Causal Relational Learning

ABSTRACT

Causal inference is at the heart of empirical research in majority of sciences and social sciences, and is critical for making sound data-driven decisions. The gold standard in causal inference is performing *controlled experiments*, which is not always feasible due to ethical, legal, or cost constraints. As an alternative, inferring causality from *observational data* has been extensively used in statistical studies in public policy or social sciences. However, the existing methods critically rely on a restrictive assumption that the population of study consists of *homogeneous units* that can be represented as a single flat table. In contrast, in many real-world settings, the study domain consists of heterogeneous units with complex relational structure, where the data is naturally represented as multiple related tables, and therefore the data consists now of an entire relational database as opposed to a single table. In this paper, we present a formal framework for causal inference from such relational data, propose a declarative language called CaRL for capturing users' assumptions and specifying causal queries using simple Datalog-like rules, and develop an underlying inference engine that answers a suite of causal queries considering *relational* and *isolated* effects of the applied treatment. We give extensive experimental evaluations on synthetic and real data, and illustrate the applicability of our method on estimating the causal effect of institutional prestige on the acceptance of papers under single-blind and double-blind review processes.

1 INTRODUCTION

The importance of causal inference for taking informed policy decisions has been long 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 will happen to revenue if the price of X is lowered?". While predictive analytics has achieved remarkable success in diverse applications, it is based on fitting a model to observational data based on associational patterns [7]. Causal inference, on the other hand, goes beyond associational patterns to the process that generates the data, thereby enabling analysts to reason about *interventions*, *policies*, and *counterfactuals*.

The gold standard in causal analysis is performing *randomized controlled experiments*, where the *subjects* or *units* of the study are assigned randomly to a treatment or a placebo (withheld from the treatment). The difference of the outcome variable between the treatment group and the control group is called *average treatment effect*, and represents the causal effect. However, controlled experiments are not always feasible due to ethical, legal, or cost constraints. An attractive alternative that has been used in statistics, economics, and social studies is to simulate controlled experiments by *Observational Studies* using any data available, called *observational data*. While we can no longer assume that the treatment has been randomly assigned, under appropriate assumptions we can still learn causal relationships. Two well established frameworks exists for this purpose, Rubin's Potential Outcome Framework [9] and Pearl's Causal Models [5]. Both have been extensively studied in the literature and used in various applications, and there is an increased interest in observational studies, as can be seen in Fig 1, which shows the result of our quick analysis on SemanticScholar, revealing an increased interest in observational studies compared to controlled experiments. Babak: [The last sentence is too long] Dan: [agreed; it should capture concisely, adn with the balance the statements "this is the real trend in causality today" and "it's a a result that *we* obtained by doing a quick analysis using SemanticScholar"]

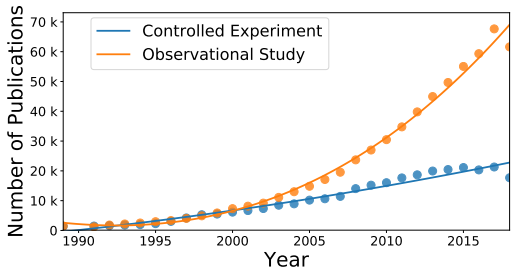

Figure 1: Number of publications on observational studies v.s. controlled experiments (obtained from SemanticScholar [10]).

However, all causal frameworks critically rely on the assumption that the units of study are sampled from a population consisting of homogeneous units that can be represented as a single flat table. In many real-world settings, the study domain consists of *heterogeneous units* that have a *complex relational structure*. In other words, the data usually consists of an entire relational database, and is not a single table. Standard notions used in causal analysis, such as units, casual DAG, covariates, no longer apply to relational data, prohibiting us from adopting existing causal inference frameworks to relational domains. We illustrate these challenges with an example.

Example 1.1. (ReviewData) *OpenReview [4] is a collection of paper submissions and their reviews to several conferences, mostly in ML and AI. What makes it interesting is that it contains the review scores for both accepted and rejected papers. Scopus [11] is a large, well maintained database of peer-reviewed literature, including scientific journals, books and conference proceedings. We crawled and integrated these two sources to produce a relational database, which we show in a simplified form in Fig. 2. Data sources like this represent a treasure trove of information for the leadership of scientific conferences and journals. For example, they can help answer questions like "does double blind achieve its desired effect"? "Does increasing (or decreasing) the page limit affect the quality of the papers and how?". To help in real-life decision, discovering associations is not sufficient, but instead decision makers would like to know if there exists a causal effect. For example, suppose a conference is currently requiring double blind submissions, and the leadership is questioning its effectiveness. If we revert to single blind submissions, would that represent an unfair advantage for authors* ***from top universities and companies (a.k.a. "prestigious institutions")*** *authors? Dan: [don't overcomplicate it. We have already given two examples above, now we need to narrow down to the question the we will eventually answer in the experiments. Whether it's about prestige of authors or prestige of institutions, that's exactly how we should formulate the question here.] Given a dataset like in Fig. 2, one can run a few SQL queries and check whether senior authors consistently get better reviews at single blind conferences than at double blind conferences, but this can only prove or disprove correlation, and not causation. Alternatively, one can try to apply the Neyman-Rubin causal model [9], but that requires us to present the data as a single table of independent units. If we do this naively on our dataset (e.g. computing the universal table [14]), then we create* interference *effects and* contagion *effects,* `Babak: [joining data sets do not create contagion and interference, these are properties of the data generative model (how the domain works)]` *Dan: [you are right. But the point to make here is that the relational data captures correctly the generative model, but if we just join them and re-define units based in the joint data, then there is intereference and contagion. Do you agree?] both of which prohibit standard causal analysis. For instance, prestige of an author not only influences their acceptance rate, but also has a* spill-over *effect on the acceptance rate of their co-authors; this is called interfere. Authors' qualifications can be* contagious *over a course of time, meaning that if a junior author collaborates frequently with a senior person then the overall quality of his/her paper may increase.*

In this paper, we propose a declarative framework for *Causal Relational Learning*, a foundation for inferring causal

Person
Bob
Carlos
Eva

Author	
Bob	s1
Eva	s1
Eva	s2
Eva	s3
Carlos	s3

Submission
s1
s2
s3

Submitted	
s1	DB
s2	AI
s3	AI

Conference
DB
AI

Prestige	
Bob	1
Carlos	0
Eva	1

Qualifications		
	Experience	h-index
Bob	10	50
Carlos	8	20
Eva	2	10

Score	
s1	0.75
s2	0.4
s3	0.1

Blind?	
DB	Single
AI	Double

Figure 2: Fragment of a relational database schema, obtained by integrating OpenReview and Scopus. (The data shown is not real.)

inference from relational domains. Specifically, we propose Relational Causal Models, which extend Pearl's Causal Models from representing causal relationship between attributes of homogeneous units, to heterogeneous units with complex relational structures. We extend Pearl's do-operator to relational causal models, to capture complex *relational interventions* and *relational causal queries*. We develop algorithms for detecting a sufficient set of *covariates* that should be adjusted for to remove confounding effects. We develop novel techniques such as *embedding relational data* for transforming relational data into a flat table amenable for easy causal inference.

Dan: [I think that this paragraph still needs work to explain our contributions] At the core of the relational causal model is a declarative language that specifies a complex causal structure in relational domains, and allows uses to formulate causal queries. As oppose to predictive analysis, domain expert knowledge plays a critical role and is fundamentally required for causal inference [6]. In our declarative language called CaRL users can specify their background knowledge and their causal queries about relational domains with simple and intuitive Datalog-like rules.

Dan: [We should have one contribution per section.] We make the following contributions: (1) We propose Relational Causal Molds, which lift Pearl's Causal Models to relational domains. (2) We propose CaRL a declarative language for capturing users' assumptions and queries using simple Datalog-like rules. (2) We develop an underlying inference engine that automatically detects a sufficient set of *covariates* that should be adjusted for to remove confounding effects. (3) We develop novel techniques such as *embedding* to transform relational data into a flat table amenable to easy causal inference. (4) we answers a suite of causal queries considering relational and isolated effects. (5) We evaluate CaRL on synthetic and real data, and illustrate the applicability of our method on estimating the causal effect of institutional

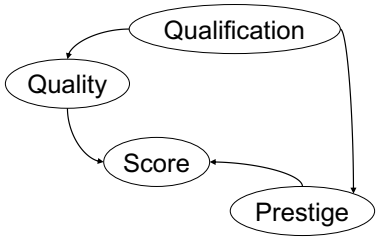

Figure 3: A simple causal DAG.

prestige on the acceptance of papers under single-blind and double-blind review processes.

2 BACKGROUND ON CAUSAL ANALYSIS

We briefly review here some basic concepts in causal analysis, covering both Pearl's casual DAGs and the Neyman-Rubin model, following [5, 9].

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

Causal DAG. The probabilistic causal model is associated to a *causal graph*, G, whose nodes are the endogenous variables $\mathbf{V}$, and whose edges are all pairs (Z, X) such that $Z \in \mathbf{Pa_V}(X)$; we write $Z \rightarrow X$ for an edge; it is usually assumed that G is acyclic and then it is called a Causal DAG. In other words, the causal DAG hides the exogenous variables (since we can't observe them anyway) and instead captures their effect by defining a probability distribution $\Pr_{\mathbf{V}}$ on the endogenous variables. This is possible under the *causal sufficiency* assumption[1]. The formula for $\Pr_{\mathbf{V}}$ is the same as that for a Bayesian network:

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

We will only refer to endogenous variables in the rest of the paper and drop the subscript $\mathbf{V}$ from $\mathbf{Pa_V}$ and $\Pr_{\mathbf{V}}$. Fig. 3 shows a simple example of a causal graph: the Score (of a paper) is affected by its Quality and by the Prestige of the author (assuming the reviews are single blind); the latter are both affected by the authors' Qualification. In this paper we will assume that the causal DAG is known[2], and one uses some observational data in order to learn the conditional probability distribution (1). Notice a fundamental limitation of traditional causal frameworks. The units need to be uniform and independent. In other words the data consists of a uniform set of tuples, called *units*, with four attributes (Qualification, Quality, Prestige, Score); each row represents one paper, its (unique?) author, and the outcome. Situations like multiple authors for the same paper, or multiple submissions by the same authors cannot be represented in this model and, in fact, would violate fundamental assumptions made by the model.

