C, D), \text{Care}(C, P)\\
\text{Death}[P] &\Leftarrow \text{Len}[P], \text{Diag}[P], \text{Dose}[D], \text{Doc}[C]\\
&\quad \text{WHERE Care}(C, P), \text{Given}(D, P)\\
\text{Len}[P] &\Leftarrow \text{Dose}[D], \text{Diag}[P] \text{ WHERE Given}(D, P)
\end{aligned}$$

NIS. The Nationwide Inpatient Sample (NIS) [14] is a dataset of hospital stays across the US, produced by the Department of Health and Human Services once annually. We use the sample for the year 2006, which comprises 8 million hospital admissions across 1035 hospitals. Each admission is associated with a hospital and the patient's demographic information, admission source, health history, performed procedures, and new diagnoses. Information available about each hospital includes size, location, and ownership. We specified a casual model in CaRL using 16 intuitive causal rules, using attributes whether the hospital is classified as large [1], patient's medical bill, etc.; we mention a few below:

$$\begin{aligned}
\text{Bill}[P] &\Leftarrow \text{Illness_Severity[P]}\\
\text{Bill}[P] &\Leftarrow \text{Private_Ownership[H] WHERE Admitted}(P, H)\\
\text{Bill}[P] &\Leftarrow \text{Surgery_Performed[P]}\\
\text{Admitted_to_large}[P] &\Leftarrow \text{Illness_Severity[P]}
\end{aligned}$$

ReviewData. ReviewData consists of 2,075 papers submitted for review between 2017 and 2019 at 10 computer science conferences and workshops, which have acceptance rates between 40%–84%. Each submission is associated with a number of referee reviews and an acceptance or rejection decision. About half of all submissions are double-blind, while the other half reveal author names to the reviewer. All submissions were unblinded after the conferences concluded. The dataset also contains an authors table, with the citation count, h-index, publishing experience (in years), and university ranking for each of the 4490 authors who contributed to a paper in the dataset. ReviewData was built by scraping, cleaning and normalizing data from OpenReview [33], Scopus [50] and the Shanghai University Rankings [39]. Scraping Scopus was done using the tool proposed in [40]. We plan to make ReviewData publicly available.

Synthetic ReviewData. We generated Synthetic ReviewData mimicking the probability distributions observed in ReviewData. The relational skeleton was generated keeping in mind the correlations we observed in the real data, *e.g.*, authors with high productivity tend to be affiliated with more prestigious institutions, and authors from more prestigious institutions tend to collaborate with each other more. However, for each paper we let the number of authors and each submission's choice of venue be determined randomly. We generated 10,000 authors with affiliations to 200 different institutions, along with 75,000 papers submitted to 100 different venues. Next, we generated two datasets to explore CaRL's performance with and without relational effects. The first dataset had a treatment effect of prestige on review score of 0 for double-blind and 1 for single-blind venues, for all submissions. In the second dataset, the isolated effects stay the same for both double- and single-blind venues while there is a constant effect of 1/2 on the review score of each submission if authors' collaborators are prestigious.

6.2 End-to-end Results

In this section we used CaRL to answer several causal queries (including all kinds defined in Section 3.3) on the real datasets and evaluated their quality. Since we do not have ground truth for this data, we discuss which results are more in agreement with the intuition or the literature in the field. We also compared CaRL's answer with more naive, correlation-based answers. All runtimes for these experiments are reported in Table 2.

MIMIC-III. We asked the following causal queries: what is the effect of not having health insurance on the mortality rate? And what is the effect of non-insurance on the length-of-stay (in the hospital)?

$$(a)\ \text{Death}[P] \Leftarrow \text{SelfPay}[P]? \qquad (b)\ \text{Len}[P] \Leftarrow \text{SelfPay}[P]? \qquad (34)$$

The treated and control groups consist of patients without

insurance (self-payers) and with insurance respectively. Table 3 shows the results for both the average treatment effect (ATE) and the naive difference of averages between the two groups. Computed naively, there is a significant difference in both mortality rate and the length of stay between insured and non-insured groups. However, after adjusting for confounders and mediators, we observe that there is almost no effect on mortality rate; in other words, care givers do not discriminate in treating patients with and without insurance. The discrepancy is due to the fact that self-payers tend to defer checking into a hospital until the problem is severe. The treatment effect on the length of stay is also attenuated compared to the estimated difference between the average of the outcomes for treated and control groups.

Causal Query	Avg. of Treated	Avg. of Control	Diff. of Averages	ATE
MIMIC 1 (34-a)	15.5%	9.8%	5.7%	0.5%
MIMIC 2 (34-b)	154.23h	244.15h	-89.92h	-26.04h
NIS 1 (35)	64%	31%	33%	-10%

Table 3: The Average Treatment Effect (ATE) compared to naively computing the difference between the averages of the treated and control groups.