***d*-Separation and Markov compatibility.** A common inference question in a causal DAG is how to determine whether a CI $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ holds. A sufficient criterion is given by the notion of d-separation, a syntactic condition $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ that can be checked directly on the graph. $\Pr$ and G are called *Markov compatible* if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ implies $(\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y}|\mathbf{Z})$; if the converse implication holds, then we say that $\Pr$ is *faithful* to G. The following is known:

Proposition 2.1. *If G is a causal DAG and* $\Pr$ *is given by Eq.(1), then they are Markov compatible.*

Interventions and do operator. An *intervention* represents, intuitively, actively setting an endogenous variable to some fixed value, and observing the effect. Pearl [5] introduced for this purpose the *do*-calculus. Formally, an intervention consists of setting a variables $\mathbf{U}$ to some values $\mathbf{U} = \mathbf{u}$, and it defines the probability distribution $\Pr(\mathbf{V}|\mathsf{do}(\mathbf{U} = \mathbf{u}))$ given by Eq. 1 where we remove all factors $\Pr(X|\mathbf{Pa}(X))$ where $X \in \mathbf{U}$. In other words we modify the causal DAG by removing all edges entering the variables on which we intervene; of course, this is different from conditioning, $\Pr(\mathbf{V}|\mathbf{U} = \mathbf{u})$. Pearl has an extensive discussion on the rationale of the *do*-calculus and describes several equivalent formulas for estimating $\Pr(\mathbf{V}|\mathsf{do}(\mathbf{U} = \mathbf{u}))$ from observed distribution.

Average Treatment Effect (ATE). The scope of Neyman-Rubin's causal model [9] is to compare the effect of a binary treatment variable T on some response variable Y, which is often measured by the following quantify known as *Average Treatment Effect (ATE)* and expressed as follows in our notation:

$$ATE(Y, T) = \mathbb{E}[Y|do(T = 1)] - \mathbb{E}[Y|do(T = 0)] \tag{2}$$

Much of the literature on the Neyman-Rubin causal model in statistics is concerned with efficient estimation of ATE from observational data.

[1] The assumption requires that, 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.

[2] There exists an extensive literature on learning the causal DAG from data.

3 RELATIONAL CAUSAL MODEL

As discussed in previous sections, the traditional frameworks for causal inference are unsuitable for capturing complex causal relationships in relational domains comprising heterogenous units, which may be different entities or relationships between those entities. In this section we propose a declarative language called *CaRL (Causal Relational Language)* that extends Pearl's Causal Model and causal graphs to relational data by allowing the user to (1) specify assumptions and backgrounded knowledge on the interactions among heterogenous units (Section 3.2), and (2) ask various causal queries (Section 3.3). We start with our data model that forms the basis for our language in Section 3.1.

3.1 Data Model

Schema. A *relational causal schema* is a tuple $\mathbf{S} = (\mathbf{P}, \mathbf{A})$, where $\mathbf{P} = \{P_1, \ldots, P_n\}$ is a standard relational schema that represents a set of *entities* $\mathbf{E}$ and their *relationships* $\mathbf{R}$, hence $\mathbf{P} = \mathbf{E} \cup \mathbf{R}$; and $\mathbf{A} = \{\mathtt{A}_1, \ldots, \mathtt{A}_k\}$ is a set of *attribute functions* with fixed arity, domain and range *Range*($\mathtt{A}$) that encode the descriptive values of the entities and their relationships.

Skeleton and observed instance Our language represents the causal relationships between the attribute functions given a set of entities and their relationships. Therefore, we refer to a standard relational database instance from schema P as a *relational skeleton*. A relational skeleton $\mathbf{\Delta}$ together with the values of a subset of observed attribute functions $\mathbf{A}$ is called an *observed relational instance*. Moreover, we distinguish between the attribute functions and the entities and their relationships by denoting the former with $P(.)$ and the latter with $\mathtt{A}[.]$.

Example 3.1. The relational causal schema corresponding to the relational data in Figure 2 is as follows:

$\mathbf{P}$ = *People*(A), *Submission*(S), *Conference*(C), *Submitted*(S, C), *Author*(A, S)

$\mathbf{A}$ = *Prestige*[A], *Qualification*[A], *Score*[S], *Quality*[S], *Blind*[C]

Here, People(A)*, Submission*(S) *and Conference*(C) *are entities; whereas Authors*(A, S) *and Submitted*(S, C) *are relationships. Moreover, Prestige*[A] *describes the prestige of the institution that the author is affiliated with measured by its ranking. Qualification*[A] *is a compact representation of an author's h-index and experience. Score*[S] *describes the average score reviewers gave to a submission:* 1 *being perfect and* 0 *being the worst possible. Quality*[S] *describes the quality of a submission (which is unobserved). Finally, Blind*(C) *shows whether a conference review policy is single-blind or double-blind. We refer to the tables in the first row of Figure 2 as the relational skeleton of the observed instance.*

3.2 Specification Language and Semantics

Now we discuss the syntax of CaRL to encode backgrounded causal knowledge in relational domains. At the core of CaRL, we have *relational causal rules* to capture the causal assumptions.

Definition 3.2. A *relational causal rule* (or simply a *rule*) over a relational schema $\mathbf{S} = (\mathbf{P}, \mathbf{A})$ has the following form:

$$\mathtt{A}[\mathbf{X}] \Leftarrow \mathtt{A}_1[\mathbf{X}_1], \ldots, \mathtt{A}_k[\mathbf{X}_k] \text{ WHERE } Q(\mathbf{Y}) \tag{3}$$

Here $\mathtt{A}, \mathtt{A}_1, \cdots, \mathtt{A}_k \in \mathbf{A}$, Q is a Boolean conjunctive query (BCQ) on the predicates in $\mathbf{P}$, $\cup_{i=1}^{k} \mathbf{X_i} \subseteq \mathbf{Y}$, and $\mathbf{X} \subseteq \mathbf{Y}$. 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*. Denote $\phi_{\mathtt{A}}$ an rule with an attribute function $\mathtt{A}$ in the head.

Definition 3.3. A *relational causal model* Φ for a relational schema $\mathbf{S} = (\mathbf{P}, \mathbf{A})$ is a collection of rules associated to $\mathbf{A}$, *i.e.*, $\Phi = \cup_{\mathtt{A} \in \mathbf{A}} \phi_{\mathtt{A}}$.

A rule is template for generating multiple *grounded rules*.

Definition 3.4. A rule $\phi_{\mathtt{A}}$ and a relational skeleton $\mathbf{\Delta}$ define a set of *grounded rules* $\phi_{\mathtt{A}}^{\mathbf{\Delta}}$ obtained by substituting the variables $\mathbf{Z} = \mathbf{X} \cup \mathbf{X}_1 \cup \ldots \cup \mathbf{X}_k$ in $\phi_{\mathtt{A}}$ with a set of constants $\mathbf{z}$ (denoted by $\mathbf{Y}/\mathbf{z}$), such that $\mathbf{\Delta} \models Q([\mathbf{Y}/\mathbf{z}])$, *i.e.*, when we substitute variables $\mathbf{Y}$ with constants $\mathbf{z}$ the Boolean query Q evaluates to true.

A relational causal model Φ together with a relational skeleton $\mathbf{\Delta}$ defines a collection of grounded rules $\Phi^{\mathbf{\Delta}} \stackrel{\text{def}}{=} \cup_{\phi_{\mathtt{A}} \in \Phi} \phi_{\mathtt{A}}^{\mathbf{\Delta}}$. To each $\Phi^{\mathbf{\Delta}}$, we associate a *grounded causal graph* $G(\Phi^{\mathbf{\Delta}})$ with vertices consisting of the set of all grounded atoms in $\Phi^{\mathbf{\Delta}}$, denoted $\mathbf{A}^{\mathbf{\Delta}}$, and directed edges $(\mathtt{A}_i(\mathbf{x}_i), \mathtt{A}_j(\mathbf{x}_j))$ if $\mathtt{A}_i(\mathbf{x}_i)$ appears in the head and $\mathtt{A}_j(\mathbf{x}_j)$ appears in the body of a grounded rule in $\Phi^{\mathbf{\Delta}}$. We denote $\mathtt{A}^{\mathbf{\Delta}} \subseteq \mathbf{A}^{\mathbf{\Delta}}$ the set of all groundings of an attribute function $\mathtt{A} \in \mathbf{A}$ in $\mathbf{A}^{\mathbf{\Delta}}$. We interpret a grounded causal graph $G(\Phi^{\mathbf{\Delta}})$ as a standard probabilistic causal models characterized by a joint probability distribution $\mathtt{Pr}(\mathbf{A}^{\mathbf{\Delta}})$ and the following conditional probability distributions, for each $\mathtt{A}[\mathbf{x}] \in \mathbf{A}^{\mathbf{\Delta}}$.

$$\mathtt{Pr}\big(\mathtt{A}[\mathbf{x}] \mid \mathbf{Pa}(\mathtt{A}[\mathbf{x}])\big). \tag{4}$$

Example 3.5. Consider the following rules that postulate a relational causal model for ReviewData in Figure 2.

$$\text{Prestige}[A] \Leftarrow \text{Qualification}[A] \text{ WHERE Author}(A) \tag{5}$$

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

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

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

The rule (5) indicates that the qualification of an author causally affects their prestige due to the fact that prestigious institutions tend to hire most qualified candidates. The rule (6) stipulates that the quality of a submission is affected by the qualifications of its authors as well-as their prestige (presumably, authors from prestigious institutions have acess

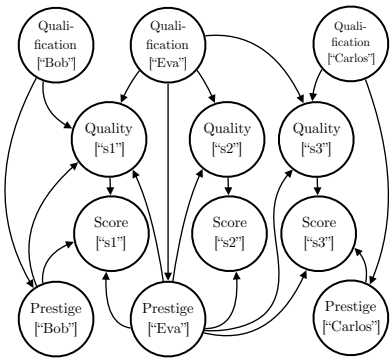

Figure 4: grounded causal graph corresponded to the grounded rules in Example 3.5.

to more resources). Finally, the rules (7) and (8) collectively stipulate that reviewers scores are based on the quality of a submission, but also potentially influenced by prestige of its authors. We implicitly assign a rule with empty body to the rest of functional attributes, meaning that they not causally influenced by the other attribute functions in the schema.