NIS. We asked the following causal query: are patients admitted to large hospitals charged more than those admitted to small hospitals? Expressed in CaRL, the query is:

$$\text{AVG\_Bill}[H] \Leftarrow \text{Admitted\_to\_Large}[P]? \tag{35}$$

The treated and control groups are large and small hospitals respectively. As before, we compared the ATE with the naive difference of the average bills of the two groups, and show the results in Table 3. The naive computation shows that the average bill at large hospitals is 33% more likely to be larger per patient (*i.e.,* less affordable). However, when computing the ATE, CaRL adjusts for the profile of the patients each hospital receives, and we obtain a surprising reversal of the trend. The reason for this discrepancy is that patients with more severe (and, thus, more costly) conditions tend to go to large hospitals, while small hospitals tend to have patients with milder conditions. In fact, the medical literature reports that, all else being equal, a larger hospital will provide more affordable treatment than a small one. One meta-analysis [9] reports that economies of scale are present in the healthcare sector and so finds support for the policy of several national governments to consolidate smaller hospitals to increase productivity and efficiency.

ReviewData. We asked two casual queries: what is the effect of an author's *prestige* on the average score of his/her submissions? And what is the effect on the submission score

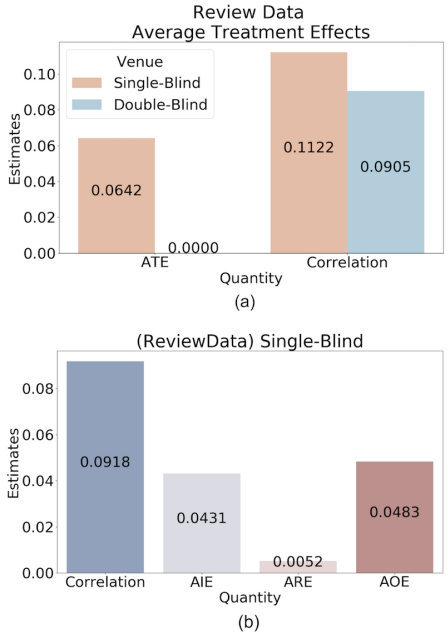

Figure 6: (a) Average treatment effect estimates and Pearson's correlation for single-blind and double-blind submissions, per query in (37) (b) Pearson's correlation, average isolated/relational/overall effect for all authors on submissions in single-blind venues, per query in (36).

when more than 1/3 of her co-authors are treated? Expressed in CaRL, the queries are:

$$\text{AVG\_Score}[A] \Leftarrow \text{Prestige}[A]? \tag{36}$$

$$\text{Score}[S] \Leftarrow \text{Prestige}[A]? \texttt{ WHEN } \textit{MORE THAN } 1/3 \texttt{ PEERS TREATED} \tag{37}$$

We ran each query twice, once on single-blind conferences, and once on double-blind; in CaRL, this is achieved by adding a `where` condition to the queries (not shown here), and computed the ATE in both cases. In addition, we also computed the Pearson correlation between the score distributions of prestigious and non-prestigious authors. The results are shown in Figure 6(a), and show a significant correlation, both for single-blind and double-blind conferences. However, CaRL found that the causal effect of prestige on submission scores was *significant* for single-blind venues, but *not significant* for double-blind venues. A naive interpretation of correlation-as-causation leads to the false conclusion that double blinding is not effective in reducing bias. While the validity of our these findings depend on the validity of the underling assumptions made in this paper, we believe they surpass naive correlation. In particular, we note that our results are in accordance with a series of controlled experiments that suggest double-blind reviewing does indeed reduce institutional prestige bias [32, 41, 52, 55].

		AIE	ARE	AOE
Single-Blind	**Estimated**	1.138	0.434	1.573
	Ground Truth	1.000	0.500	1.500
Double-Blind	**Estimated**	0.101	0.429	0.538
	Ground Truth	0.000	0.500	0.500

Table 4: Averages for isolated, relational and overall effects for Synthetic ReviewData by query in (36).

Method	Embedding	*Single-Blind*		*Double-Blind*	
		Estimated	True	Estimated	True
CaRL	Mean	1.124 ± 0.43	1.00	0.192 ± 0.40	0.00
	Median	1.119 ± 0.36	1.00	0.115 ± 0.37	0.00
	Moment Summary	1.020 ± 0.36	1.00	0.109 ± 0.32	0.00
	Padding	1.011 ± 0.29	1.00	0.013 ± 0.30	0.00
Universal Table	N/A	0.54 ± 0.73	1.00	0.201 ± 0.64	0.00

Table 5: Comparing the sensitivity of the quality of query answer to different choice of embeddings on Synthetic ReviewData, using the query in (37).