The relational causal model together with the toy relational skeleton instance in Figure 2 generates the following grounded rules:

$$\text{Prestige}[\text{“}Bob\text{”}] \Leftarrow \text{Qualification}[\text{“}Bob\text{”}] \tag{9}$$
$$\text{Prestige}[\text{“}Carlos\text{”}] \Leftarrow \text{Qualification}[\text{“}Carlos\text{”}] \tag{10}$$
$$\text{Prestige}[\text{“}Eva\text{”}] \Leftarrow \text{Qualification}[\text{“}Eva\text{”}] \tag{11}$$
$$\text{Quality}[\text{“}s1\text{”}] \Leftarrow \text{Qualification}[\text{“}Bob\text{”}], \text{Qualification}[\text{“}Eva\text{”}] \tag{12}$$
$$\text{Quality}[\text{“}s2\text{”}] \Leftarrow \text{Qualification}[\text{“}Eva\text{”}] \tag{13}$$
$$\text{Quality}[\text{“}s3\text{”}] \Leftarrow \text{Qualification}[\text{“}Carlos\text{”}], \text{Qualification}[\text{“}Eva\text{”}] \tag{14}$$
$$\text{Score}[\text{“}s1\text{”}] \Leftarrow \text{Quality}[\text{“}s1\text{”}], \text{Prestige}[\text{“}Bob\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \tag{15}$$
$$\text{Score}[\text{“}s2\text{”}] \Leftarrow \text{Quality}[\text{“}s2\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \tag{16}$$
$$\text{Score}[\text{“}s3\text{”}] \Leftarrow \text{Quality}[\text{“}s3\text{”}], \text{Prestige}[\text{“}Carlos\text{”}], \text{Prestige}[\text{“}Eva\text{”}] \tag{17}$$

The rules in Example 3.5 result in the grounded causal graph shown in Figure 4. Note that our language allows for recursive rules:

$$\text{Qualification}[A] \Leftarrow \text{Qualification}[A'] \text{ WHERE Collaborator}(A, A') \tag{18}$$

Where, Collaborator(A, A') is a derived predicates that shows whether two authors were collaborators in the past. The rule (18) states that an author's qualifications has a feedback nature, i.e., the qualification of an author can influence the qualifications of his/her collaborators over time. Even though CaRL is equipped to handle cyclic causal dependency such as 18, using *equilibrium of discrete Markov chains* for brevity we discuss recursive rules only in the full technical report. [SR: do we want to mention cyclic rules? we have not discussed those details yet. How about saying out of scope of the current paper?]

3.3 Query Language

This section describes different types of causal queries supported in CaRL. The semantics of the queries will be explained in Section 4.

Average treatment effect (ATE). In CaRL, queries about ATE (see (2)) of a binary attribute function $T[\mathbf{X}]$ on a response attribute function $Y[\mathbf{X}']$ expressed as follows:

$$Y[\mathbf{X}'] \Leftarrow T[\mathbf{X}]? \tag{19}$$

For example, the query

$$\textit{Score}[A] \Leftarrow \textit{Prestige}[A]? \tag{20}$$

compute the ATE of *Prestige* on *Score*, *i.e.*, it compares papers' scores in two hypothetical worlds in which all authors are, and are not affiliated with prestigious institutions.

Aggregated Response. In CaRL one can define aggregated responses such as AVG_$score[A]$, the average submission scores of an authors, and answer queries about ATE of a treatment on an aggregated outcome. For example the following query quantifies the ATE of *Prestige* on AVG_$score[A]$:

$$\text{AVG\_}score[A] \Leftarrow \textit{Prestige}[A]? \tag{21}$$

Relational and Isolated Effects. In relational domain units that are relationally connected can have a causal influence on each other. For example, Prestige of an authors not only influences their average submission scores but also their collaborators average submission scores. To measure such complex relational causal interactions CaRL support queries about *relational* and *isolated* effects, expressed as follows (will be defined rigorously in Section 4):

$$Y[\mathbf{X}'] \Leftarrow T[\mathbf{X}] \text{ ? WHEN } \langle cnd \rangle \text{ PEERS TREATED} \tag{22}$$

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

$$\langle cnd \rangle \leftarrow \langle \text{LESS} \mid \text{MORE} \rangle \text{ THAN } k\% \mid \text{AT } \langle \text{MOST} \mid \text{LEAST} \rangle \text{ k} \mid \text{EXACTLY k} \mid \text{ALL} \mid \text{NONE} \tag{23}$$

For example, the query

$$\textit{Score}[S] \Leftarrow \textit{Prestige}[A]? \text{ WHEN ALL PEERS TREATED} \tag{24}$$

compute both isolated and relation effect prestige on reviewer scores if all coauthors of an author are from prestigious institutions. The isolated effect measures to what extent an author review score is influenced by their own prestige. Whereas, the relational effect quantifies to what extent the author review scores are influenced by the prestige of their collaborators.

3.4 Structural Homogeneity

Recall from Section 2 that in standard probabilistic causal models the conditional probability distribution $\Pr(X|\mathbf{Pa}(X))$ for each $X \in \mathbf{V}$, captures the underlying structural equation F_x associated to X as a probabilistic mapping. Since, different

variables have different corresponding structural equations, it is natural to associate different conditional probability distribution to each variable. For instance in Figure 3 the structural equations corresponded to *Score* and *Prestige* models two different functions with different inputs and outputs. Also note that these conditional probability distributions are unknown and must be estimated from available data.

However, in CaRL assuming different structural equations, and thereby different conditional probability distribution $\mathsf{Pr}\big(\mathsf{A}[\mathbf{x}] \mid \mathbf{Pa}(\mathsf{A}[\mathbf{x}])\big)$ for each $\mathsf{A}[\mathbf{x}] \in \mathbf{A}^{\Delta}$, is pathological. Here, the number of grounded atom are not fixed a priory and depends on the size of the skeleton (the relational database); we no longer have the attribute *Score*, but instead we have *Score*["*s*1"], *Score*["*s*2"], ... Under the unreasonable assumption that the structural equations associated to grounded atom are different, the estimation of the corresponding conditional probability distributions $\mathsf{Pr}\big(\mathsf{A}[\mathbf{x}] \mid \mathbf{Pa}(\mathsf{A}[\mathbf{x}])\big)$, and hence causal inference becomes impossible. Therefore in this paper we make the following reasonable *Structural Homogeneity assumption*:

- **Structural Homogeneity:** All the grounded atoms $\mathsf{A}[\mathbf{x}] \in \mathsf{A}^{\Delta}$ of an attribute function $\mathsf{A} \in \mathbf{A}$ share the same structural equation, which models a function with variable input size that determines the value of $\mathsf{A}[\mathbf{x}]$, given its endogenous and exogenous parents.

For instance, in Example 3.5, we assume the underlying structural equation for submission's scoring is the same for all submissions.

The structural homogeneity assumption, however, is not easily captured by associating the same conditional probability distribution corresponding to all grounded atoms $\mathsf{A}[\mathbf{x}] \in \mathsf{A}^{\Delta}$. Different grounded atoms can have different number of parents in the grounded causal digram. For instance, consider the atoms *Score["s1"]* and *Score["s2"]* from (15) and (17). In this toy example, *Score["s1"]* depends on the *Prestige* of two authors *"Eva"* and *"Bob"*, whereas *Tutor["s2"]* depends on the *Prestige* of one author only: *"Eva."* This requires a *unification* of the parents of the same type using another layer of functions ψ (using aggregates, or in general, an embedding, see Section **??**).

Formally, instead of using (4), we assume the structural equations associated to the groundings of an attribute function A can be captured using the following conditional probability distribution shared by all $\mathsf{A}[\mathbf{x}] \in \mathsf{A}^{\Delta}$:

$$\mathsf{Pr}\Big(\mathsf{A}[\mathbf{x}] \mid \Psi^{\mathsf{A}}\big(\mathbf{Pa}(\mathsf{A}[\mathbf{x}])\big)\Big) \tag{25}$$

where, Ψ^{A} is a collection of mappings that projects the parents of $\mathsf{A}[\mathbf{x}]$ into a low-dimensional vector with fixed dimensionality for all $\mathsf{A}[\mathbf{x}] \in \mathsf{A}^{\Delta}$. Intuitively, we assume that the mappings provide *sufficient statistics* for evaluation of the underling structural questions corresponded to all $\mathsf{A}[\mathbf{x}] \in \mathsf{A}^{\Delta}$. Throughout this paper we assume $\Psi^{\mathsf{A}}_{\mathbf{V}}$ is known and consists of set of mappings $\{\psi^{\mathsf{A}}_{\mathsf{A}_1}, \psi^{\mathsf{A}}_{\mathsf{A}_2}, \ldots\}$ where, each $\psi^{\mathsf{A}}_{\mathsf{A}_1}$ is an *embedding function* that map a subset of $\mathbf{Pa}(\mathsf{A}[\mathbf{x}])$ consisting of the grounding of the attribute function A_i into a low-dimensional *embedding space* with fixed dimensionality. In section **??** we discuss a set of techniques to learn Ψ^{A} from data.

Example 3.6. Let us assume whether a submission is accepted or not is a function of the result of applying a mapping $\psi^{Score}_{Prestige}$ to the vector of the *Prestige* of the submission authors, of the *Prestige* of the authors of the submission into a vector with a fixed dimension for all submissions, e.g., a vectors consisting of the average (weighted by the authors position in the submission) *Prestige*s and the number of authors. In our running example the vector $\langle 1, 1\rangle$ and $\langle 0\rangle$, respectively corresponded to the prestige of authors of the submissions s_1 and for s_2. Using average and size of the vector, the embedded vectors become $\langle 1, 2\rangle$ and $\langle 0, 1\rangle$, respectively.

Note that in this paper we work on the common assumption of causal sufficiency and acyclicity of $G(\Phi^{\Delta})$. The following factorization, which forms a basis for defining and identifying the effect of interventions (see Section 4 and 5), immediately implied from (25) and Markov compatibility of $\mathsf{Pr}(\mathbf{A}^{\Delta})$ and the grounded causal graph $G(\Phi^{\Delta})$.

$$\mathsf{Pr}(\mathbf{A}^{\Delta}) = \prod_{\mathsf{A}[\mathbf{x}] \in \mathbf{A}^{\Delta}} \mathsf{Pr}\Big(\mathsf{A}[\mathbf{x}] \mid \Psi^{\mathsf{A}}\big(\mathbf{Pa}(\mathsf{A}[\mathbf{x}])\big)\Big) \tag{26}$$

Structural Heterogeneity. In practice, relational data may consist of data collected from different domain, *e.g.*, different organization, cooperation, countries, cities, etc that could lead to the violation of the Structural homogeneity assumption. For instance, its is reasonable to assume the the structural question associated to scoring is different across single-blind and double-blind conferences. Such situations can be expressed in CaRL by postulating different rule at *different granularity* in which the homogeneity assumption perceived to hold, same organization, all double-blind conferences, all conference in computer science, etc. For example, the heterogeneity of paper scoring can be captured using the following rules:

$$\textit{SBlind\_Score}[S] \Leftarrow \textit{Prestige}[A] \text{ WHERE } \textit{Contributor}(A, S)$$
$$\textit{SBlind\_Score}[S] \Leftarrow \textit{Quality}[S] \text{ WHERE } \textit{Submission}(S)$$
$$\textit{DBlind\_Score}[S] \Leftarrow \textit{Quality}[S] \text{ WHERE } \textit{Submission}(S)$$

that show the generative models for scoring are different in double-blind and single-blind conferences. Note that CaRL supports a compact way for representing heterogeneous domains (to be included in full technical report).

In general for causal inference from relational data it is indispensable to assume that the homogeneity holds at some granularity level.

4 SEMANTIC FOR CAUSAL QUERIES

This section defines the semantics of the causal queries described in Section 3.3. 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 $\mathbb{U}_{\mathtt{A}}$ 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 the authors, e.g., {"Bob", "Eva", "Carlos"}, whereas $\mathbb{U}_{Score}$ consists of all the submissions, e.g., {*"s1", "s2"*}. We refer to each element $\mathbf{x} \in \mathbb{U}_{\mathtt{A}}$ as a *unit* of an attribute function $\mathtt{A}$.

4.1 Treatment and response attribute function

In a causal analysis, we are given a pair of attribute functions $T[\mathbf{X}], Y[\mathbf{X}'] \in \mathbf{A}$, the goal is to estimate the effect of intervening on the attribute function T of the *treated units* $\mathbb{U}_T$, on the attribute function Y of the *response units* $\mathbb{U}_Y$, hence we call T the *treatment attribute function* and Y the *response attribute function*. For example, to study the effect of prestige on submission scores, *Prestige*$[A]$ is the treatment and *Score*$[S]$ is the response attribute function. Here, the treated units are $\mathbb{U}_{Prestige}$ (all authors) and the response units are $\mathbb{U}_{Score}$ (all submissions). For exposition, we assume $Range(T) = \{0, 1\}$ and $Range(Y) = \mathbb{R}$.

Formally, 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 $\mathsf{do}(T(\mathbf{x}_i) = t_i)$ 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 an intervention in which *hypothetically changes the authors' afflictions to a prestigious one.* 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 the elements in the set $\mathbb{S}$.

Here we emphasize the power of our framework for causal analysis on relational data with two key observations:

- **Non-uniform treated and response units.** *The treated and response units can correspond to different entities or relationships.* In our example, they correspond to entities for S and entities of (A, S), but if we want to understand the intervention of prestige on a conference's average submission score, it could be two entities A and C that are not directly related.
- **Independent treated units.** *Each treated unit can be subjected to an intervention* independently *of all other treated units.* In our example, we can have a treatment vector $\vec{1}$ where we force *all* to be afflicted with prestigious institutions, as well as another treatment vector $(1, 0, 0, \ldots)$ where we only force the "*Bob*" to be affiliated to a prestigious institution.

Our goal is to answer a set of queries that compare the *average response* of the units $\mathbb{U}_Y$ to two alternative intervention strategies $\vec{t}$ and $\vec{t}'$ applied to the treated units $\mathbb{U}_Y$ as we discuss below.

4.2 Average Treatment Effect (ATE) on attributes

The primary causal query in CaRL is that of average treatment effect (ATE) as in (19), which is defined as follows:

$$ATE(T, Y) \stackrel{\text{def}}{=} \sum_{\mathbf{x}' \in \mathbb{U}_Y} \frac{1}{m} (\mathbb{E}[Y[\mathbf{x}'] \mid \mathsf{do}(T[\mathbb{U}_T] = \vec{0})] - \mathbb{E}[Y[\mathbf{x}'] \mid \mathsf{do}(T[\mathbb{U}_T] = \vec{1})]) \quad (27)$$

Intuitively, ATE compares the expected response of the response units in two regimes of interventions in which all units receive treatment and all of them do not receive treatment respectively. For example, $ATE(\textsc{Prestige}, \textsc{Score})$ compare papers' scores under two interventions in which all authors are and are not affiliated with prestigious institutions.

4.3 Extending attribute functions and schema with aggregates

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

$$\begin{aligned} \text{AGG\_A}[\mathbf{W}] &\Leftarrow \mathtt{A}[\mathbf{X}] \text{ WHERE } Q(\mathbf{Z}) \\ \text{AGG\_A}[\mathbf{W}'] &\Leftarrow \mathit{AGG}\text{\_A}[\mathbf{W}] \text{ WHERE } Q(\mathbf{Z}) \end{aligned} \quad (28)$$

Here, $\mathbf{Z} \supseteq \mathbf{X}' \cup \mathbf{W}$ and AGG is an aggregate function on $\mathtt{A}$, e.g., AVG (average) and VAR (variance). The new aggregated attribute functions *AGG*_A are included in the extended attribute functions $\mathbf{A}$ (for simplicity, we use $\mathbf{A}$ for both given and extended attribute functions). The semantic of aggregated rules is defined similar to rules in Section 3, *i.e.*, they define a set of grounded rules with corresponding vertices and edges in the grounded causal graph $G(\Phi^{\Delta})$ as described in Section 3. However, instead of a conditional probability distribution, a deterministic function $\text{AGG}(\mathbf{Pa}(\text{AGG\_}Y[\mathbf{w}]))$, will be associated to each $\text{AGG\_}Y[\mathbf{w}] \in \text{AGG\_}Y^{\Delta}$. For example, the following aggregate rule defines the average acceptance rate for each author.

$$\text{AVG\_}\mathit{Score}[A] \Leftarrow \mathit{Score}[S] \text{ WHERE } \mathit{Contributor}(A, S) \quad (29)$$

Find the grounded causal graph that incorporates (29) in Figure 5.

4.4 Average relational and isolated effects on attributes and aggregates

In order to formalize isolated and relational effects as described in Section 3.3, we need to establish a one-to-one correspondence between treated and response units by using aggregations carefully. To this end, first we define relational paths.

Definition 4.1. Given a relational causal schema $\mathbf{S} = (\mathbf{P}, \mathbf{A})$ consisting of a set for entities and relations $\mathbf{P} = \mathbf{E} \cup \mathbf{R}$, and a set of attribute functions $\mathbf{A}$ (possibly extended with aggregates), a *relational path* is a sequence of entities and relationships of the following form:

$$\mathcal{P} : E_1(X_1) \xleftrightarrow{R_1(X_1,X_2)} E_1(X_2) \cdots E_{\ell-1}(X_{\ell-1}) \xleftrightarrow{R_{\ell-1}(X_{\ell-1},X_\ell)} E_\ell(X_\ell) \tag{30}$$

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, $\mathit{Conference}(C) \xleftrightarrow{\mathit{Submitted}(S,C)} \mathit{Submission}(S)$ is a relational path in our example. The treated and response units corresponding to a 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 that T and Y describe either as the endpoints in the path or as the labels of the edges at the ends of the path. For example, for $T[\mathbf{X}] = \mathit{Prestige}[A]$ and $Y[\mathbf{X}'] = \mathit{Score}[S]$, the treatment is an attribute function of the entity $\mathit{Author}(A)$, the response is an attribute function of the relationship $\mathit{Contributor}(A, S)$, and the treated and response units are relationally connected by the following relational path:

$$\mathit{Author}(A) \xleftrightarrow{\mathit{Contributor}(A,S)} \mathit{Submission}(S) \tag{31}$$

In this paper, we make the natural assumption that the treated and response units are relationally connected by at least one relational path as in (30). Then the treated and response units can be unified using the aggregated response AGG_$Y[\mathbf{X}]$ defined with the following aggregate rule that maps attribute Y in entity(s) X' to entity(s) X:

$$\text{AGG\_}Y[\mathbf{X}] \Leftarrow Y[\mathbf{X}'] \text{ WHERE } R_1(X_1, X_2), \ldots, R_{\ell-1}(X_{\ell-1}, X_\ell) \tag{32}$$

For example, for unifying the treated and response units associated to $T[\mathbf{X}] = \mathit{Prestige}[A]$ and $Y[\mathbf{X}'] = \mathit{Score}[S]$, the aggregate rule associated with the relational path in (31) coincides with (29).

Therefore, from now on, we assume 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 $\mathbb{U}_{(T,Y)} = \mathbb{U}_T = \mathbb{U}_Y$.

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 4.3.

Definition 4.2. 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 $\mathbf{x}' \in \mathbb{P}(\mathbf{x})$, there exists a directed path from $T[\mathbf{x}']$ to $Y[\mathbf{x}]$ in $G(\Phi^\Delta)$.

For example, for $T[\mathbf{X}] = \mathit{Prestige}[A]$ and $Y[\mathbf{X}'] = \text{AVG\_}\mathit{Score}[A]$, $\mathbb{P}(\text{“}\mathit{Bob}\text{”}) = \{\text{“}\mathit{Eve}\text{”}\}$ and $\mathbb{P}(\text{“}\mathit{Eve}\text{”}) = \{\text{“}\mathit{Bob}\text{”}, \text{“}\mathit{Carlos}\text{”}\}$. In practice, the relational causal model is expected to form relational peers $\mathbb{P}(\mathbf{x})$ such that they consist only of units that are in some *relational proximity* of $\mathbf{x}$, e.g., authors from the same institution, same research interests, etc.

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 $\vec{t} = (t_1, t_2 \ldots)$.

$$Y_{\mathbf{x}}(t, \vec{t}) \stackrel{\text{def}}{=} \mathbb{E}[Y[\mathbf{x}] \mid \mathsf{do}\big(T[\mathbf{x}] = t\big), \mathsf{do}\big(T[\mathbb{P}(\mathbf{x})] = \vec{t}\big)] \tag{33}$$

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 either has a natural ordering or is ordering-invariant. However, as we discuss after (23), we allow several relaxations on the size and type on $\vec{t}$ in our framework.

Overall, Isolated and Relational Effects. The query (22) in CaRL, compute the following quantities that compare the overall, isolated and relational effects of two alternative intervention strategies $(t, \vec{t})$ and $(t', \vec{t}')$:

- The average *isolated causal effect*:

$$\text{AIE}(t; t' \mid \vec{t}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t}) \tag{34}$$

- The average *relational causal effect*:

$$\text{ARE}(\vec{t}; \vec{t}' \mid t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t, \vec{t}') \tag{35}$$

- The average *overall causal effect*:

$$\text{AOE}(t, \vec{t}; t', \vec{t}') \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t, \vec{t}) - Y_{\mathbf{x}}(t', \vec{t}') \tag{36}$$

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 $\mathit{Prestige}[A]$ on AVG_$\mathit{Score}[A]$ fixes the prestige of an author such as “*Bob*” and compares the counterfactual response AVG_Score[“*Bob*”] under two regimes of interventions in

which the relational peers of "*Bob*", e.g., "*Eve*", receive two different treatment strategies, e.g., all of them have prestigious affiliation versus non-of them have prestigious affiliations. Notice that the overall causal effect is an extension of ATE (27) for two arbitrary treatment strategies. Indeed, ATE coincides with AOE$(1, \vec{1} \mid 0, \vec{0})$, when the treated and response units are unified. The following preposition shows the connection between relational, isolated and overall effects.