Given its primarily networked structure, ReviewData offers a great opportunity to compute peer effects. (In contrast, there are no relational peers for the causal queries on MIMIC-III and NIS.) We computed the effect of prestige across peers on review scores in single-blind conferences, and used CaRL to compute the isolated, the relational, and the overall effects as in (37). Figure 6(b) reveals that the isolated effect (AIE) is more significant than the relational effect (ARE), meaning that an author's own prestige has a stronger effect on his or her average submission score than their collaborators' prestige has, as we might expect. Furthermore, one can verify that we obtained AOE = AIE + ARE, which independently conforms with Proposition 4.4.

6.3 Quality of Estimates

As the ground truth is not known for the real datasets, we use Synthetic ReviewData to evaluate the quality of the estimates CaRL provides. We report estimated and true ATE, ARE, AIE and AOE to scrutinize CaRL's performance. As seen in Table **??**, CaRL is able to disentangle the isolated and relational effects present in Synthetic ReviewData. It is able to do so for both sub-populations, which have different generative rules. The different estimates are correctly recovered, and the property AOE = AIE + ARE from Proposition 4.4 is again respected.

To test the ability to utilize relational structure, we computed the treatment effect estimates to the causal queries (37) using CaRL and compared to propensity score matching on the universal table obtained by joining all base relations. Table **??** compares the estimates by these two approaches with the ground truth. As shown, in all tested cases CaRL approximately recovered the ground truth within a reasonable error bound. However, causal inference on the universal table resulted in an *incorrect ATE with a considerable variance.* This experiment reveals that ignoring the relational structure in relational domains can lead to incorrect estimates and erroneous conclusions.

6.4 Sensitivity to Embeddings

Assessing the effect of embeddings requires access to the ground truth, so we restrict ourselves to testing on Synthetic ReviewData in this subsection. Table **??** shows that while CaRL consistently recovers the ATE, the correct choice of embedding can improve its performance. We observe that simple embeddings (such as mean or median) recovered approximately the true average treatment effect. However, their estimate was less centered around the ground truth compared to embeddings like padding or moment summarization. While padding had the tightest variance, moment summarization also showed promising results. These trends apply regardless of whether we consider single- or double-blind venues, each of which has different generative models and ground truths. It is important to note that moment summarization is one of the simpler approaches for set embedding. Additionally, the padding technique tends to create vectors that grow in proportion to the size of the relational skeleton, which limits to its applicability. In future work, we aim to develop principled learning approach for finding efficient embeddings using graph representation learning and graph embedding.

7 CONCLUSIONS AND FUTURE WORK

We introduced the Causal Relational Learning framework for performing causal inference on relational data. This framework allows users to encode background knowledge using a declarative language called *CaRL (Causal Relational Language)* using simple Datalog-like rules, and ask various complex causal queries on relational data. CaRL is designed for researchers and analysts with a social science, healthcare, academic or legal background who are interested inferring causality from a complex relational data. CaRL adds on to existing causal inference literature by relaxing the *unit-homogeniety assumption* and allowing the confounders, treatment units and outcome units to be of different kinds. We evaluated CaRL's completeness and correctness on real-world and synthetic data from academic and healthcare domains. CaRL is successfully able to recover the treatment effects for complex causal queries that may require multiple joins and aggregates.

Acknowledgments. We thank the anonymous reviewers for their feedback. This work is supported in part by NSF awards IIS-1552538, AITF-1535565, IIS-1614738, IIS-1703281, IIS-1703331, IIS-1703431, CCF-1740850, IIS-1907997, and NIH award R01EB025021.

REFERENCES

[1] Agency for Healthcare Research and Quality. NIS data elements: Bedsize Categories.

[2] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[3] David T. Arbour, Dan Garant, and David D. Jensen. Inferring network effects from observational data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 715–724, 2016.

[4] Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.*, 11(4):1912–1947, 12 2017.

[5] Andrey Balmin, Thanos Papadimitriou, and Yannis Papakonstantinou. Hypothetical queries in an olap environment. In *VLDB*, volume 220, page 231, 2000.

[6] Abhijit V Banerjee, Abhijit Banerjee, and Esther Duflo. *Poor economics: A radical rethinking of the way to fight global poverty*. Public Affairs, 2011.

[7] Daniel Deutch, Zachary G Ives, Tova Milo, and Val Tannen. Caravan: Provisioning for what-if analysis. In *CIDR*, 2013.

[8] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[9] Monica Giancotti, Annamaria Guglielmo, and Marianna Mauro. Efficiency and optimal size of hospitals: Results of a systematic search. *PLOS ONE*, 12(3):e0174533, March 2017.

[10] Bryan S Graham, Guido W Imbens, and Geert Ridder. Measuring the effects of segregation in the presence of social spillovers: A nonparametric approach. Technical report, National Bureau of Economic Research, 2010.