Proposition 4.3. The average overall effect can be decomposed into the average isolated and average relational effects as follows:

$$\begin{aligned} \text{AOE}(t, \vec{t}; t', \vec{t}') &= \text{AIE}(t, t' \mid \vec{t}) + \text{ARE}(\vec{t}, \vec{t}' \mid t') \\ &= \text{AIE}(t, t' \mid \vec{t}') + \text{ARE}(\vec{t}, \vec{t}' \mid t) \end{aligned} \tag{37}$$

Relaxations on the treatment vector to the peers. In the definitions of the average isolated, relational, and overall causal effects in (34), (35), and (36), 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 $\mathbf{x}$, as well as we do not need all units $\mathbf{x}$ to have the same number of peers in $\mathbb{P}(\mathbf{x})$. As the grammar defined in (23) 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 $\mathbb{P}(\mathbf{x})$.

5 ANSWERING CAUSAL QUERIES

The query answering component of CaRL consists of the *covariate detection* (Section 5.1) and *covariate adjustment* (Section 5.2). The goal of covariate detection is to find 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 such that adjustment can be performed using standard methods.

5.1 Covariate Detection

Given a treatment and a response attribute functions $T[\mathbf{X}]$ and $Y[\mathbf{X}']$, the quantity $\texttt{Pr}\Big(Y[\mathbf{x}'] = y \mid \texttt{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big)$ for a response unit $\mathbf{x}' \in \mathbb{U}_Y$ and a subset of treated units $\mathbb{S} \subseteq \mathbb{U}_T$, is primitive to all causal queries defined in Section 4. For example, ATE (27) considers $\mathbb{S} = \mathbb{U}_T$, whereas AIE (34), ARE (35), and AOE (36) consider $\mathbb{S} = \{\mathbf{x}\} \cup \mathbb{P}(\mathbf{x})$. Our goal is to use the assumptions encoded in a GCD $G(\Phi^{\Delta})$ to rewrite $\texttt{Pr}(Y[\mathbf{x}'] \mid \texttt{do}(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}))$ into an equivalent probabilistic formula in terms of a set of carefully chosen *observed covariates* to adjust for. If such rewriting exists, then

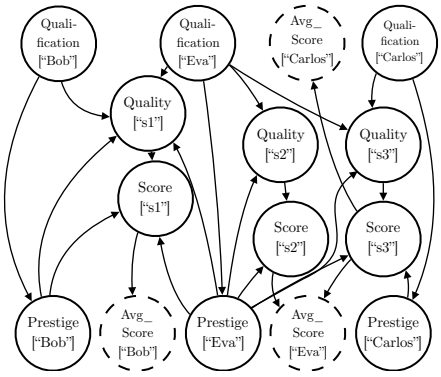

Figure 5: Extending the GCD in Fig 4 with the aggregated attribute *AVG_Score*[*A*].

$\texttt{Pr}\Big(Y[\mathbf{x}'] = y \mid \texttt{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big)$ can be estimated from data, thereby the causal queries can be answered.

First, we extend a GCD $G(\Phi^{\Delta})$ to account for the embedding functions $\Psi^{\mathsf{A}} = \{\psi^{\mathsf{A}}_{\mathsf{A}_1}, \psi^{\mathsf{A}}_{\mathsf{A}_2}, \ldots\}$ for each $\mathsf{A} \in \mathbf{A}$, that map the parents $\mathbf{Pa}(\mathsf{A}(\mathbf{x}))$ of $\mathsf{A}(\mathbf{x})$ in GCD $G(\Phi^{\Delta})$ into a low-dimensional embedding space as given in (**??**). For instance, in Example 5.1, $\psi^{Score}_{Prestige}$ embeds the *Prestige* of the (possibly varying number of) authors of submissions into vectors with a fixed dimension for all submissions. The motivation is to enable the search for a low-dimensional embedding representation of covariates, which in turn enables efficient query answering from finite data (see Section 5.2). To this end, with abuse of notation, we extend the set of attribute functions $\mathbf{A}$ with a collection of *relational embedding* attribute functions $\Psi^{\mathsf{A}} = \{\psi^{\mathsf{A}}_{\mathsf{A}_1}[\mathbf{X}], \psi^{\mathsf{A}}_{\mathsf{A}_2}[\mathbf{X}], \ldots\}$ for each $\mathsf{A} \in \mathbf{A}$, such that for each $\mathbf{x} \in \mathbb{U}_{\mathbf{A}}$, $\psi^{\mathsf{A}}_{\mathsf{A}_i}[\mathbf{x}]$ corresponds to the result of applying the mapping $\psi^{\mathsf{A}}_{\mathsf{A}_i}$ to a subset of $\mathbf{Pa}(\mathsf{A}(\mathbf{x}))$ consisting of the grounding of A_i. Then, we augment a GCD $G(\Phi^{\Delta})$ by inserting the grounding of the relational embeddings as intermediate vertices between a ground atom and its parents.

Example 5.1. In Example 3.6, the relational relational embedding attribut $\psi^{Score}_{Prestige}(S)$ now corresponds to the result of applying the function $\psi^{Score}_{Prestige}$ to map the *Prestige* attributes of the authors of a submission (of the *Author* entities) to a new attribute of the corresponding *Submission* entities (see Figures **??**, 4, 5 for the input GCDs, and Figure 6 for the output GCD with augmented attributes).

We show the following theorem in the technical report.

Theorem 5.2 (Relational Adjustment Formula). Given an augmented GCD $G(\Phi^{\Delta})$, the quantity $\texttt{Pr}\Big(Y[\mathbf{x}'] = y \mid \texttt{do}\big(T[\mathbb{S}] =$

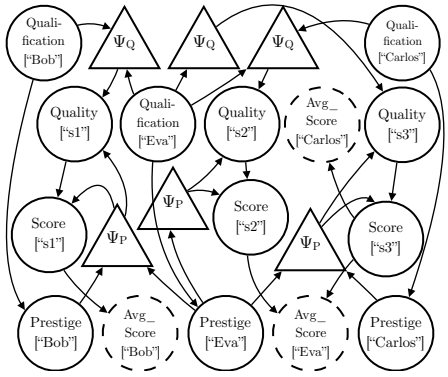

Figure 6: Augmenting the GCD in Fig 5, for clarity $\psi^{Quality}_{Qualifications}[S]$ represented as ψ_Q and $\psi^{Quality}_{Prestige}[S]$ and $\psi^{Score}_{Prestige}[S]$ compactly represented as ψ_P

$\vec{t}_{\mathbb{S}}\big)\Big)$ is given by the following *relational adjustment formula*:

$$\mathtt{Pr}\Big(Y[\mathbf{x}'] = y \mid \mathtt{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big) = \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \mathtt{Pr}\big(Y[\mathbf{x}'] = y \mid \mathbf{Z} = \mathbf{z}, T[\mathbb{S}'] = \vec{t}_{\mathbb{S}'}\big)\mathtt{Pr}(\mathbf{Z} = \mathbf{z}) \tag{38}$$

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

$$\Big(Y[\mathbf{x}'] \perp\!\!\!\perp \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big) \;\big|_{G(\Phi^{\Delta})}\; \bigcup_{\mathbf{x}\in\mathbb{S}} T[\mathbf{x}],\ \mathbf{Z}\Big) \tag{39}$$

Notice that the relational adjustment formula ignores the treated units in $\mathbb{S}\backslash\mathbb{S}'$, simply because, $\mathtt{Pr}\Big(Y[\mathbf{x}'] = y \mid \mathtt{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big) = \mathtt{Pr}\Big(Y[\mathbf{x}'] = y \mid \mathtt{do}\big(T[\mathbb{S}'] = \vec{t}_{\mathbb{S}'}\big)\Big)$. The following observation immediately follows:

Observation 5.1. *The d-separation condition (39) always satisfied for* $\mathbf{Z} = \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big)$ *(any subset of vertices d-separates itself from the rest of the vertices). Hence, adjusting for the joint distribution of the parents of the treated atoms is* always *sufficient for identification. However, if some of the parents are unobserved or redundant for adjustment, one can use the condition (39) to select a* minimal set of observed covariates *that are sufficient for identification.*

We illustrate with an example.

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

$$\mathtt{Pr}\Big(Score[s] = 1 \mid \mathtt{do}\big(Prestige[\{\text{“}Bob\text{”}, \text{“}Eve\text{”}, \text{“}Carlos\text{”}\}] = \vec{1}\big)\Big) \tag{40}$$

where we intervene on all three authors in the example. By applying Theorem 5.2 and Observation 5.1 for submission s = “s_1” and $\mathbf{Z}$ ={*Qualifications*[“*Bob*”], *Qualifications*[“*Eve*”]} we obtain,

$$(40) = \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \mathtt{Pr}\big(Score[\text{“}s_1\text{”}] = 1 \mid \mathbf{Z} = \mathbf{z}, Prestige[\{\text{“}Bob\text{”}, \text{“}Eve\text{”}\}] = \vec{1}\big)\mathtt{Pr}(\mathbf{Z} = \mathbf{z}) \tag{41}$$

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

$$(40) = \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \mathtt{Pr}\big(Score[\text{“}s_2\text{”}] = 1 \mid \mathbf{Z} = \mathbf{z}, Prestige[\{\text{“}Eve\text{”}\}] = 1\big)\mathtt{Pr}(\mathbf{Z} = \mathbf{z}) \tag{42}$$

For estimating *ATE*(*Quality*, *Score*) (assuming quality is observed), by applying (39) we obtain that for each submission s, $\mathtt{Pr}\Big(Score[\text{“}s\text{”}] = 1 \mid \mathtt{do}\big(Prestige[\mathbb{U}] = \vec{1}\big)\Big)$ can be estimated by adjusting for the embedded attribute functions $\mathbf{Z}$ = $\{\psi^{Score}_{Prestige}[\text{“}s\text{”}], \psi^{Score}_{Qualifications}[\text{“}s\text{”}]\}$.

For estimating *ATE*(*Quality*, *AVG_Score*) (the effect on average acceptance rate of an author) we need to estimate $\mathtt{Pr}\Big(AVG\_Score[A] = y \mid \mathtt{do}\big(Prestige[\mathbb{U}] = \vec{1}\big)\Big)$, for each author. According to Theorem (39), this can be done by adjusting for the joint distribution of the qualifications of *all* of their past coauthors, which is potentially very high-dimensional.

Algorithm 1: Constructing unit-table.

Input: An augmented GCD $G(\Phi^{\Delta})$, a treated and outcome attribute functions attribute function $T[\mathbf{X}]$ and $Y[\mathbf{X}']$.
Output: The unit table $\mathbf{D}(Y, \mathbf{H}_T, \mathbf{H}_{\mathbf{Z}})$

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 $\hat{G}(\Phi^{\Delta})$ that satisfies d-separation statement in Eq (39)
4 $\quad \psi_T \leftarrow \psi^T_Y(\langle T[\mathbf{x}_1],\ ldots, T[\mathbf{x}_{|\mathbb{U}'_T|}]\rangle)$
5 $\quad \Psi_{\mathbf{Z}} \leftarrow \Psi^{\mathbf{Z}}_Y(\mathbf{Z})$
6 $\quad$ Insert the tuple $(Y[\mathbf{x}], \psi_T[\mathbf{x}], \Psi_{\mathbf{Z}}[\mathbf{x}]))$ to unit table $\mathbf{D}$

5.2 Covariate Adjustment

This section addresses the following challenges that arise in estimating the causal queries in Section 4 using the relational adjustment formula (38) (see Section 5.1).

(1) High dimensional distributions. The first challenge is that the relational adjustment formula (38) consists of two potentially high-dimensional distributions $\mathtt{Pr}\big(Y[\mathbf{x}] = y \mid \mathbf{Z} = \mathbf{z}, T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)$ and $\mathtt{Pr}(\mathbf{Z} = \mathbf{z})$ that are unknown and must be estimated from limited data (see Example 5.3).

(2) Variability of treatment and covariate vectors. The causal queries need to compute *averages* across all response units. Hence, we need to estimate $\mathtt{Pr}\Big(Y[\mathbf{x}'] = y \mid \mathtt{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big)$ for all $\mathbf{x}' \in \mathbb{U}_Y$. This is challenging since Theorem 5.2 can lead to

Unit	Outcome (Y)	Embedded Authors' Treatments (ψ_T^Y)		Embedded Authors' Covariates (ψ_Z^Y)	
Submission ID	Score	Prestige (AVG)	Authors (COUNT)	Experience (AVG)	H-index (AVG)
S_1	0.8	0.5	2	6	30
S_2	0.15	0.2	1	2	10
S_3	0.2	0.5	2	5	15

Unit	Outcome (Y)	Embedded Collaborators' Treatments (ψ_T^Y)		Embedded Collaborators' Covariates(ψ_Z^Y)	
Author ID	AVG_Score	Prestige (AVG)	Centrality (COUNT)	Experience (AVG)	H-index (AVG)
Bob	1	1	1	2	10
Carlos	0	1	1	2	10
Eve	0.5	2	0.5	9	35

Table 1: The unit table for $T[\mathbf{X}] = Prestige[A]$ and $Y[\mathbf{X}'] = AVG\_Score[A]$. The unit table for $T[\mathbf{X}] = Prestige[A]$ and $Y[\mathbf{X}'] = Score[A]$.

different probabilistic formulas for different response units. Then even under the homogeneity assumption, the corresponding probabilistic formulas must be estimated for each $\mathbf{x}' \in \mathbb{U}_Y$, which is not feasible. For instance, in Example 5.3 (41) and (42) must be estimated separately.

To address the aforementioned issues, similar to Section 3, we use a set of embedding functions ψ_T^Y and $\Psi_{\mathbf{Z}}^Y$, respectively to project the treatment and covariates vectors into a low-dimensional embedding space with *fixed dimensionality* and for all response units. Note that it implies from the homogeneity assumption in Section 3 that the embedding representation of the treated and covariates are *identically distributed*. This enables us to transform a relational instance to a low-dimensional flat-table that suffices for query answering.

Unit-Table. Given a GCD $G(\Phi^{\Delta})$, a 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 with schema $\mathbf{D}(Y, \psi_T^Y, \psi_{\mathbf{Z}}^Y)$, 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 $\psi_T^Y[\mathbf{X}']$, $\psi_{\mathbf{Z}}^Y[\mathbf{X}']$ (with abuse of notation) are relational embedded attribute functions corresponded to the result of applying $\psi_T^Y, \psi_{\mathbf{Z}}^Y$, respectively to the treatment and covariate vectors.

Example 5.4. Table **??** shows the unit-table corresponding to $T[\mathbf{X}] = Prestige[A]$ (the *submissions* constitute the response units) and $Y[\mathbf{X}'] = \text{Score}[S]$, whereas Table 1 shows the unit-table associated with $T[\mathbf{X}] = Prestige[S]$ and $Y[\mathbf{X}'] = AVG\_Score[A]$ (the *Authors* constitute the response units). In these two tables, simple mappings as average and the number of treated units are used for embedding.

By rewriting the RHS of the relational adjustment formula (38) in term 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:

$$\sum_{\mathbf{z} \in Dom(\psi_{\mathbf{Z}}^Y)} \Pr(Y = y \mid \psi_{\mathbf{Z}}^Y = \mathbf{z}, \psi_T^Y = \vec{t}^e_{\mathbb{S}'}) \Pr(\psi_{\mathbf{Z}}^Y = \mathbf{z}) \qquad (43)$$

Now, all causal queries in Section 4 can be estimated using standard techniques in statistics for estimating (43) using the unit table $\mathbf{D}$.

Dataset	Att. [#]	Rows [#]	Embedding Choice			Query Ans.
			Predefined	RNN	Forest	
ReviewData	7	6K	0.2	480	2.5	180
SyntheticData	7	300K	11	24,000	x	400

Table 2: Runtime in seconds for experiments.

Note that the idea of embedding addresses both issues of high dimensionality and variable size of the treatments and covariates corresponding to the response units, thereby makes the estimation of all causal queries feasible. However, (43) only approximates (38), hence the quality of the answers depend on whether the embeddings preserve sufficient statistics. In Section **??**, we develop several techniques to learn this embeddings from data.

6 EXPERIMENTS

This section presents experiments that evaluate the feasibility and efficacy of CaRL. We aim to answer the following questions: **Q1:** What is the end-to-end performance of CaRL ? In particular, can the system (1) avoid confusing correlation with causation; and (2) consider the relational structure in multi-relational settings rather than naively applying traditional causal inference methods on the *universal table* obtained by joining all base relations? **Q2:** What is the sensitivity of query answering component of CaRL to different embedding methods? Additionally, we report the performance of CaRL in Table 2.

6.1 Setup

This section describes the datasets and causal queries we used in the experiments.

Real Dataset. The ReviewData consists of 2,075 papers submitted for review across 10 computer science conferences and workshops. Each submission is associated with a number of referee reviews, as well as an acceptance or rejection decision. About half of all submissions are double-blind, while in the other half the names of the authors are revealed. All submissions have been 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 that contributed to a paper in the dataset.

Synthetic Dataset. Babak: [@Harsh] To answer both questions we needed the ground truth, so we generated SyntheticData data according to a slightly modified version of the relational causal model in Example 1.1. . Specifically, ... explain how you generated the relational skeleton .... and then ... given them the ground truth.

Causal Queries. We used both ReviewData and SyntheticData to answer the following causal queries.

$$Score[S] \Leftarrow Prestige[A] \; ? \tag{44}$$

$$Score[S] \Leftarrow Prestige[A] \; ? \; \texttt{WHEN} \; \langle LESS \; THAN \; 35\% \rangle \; \texttt{PEERS TREATED} \tag{45}$$

where, (44) quantifies *ATE* of *Prestige* on *Score*, and (45) quantifies the *ATE*, ARE, AIE effect of having less than and more than one third of the prestigious collaborator on submissions scores (cf. 4). The threshold of 1/3 was as chosen as it resulted maximum correlation between prestige and review score. However, in addition to *ATE*, ARE, AIE and AOE as defined in Section 4, we also report these quantities as per each level of covariates $\mathbf{Z} = \mathbf{z}$. We denote the corresponding conditional quantities, respectively as *CATE*, *C*ARE, *C*AIE and *C*AOE. For instance *CATE* quantifies the *ATE* of prestige on submission scores as per each level of authors qualifications (the covariates). The motivation is to explore the variability of causal effects across different sub populations. Also note we used Pearson's correlation coefficients to measure the association between the *Prestige* and *Score* and *AVG_Score*.

Note that for estimating the *ATE*, since the ReviewData is sparse (*i.e.*, there are only a few submissions with and without prestigious authors), we further aggregated the treatment vector associated to each authors (as in Figure1). Specifically, we considered submissions to be treated if 1/3 of the authors are from prestigious universities and untreated otherwise.).

Babak: [@harsh: is there any sub-population for which the effect is significantly large? Then we should report that.]

Answering Queries and Embeddings. For query answering we used Algorithm 1.1 for creating unit tables. The tables were generated using the same relational causal model as in Example 1.1. Then the unit tables used for estimating the relational adjustment formula as in (43) for both treatment vectors the queries concern about.. Note that we used Random Forest Regression for estimating the underling conditional probability stributio For emben.ddings we used basic for ediaggregation functions such as mean and median together with the cardinally of the vectors to account for the underling topology of the relational skeleton (e.g, number of authors, number of co authors). We also the vector of K moments for embedding, *i.e.*, the vector consist of mean, variance, skewness, etc., when we learn k automatically from data. In general, when the vectors are ordered or have temporal pattern the basic aggregations and moment summarization lead to information loos, which would negatively impact the quality of query answer. However, for experiment all these methods turned out to be effective. In future we plan to use techniques such as convolutional graph neural networks and Recurrent neural networks.

CaRL

covariate detection and creating the corresponding the unit table. For estimating the conditional probability distribution involved in the probabilistic formulas (43) we used Random Forest regression.

.

6.2 Q1: End-To-End Results

For this experiment we report the answers to our causal queries on both real and synthetic data. We used the the embedding method for computing these results.

ReviewData. Figure 7(top), compares the ATE as in (44) and the naive correlation between the prestige and scores across the single and double-blind conferences. The results shows a significant correlation in both single and double blind conferences. However, after conducting a sound causal analysis using CaRL, it turned out that there is no significant causal effect double-blind conferences. However, the result shows a significant causal effect turned for single-blind conferences. This results suggest double-blind reviewing can be effective in reducing bias toward prestigious institutions

To further investigate the effect of prestige on review score in single-blind conferences, we used CaRL compute isolated and relational and overall effect of prestige on average review score as in (45). Figure 7(button) reveals that the isolated effect (AIE) is more significant than the relational effect (ARE), meaning that author's own prestige have stronger effect on their average submission score than their collaborators prestige, which is expected. Furthermore, one can verify that approximately AOE = AIE + ARE, which confirms Theorem 4.3.

Notice the response variable underling the analysis in Figure 7(top) is the review score of each paper, whereas in Figure 7(bottom) is the average review scores for each author (corresponded to the tables in Figure 1). Therefore, the correlations in obtained in Figure 7(top) and Figure 7(bottom) are diffident. by naively interpreting these correlation as causation and might arrive at the wrong conclusion that double-blinding is not effective in reducing bias. While the validity

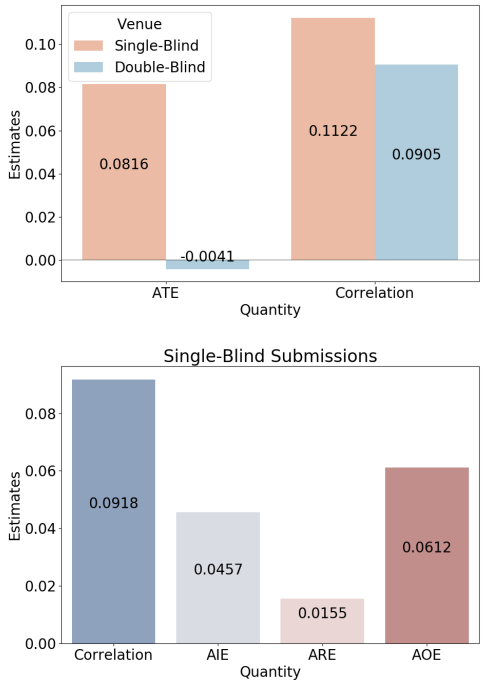

Figure 7: Average treatment effects estimates and difference of sample means of treated and control units for single-blind and double-blind submissions.