[11] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[12] M Elizabeth Halloran and Michael G Hudgens. Causal inference for vaccine effects on infectiousness. *The International Journal of Biostatistics*, 8(2):1–40, 2012.

[13] M Elizabeth Halloran and Claudio J Struchiner. Causal inference in infectious diseases. *Epidemiology*, 6(2):142–151, 1995.

[14] Healthcare Cost and Utilization Project (HCUP). HCUP Nationwide Inpatient Sample (NIS), 2006.

[15] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236, 2007.

[16] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):pp. 945–960, 1986.

[17] Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101(475):901–910, 2006.

[18] Stefano M Iacus, Gary King, Giuseppe Porro, et al. Cem: software for coarsened exact matching. *Journal of Statistical Software*, 30(9):1–27, 2009.

[19] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[20] Laks VS Lakshmanan, Alex Russakovsky, and Vaishnavi Sashikanth. What-if olap queries with changing dimensions. In *International Conference on Data Engineering*, pages 1334–1336. IEEE, 2008.

[21] Sanghack Lee and Vasant Honavar. On learning causal models from relational data. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[22] Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. A sound and complete algorithm for learning causal models from relational data. *arXiv preprint arXiv:1309.6843*, 2013.

[23] Marc Maier, Katerina Marazopoulou, and David Jensen. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*, 2013.

[24] Marc Maier, Brian Taylor, Huseyin Oktay, and David Jensen. Learning causal models of relational domains. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[25] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F. Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *Proc. VLDB Endow. (PVLDB)*, 4(1):34–45, 2010.

[26] Alexandra Meliou, Wolfgang Gatterbauer, Suman Nath, and Dan Suciu. Tracing data errors with view-conditioned causality. In *ACM SIGMOD International Conference on Management of data*, pages 505–516, 2011.

[27] Alexandra Meliou, Sudeepa Roy, and Dan Suciu. Causality and explanations in databases. *Proceedings of the VLDB Endowment*, 7(13):1715–1716, 2014.

[28] Alexandra Meliou and Dan Suciu. Tiresias: the database oracle for how-to queries. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 337–348. ACM, 2012.

[29] Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks, and chain graphs. *arXiv preprint arXiv:1812.04990*, 2018.

[30] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J van der Laan. Causal inference for social network data. *arXiv preprint arXiv:1705.08527*, 2017.

[31] Elizabeth L Ogburn, Tyler J VanderWeele, et al. Causal diagrams for interference. *Statistical science*, 29(4):559–578, 2014.

[32] Kanu Okike, Kevin T Hug, Mininder S Kocher, and Seth S Leopold. Single-blind vs double-blind peer review in the setting of author prestige. *JAMA*, 316(12):1315–1316, 2016.

[33] OpenReview. https://openreview.net.

[34] Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. Malts: Matching after learning to stretch. *arXiv preprint arXiv:1811.07415*, 2018.

[35] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

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

[37] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[38] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[39] Shanghai University Ranking. http://www.shanghairanking.com.

[40] Michael E. Rose and John R. Kitchin. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX*, 10:100263, July 2019.

[41] Joseph S Ross, Cary P Gross, Mayur M Desai, Yuling Hong, Augustus O Grant, Stephen R Daniels, Vladimir C Hachinski, Raymond J Gibbons, Timothy J Gardner, and Harlan M Krumholz. Effect of blinded peer review on abstract acceptance. *JAMA*, 295(14):1675–1680, 2006.

[42] Sudeepa Roy and Dan Suciu. A formal approach to finding explanations for database queries. ACM SIGMOD International Conference on Management of Data, 2014.

[43] Donald B Rubin. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1970.

[44] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[45] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[46] Donald B Rubin et al. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

[47] Babak Salimi, Leopoldo Bertossi, Dan Suciu, and Guy Van den Broeck. Quantifying causal effects on query answering in databases. In *TaPP*, 2016.

[48] Babak Salimi and Leopoldo E. Bertossi. From causes for database queries to repairs and model-based diagnosis and back. In *International Conference on Database Theory*, pages 342–362, 2015.

[49] Babak Salimi, Harsh Parikh, Moe Kayali, Sudeepa Roy, Lise Getoor, and Dan Suciu. Causal relational learning. *arXiv e-prints, arXiv:2004.03644, https://arxiv.org/abs/2004.03644*, 2020.

[50] Scopus. https://www.scopus.com/.

[51] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239, 2011.

[52] Richard Snodgrass. Single- versus double-blind reviewing: an analysis of the literature. *ACM Sigmod Record*, 35(3):8–21, 2006.

[53] Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

[54] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.

[55] Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[56] Tyler J VanderWeele and Eric J Tchetgen Tchetgen. Bounding the infectiousness effect in vaccine trials. *Epidemiology*, 22(5):686, 2011.