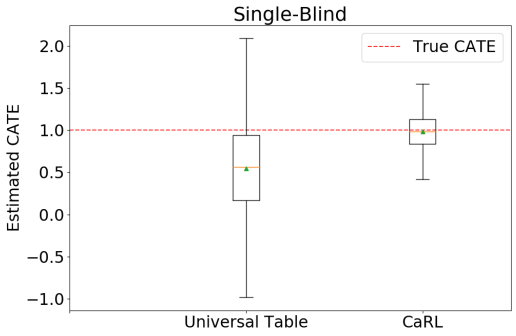

Figure 8: Universal Table vs CaRL

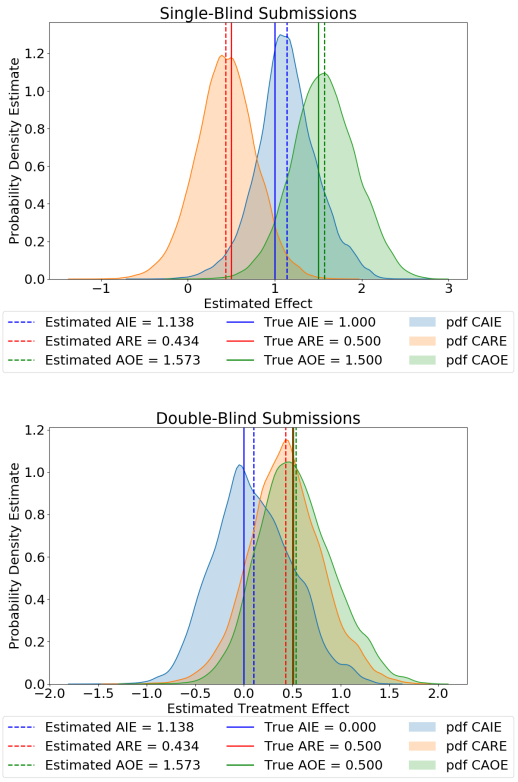

Figure 9: Probability density function and corresponding averages for isolated treatment effect, relational treatment effect and total effects for (a) Single-Blind submission and (b) Double-blind submission.

of results depend on the validity of the underling assumption, are indeed insightful and goes way beyond associational analysis. We note that our results are in accordance with a series of controlled experiments that suggest double-blind reviewing reduces institutional prestige bias [3, 8, 12, 13].

SyntheticData. To further investigate the efficacy of CaRL, we used it to answer the same queries from SyntheticData for which we knew the ground truth. Figure 8, compares *CATE* (conditional *ATE*) of prestige on submission scores on SyntheticData computed using CaRL with the naive computation of CATE on the universal table obtained by joining all based relations. As shown, CaRL approximately recovers the ground truth. The variance is expected and is the product of limited data. However, naive causal analysis on the universal table leads to *incorrect ATE with huge variance.* This experiment reveals that ignoring the relational structure in relational domains can lead to incorrect conclusions.

Furthermore, Figure 9 compares the answers to both queries in (44) and (45) computed with CaRL with the ground truth. As shown in all cases CaRL approximately recovered the the ground truth. Note that the probability density functions shows Babak: [@Harsh]

6.3 Q2: Sensitivity and Learnability of Embeddings

Figure 10 box-plots shows the distribution of estimated conditional average treatment effects (CATEs) under different choices of embedding functions for a) single-blind venue

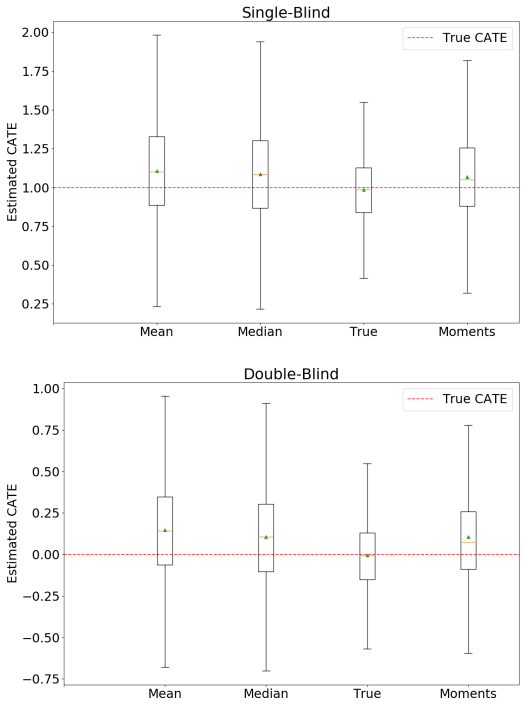

Figure 10: Comparing the sensitivity of of the quality of query answer (CATE) to different choice of embeddings.

submissions, and b) double-blind venue submissions. Here, “Mean" refers to using mean of the set as the hardcoded embedding of the set, “Median" refers to using median of the set as the hardcoded embedding of the set. “True" refers to using harmonic mean as the embedding for authors’ experience and weighted mean of the log of authors’ citation of each paper as the embedding, which are close to true generative model. The learned embedding uses optimal “Moments" of the set as the embedding. Here, we primarily observe that even the simple embeddings like mean or median are able to recover the true average treatment effect approximately. While the true embedding is able to recover almost the exact truth, the learned moment summarization embedding (which learns the optimal number of moments for accurate outcome prediction) also shows promising results. It is important to note that moment summarization is one of the most basic learning approach for set embedding and more complex approaches using recurrent neural networks or kernel density estimators can also be used. Across, both single and double-blind reviews estimates, we observe from Figure 10 that the estimated mean is almost approximately same for all methods while the variances for complex embedding is the smallest followed by the variance for the learned moment summarizing embedding.

Next, we generated the data using the second DGP where the review score has both a constant isolated effect of median prestige of the authors equal to 1 and a constant relational effect of mean prestige of the coauthors of the authors equal to 0.5.

7 RELATED WORK AND DISCUSSIONS

Statistical Relational Learning (SRL) aims at unifying logic and probability to model a joint probability distribution over relational data amenable for probabilistic reasoning [2]. A plethora of models such as PRMs [1], PSL, DAPER, PBNs, BLPs and MLNs has been proposed for SRL in the literature. While our proposed framework for Causal Relational Learning, CaRL, shares some similarities with some of these models, it is fundamentally motivated for causal reasoning. The closest SRL model to CaRL are PRMs and DAPER that extend the standard entity-relationship (ER) model to incorporates probabilistic dependencies. They can also be seen as a first-order extension of Bayesian Networks to relational data. However, there are two primary differences between CaRL and these models. First, PRMs are dependency models that essentially define a joint probability distribution over all possible completions of a partially specified instance. In contrast CaRL models the generative process of relational data. Here, the uncertainty is induced by hidden noise variables, *and prevails even when the entire relational data is available.* There has been prior attempt to use use these models for causal inference from relational data, e.g.,

8 CONCLUSIONS

TODOS:

- Do we need FKs and other constraints?
- Do we need to define ER model?

REFERENCES

[1] Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, and Benjamin Taskar. Probabilistic relational models. *Introduction to statistical relational learning*, 8, 2007.

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

[3] 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.

[4] OpenReviwe. https://openreview.net.

[5] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

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

[8] 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.

[9] 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.

[10] Semantic Scholar. www.semanticscholar.org/.

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

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

[13] 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.

[14] Jeffrey D. Ullman. *Principle of Database Systems*. Pitman, 2nd edition, 1982.

9 SUPPLEMENTARY MATERIAL

This section includes the proof of theorems and propositions in the paper.

Proof of Proposition **??**. (Sketch) Markov compatibility $\Pr(\mathbf{A}^{\Delta})$ and $G(\Phi^{\Delta})$ implied from the the fact that $\Pr(\mathbf{A}^{\Delta})$ satisfies *global Markov property, i.e.*, for any set of ground atoms $\mathbf{N}$ that are non-descendants of a ground atom $\mathsf{A}[\mathbf{x}]$ in $G(\Phi^{\Delta})$ it holds that (*) $\big(\mathsf{A}[\mathbf{x}] \perp\!\!\!\perp \mathbf{N} \mid \mathbf{pa}(\mathsf{A}[\mathbf{x}])\big)$ (implied from the functional dependence between $\mathbf{pa}(\mathsf{A}[\mathbf{x}])$ and $\mathsf{A}[\mathbf{x}]$ and the assumption of causal sufficiency. Furthermore, it follows from the assumption in Eq 25 that (**) $\big(\mathsf{A}[\mathbf{x}] \perp\!\!\!\perp \mathbf{Pa}(T[\mathbf{x}]) \perp\!\!\!\perp | \Psi^{\mathsf{A}}\big(\mathbf{Pa}(\mathsf{A}[\mathbf{x}])\big)$.

The factorization in Eq 26 immediately follows from the following independence obtained by applying the contraction axiom in Graphoid to (*) and (**).

$$A(\mathbf{x}) \perp\!\!\!\perp \mathbf{N} \mid \Psi^{\mathsf{A}}\big(\mathbf{Pa}(\mathsf{A}[\mathbf{x}]\big) \tag{46}$$

□

Proof of Theorem 5.2. (Sketch) Since Ψ^{Δ} are deterministic functions the ground random variables $\mathbf{A}^{\Delta}$, they are also random variables, hence we can define the joint probability distribution $\Pr(\mathbf{A}^{\Delta}, \Psi^{\Delta})$. Also note that that the parents of an atom in the augmented GCD $\hat{G}(\Phi^{\Delta})$ corresponded to the embedded parents of the same node in GCD $G(\Phi^{\Delta})$.

Since each atomic intervention $do(T[\mathbf{x}_i] = t_i)$ modifies the augmented GCD $\hat{G}(\Phi^{\Delta})$ by removing the parents of $T[\mathbf{x}_i]$ from $\hat{G}(\Phi^{\Delta})$ (implied from the factorization Eq 26), the post intervention distribution $\Pr\Big(\mathbf{A}^{\Delta}, \Psi^{\Delta} \mid \mathsf{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big)$ can be obtained from the pre-intervention (observed) distribution $\Pr(\mathbf{A}^{\Delta})$ by removing all factors $\Pr\Big(\mathsf{A}(\mathbf{x}), \mid \mathbf{Pa}(\mathsf{A}(\mathbf{x}))\Big)$, from $\Pr(\mathbf{A}^{\Delta}, \Psi^{\Delta})$ (cf. Eq 26), hence we obtain the following:

$$\Pr\Big(\mathbf{A}^{\Delta}, \Psi^{\Delta} \mid \mathsf{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big) = \frac{\Pr(\mathbf{A}^{\Delta})}{\prod_{\mathbf{x}\in\mathbb{S}} \Pr\Big(\mathsf{A}(\mathbf{x}) \mid \Psi^{\mathsf{A}}\big(\mathbf{Pa}'(\mathsf{A}(\mathbf{x}))\big)\Big)} \tag{47}$$

The following factorization implied by the chain rule of probability:

$$\Pr(\mathbf{A}^{\Delta}, \Psi^{\Delta}) = \Pr\Big(\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big)\Big) \Pr\Big(T[\mathbf{x}_1] \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}(T[\mathbf{x}])\Big) \tag{48}$$

$$\Pr\Big(T[\mathbf{x}_2] \mid T[\mathbf{x}_1], \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}(T[\mathbf{x}])\Big) \tag{49}$$

$$\vdots$$

$$\Pr\Big(T[\mathbf{x}_i] \mid \bigcup_{j=0}^{i-1} T[\mathbf{x}_j], \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big)\Big) \tag{50}$$

$$\vdots$$

$$\Pr(\overline{\mathsf{A}^{\Delta}}, \overline{\Psi^{\Delta}} \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big), \bigcup_{\mathbf{x}\in\mathbb{S}} T[\mathbf{x}]) \tag{51}$$

where $\overline{\mathsf{A}^{\Delta}} \cup \overline{\Psi^{\Delta}}$ consist of all ground atoms in $\mathbf{A}^{\Delta} \cup \Psi^{\Delta}$ except for $\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big)$ and $\bigcup_{\mathbf{x}\in\mathbb{S}} T[\mathbf{x}]\big)\}$ The above factorization is valid if $\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{Pa}\big(T[\mathbf{x}]\big)$ and $\bigcup_{\mathbf{x}\in\mathbb{S}} T[\mathbf{x}]$ are disjoint, which is true in acyclic $\hat{G}(\Phi^{\Delta})$.

The following implied from the factorization in Eq. 51, the generic conditional independence in Eq. 46, and Eq 47, and marginalization we obtain the following:

$$\Pr\Big(Y[\mathbf{x}'] = y \mid \mathsf{do}\big(T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big)\Big) = \sum_{\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)} \Pr\big(Y[\mathbf{x}'] = y \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}(T[\mathbf{x}]), T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\big) \Pr\big(\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\big(T[\mathbf{x}]\big)\big) \tag{52}$$

Now, given a set $\mathbf{Z} \subseteq \Psi^{\Delta}$, we can rewrite the RHS of Eq. 52 into the following equivalent formula:

$$
\begin{aligned}
&\text{RHS} = \\
&\sum_{\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)} \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \Pr\left(Y[\mathbf{x}'] = y \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right), T[\mathbb{S}] = \vec{t}_{\mathbb{S}}, \mathbf{Z} = z\right) \\
&\qquad \Pr(\mathbf{Z} = z \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right), T[\mathbb{S}] = \vec{t}_{\mathbb{S}}) \Pr(\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right))
\end{aligned} \tag{53}
$$

Now from the conditional independence in Eq.39 and the conditional independence statments $T[\mathbf{x}] \perp\!\!\!\perp \mathbf{Z}, \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right) \mid \mathbf{pa}(T[x])$ (implied from the generic conditional independence in Eq 46)) for each $\mathbf{x} \in \mathbb{S}$, Eq **??** can be simplified as follows:

$$
\begin{aligned}
&\text{RHS} = \\
&\sum_{\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)} \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \Pr\left(Y[\mathbf{x}'] = y \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right), T[\mathbb{S}] = \vec{t}_{\mathbb{S}}, \mathbf{Z} = z\right) \\
&\qquad \Pr(\mathbf{Z} = z \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right), T[\mathbb{S}] = \vec{t}_{\mathbb{S}}) \Pr(\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)) \\
&= \sum_{\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)} \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \Pr\left(Y[\mathbf{x}'] = y \mid T[\mathbb{S}] = \vec{t}_{\mathbb{S}}, \mathbf{Z} = z\right) \\
&\qquad \Pr(\mathbf{Z} = z \mid \bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)) \Pr(\bigcup_{\mathbf{x}\in\mathbb{S}} \mathbf{pa}\left(T[\mathbf{x}]\right)) \\
&= \sum_{\mathbf{z}\in Dom(\mathbf{Z})} \Pr\left(Y[\mathbf{x}] = y \mid \mathbf{Z} = \mathbf{z}, T[\mathbb{S}] = \vec{t}_{\mathbb{S}}\right) \Pr(\mathbf{Z} = \mathbf{z})
\end{aligned} \tag{54}
$$

This completes the proof. □

9.0.1 Data-Generative Process. We now briefly explain the data generative process (DGP) for the synthetic data experiments and highlight the different levels of complexity. We generate each institution i in the set of Institutes using the following generator:

$$
\text{Prestige}_i \sim \text{exponential}(\lambda = 0.1).
$$

Similarly, an author a in set of Authors is generated using the following generative process:

$$
\begin{aligned}
\text{Experience}_a &\sim \text{Bernoulli}(1/3) \times \text{Exponential}(3/4) \\
&+ \text{Bernoulli}(1/2) \times \text{Normal}(20, 4) \\
&+ \text{Bernoulli}(1/4) \times \text{Normal}(35, 1) \\
\text{Gender}_a &\sim \text{Bernoulli}(1/2) \\
\text{Citation}_a &\sim \text{Bernoulli}(1/2) \times \text{Poisson}(100) \\
&+ \text{Bernoulli}(1/3) \times \text{Poisson}(200) \\
&+ \text{Bernoulli}(1/5) \times \text{Poisson}(500) \\
&+ \text{Bernoulli}(1/16) \times \text{Poisson}(1000) \\
\text{Expertise}_a &\sim \text{Multinomial}(1, \{1/10 \ldots 1/10\}_{10}).
\end{aligned}
$$

For each author a we generate their affiliation to an institute using following process. We use this affiliated institute's prestige as author's prestige in our experiments:

$$
\begin{aligned}
\text{AfP}(a, i) &= expit\Big(\frac{\text{Prestige}_i \text{Citation}_a}{10} \\
&+ \frac{305}{\text{Prestige}_i(\text{Citation}_a + 1)} \\
&- \frac{305}{10} \frac{\text{Prestige}_i}{\text{Citation}_a + 1} \\
&- 305 \frac{\text{Citation}_a}{\text{Prestige}_i}\Big) \times \frac{30}{\text{Experience}_a} \\
\text{Affiliation}_a &\sim \text{Dice-Roll}\Big(N = |\text{Institutes}|, \\
&\quad P = \{\text{AfP}(a, i)\}_{i\in\text{Institutes}}\Big) \\
\text{Prestige}_a &= \text{Prestige}_{\text{Affiliation}_a}
\end{aligned}
$$

Similarly, we generate each conference c in set of conferences as follows:

$$
\begin{aligned}
\text{Impact}_c &\sim \text{Exponential}(0.1) \\
\text{Area}_c &\sim \text{Multinomial}(1, \{1/10 \ldots 1/10\}_{10}) \\
\text{SingleBlind}_c &\sim \text{Bernoulli}(2/3).
\end{aligned}
$$

Finally, for each paper p in the set of paper submissions, the number of authors are randomly drawn from 1+Exponential(2/5) and then for the purpose of simulation we let the authors for each paper be randomly chosen without replacement from the set of authors. We also let each paper be submitted to any conference uniformly at random. This is just to keep the generation process less complicated while driving the major point home. Other parameters for the paper are generated as follows:

$$
\begin{aligned}
\text{Quality}_p &= \text{expit}\left(\sum_{a\in\text{Authors}_p} e^{-2a} \log\left(\frac{30 \times \text{Citation}_a}{\text{Experience}_a + 5} + 1\right)\right) \\
\text{Review}_p^{(1)} &= 3 + \big(\text{SingleBlind}_{\text{Conference}_p} \\
&\times \big(median(\{\text{Prestige}_a\}_{a\in\text{Authors}_p}) > 10\big)\big) \\
&+ \frac{20}{3}\Big(\text{Normal}\Big(log(\text{Quality}_p + 1) \\
&- \log(\text{Impact}_{\text{Conference}_p})/10, 0.1 * \text{Quality}_p\Big)\Big).
\end{aligned}
$$

We can observe from the generative model of the review score of a paper that for a submission in single blind conference the treatment effect of the median prestige of the authors more than 10 is 1 while for a submission in a double-blind conference, the analogous treatment effect is 0. In this data generation setup, there is no relational effect. In the following section, we will show CaRL's ability to recover true treatment effect under different embeddings of paper's authors' features.

We define a second possible data generator for review score of the paper where we have a positive relational effect:

$$
\begin{aligned}
\text{Review}_p^{(2)} &= 3 + \big(\text{SingleBlind}_{\text{Conference}_p} \\
&\times \big(median(\{\text{Prestige}_a\}_{a\in\text{Authors}_p}) > 10\big)\big) \\
&+ \frac{20}{3}\Big(\text{Normal}\Big(log(\text{Quality}_p + 1) \\
&- \log(\text{Impact}_{\text{Conference}_p})/10, 0.1 * \text{Quality}_p\Big)\Big) \\
&+ \frac{1}{2}(mean(\{\text{Prestige}_{a'}\}_{a'\in(\cup_{a\in\text{Authors}_p}\text{Coauthor}_a \setminus \text{Authors}_p)}) > 10).
\end{aligned}
$$

In the above mentioned DGP, the isolated treatment effect for single blind submissions is 1 while it is 0 for double blind submission. However, the relational treatment effect of the mean prestige of coauthors being more that 10 is 0.5 for both single-blind and double-blind submissions.

Proof.

$$\begin{aligned}\mathrm{AOE}(t,\vec{t};t',\vec{t}') &= \frac{1}{n}\sum_{\mathbf{x}\in\mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t,\vec{t}) - Y_{\mathbf{x}}(t',\vec{t}') \\ &= \frac{1}{n}\sum_{\mathbf{x}\in\mathbb{U}_{(T,Y)}} Y_{\mathbf{x}}(t,\vec{t}) - Y_{\mathbf{x}}(t',\vec{t}) + Y_{\mathbf{x}}(t',\vec{t}) - Y_{\mathbf{x}}(t',\vec{t}') \\ &= \mathrm{AIE}(t,t' \mid \vec{t}) + \mathrm{ARE}(\vec{t},\vec{t}' \mid t')\end{aligned}$$

Similarly the second equality in (37) holds. □