Demonstration of Inferring Causality from Relational Databases with CaRL

Moe Kayali, Babak Salimi, Dan Suciu
{kayali, bsalimi, suciu}@cs.washington.edu
University of Washington

ABSTRACT

Understanding cause-and-effect is key for informed decision-making. The gold standard in causal inference is performing controlled experiments, which may not always be feasible due to ethical, legal, or cost constraints. As an alternative, inferring causality from observational data has been extensively used in statistics and 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 that are best represented using relational databases. We propose and demonstrate CaRL: an end-to-end system for drawing causal inference from relational data. In addition, we built a visual interface to wrap around CaRL. In our demonstration, we will use this GUI to show a live investigation of causal inference from real academic and medical relational databases.

PVLDB Reference Format:
Moe Kayali, Babak Salimi, Dan Suciu. Demonstration of Inferring Causality from Relational Databases with CaRL. *PVLDB*, 13(12): 2985-2988, 2020.
DOI: https://doi.org/10.14778/3415478.3415525

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 [9]. 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?").

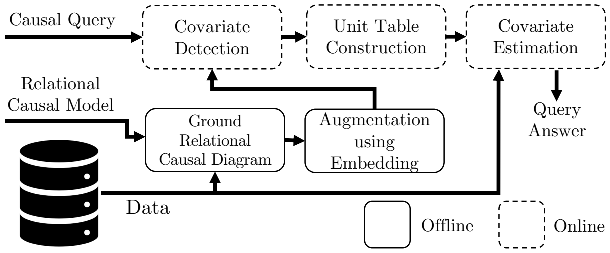


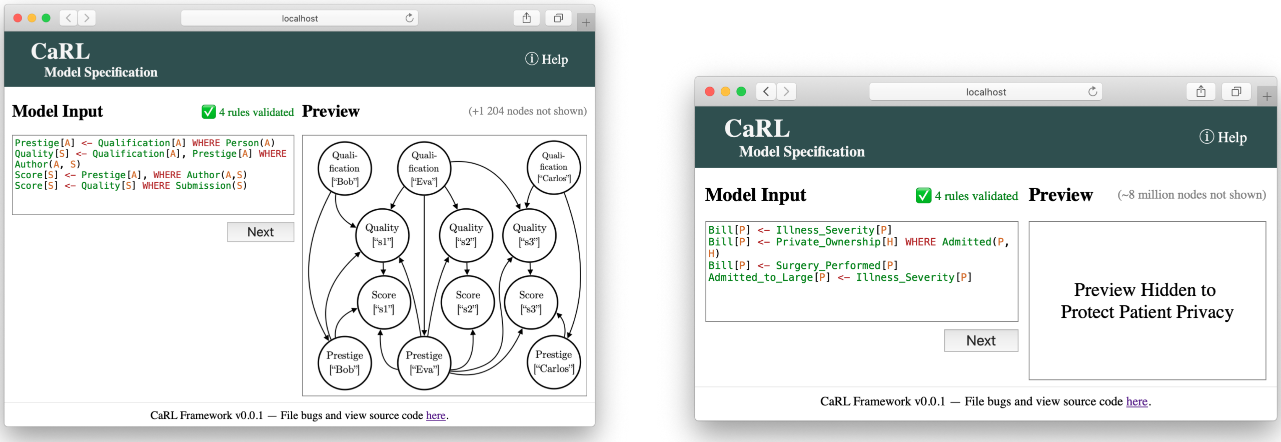
Figure 1: CaRL architecture.

Rubin's Potential Outcome Framework [10] and Pearl's Causal Models [9] are two well-established frameworks which have been extensively explored in the literature and used in various applications for estimating causal effects in observational studies [2, 11, 8, 7, 1]. These 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 [6, 1]. 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 demonstration on real data [5], 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 the subjects who receive a treatment—no longer readily apply to relational data, prohibiting us from adopting existing causal inference frameworks to relational domains.

The CaRL project aims to provide a foundation for causal inference from relational data. At the heart our framework is a declarative language that allows researchers to represent domain knowledge and assumptions as well as ask causal queries on relational data. Our framework gives semantics to complex causal queries where the treatment units and outcome units might be of different types and controlling for confounding may require performing multiple joins and aggregates. Using CaRL, we can answer complex causal queries like: "What is the effect of not having insurance on mortality of a patient?"

In this demonstration, the attendees will go through the three key steps of the causal inference from relational data:

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415525



```
Prestige[A] <- Qualification[A] WHERE Person(A)
Quality[S] <- Qualification[A], Prestige[A] WHERE
Author(A, S)
Score[S] <- Prestige[A], WHERE Author(A,S)
Score[S] <- Quality[S] WHERE Submission(S)
```

localhost
CaRL
Model Specification
ⓘ Help
Model Input
✅ 4 rules validated
Preview
(~8 million nodes not shown)
Next
Preview Hidden to Protect Patient Privacy
CaRL Framework v0.0.1 — File bugs and view source code here.

```
Bill[P] <- Illness_Severity[P]
Bill[P] <- Private_Ownership[H] WHERE Admitted(P,
H)
Bill[P] <- Surgery_Performed[P]
Admitted_to_Large[P] <- Illness_Severity[P]
```

a) ReviewData dataset

b) NIS dataset

Figure 2: Relational causal model specification via CaRL on two datasets.

(1) Articulating and encoding their domain knowledge using a declarative language that consists of Datalog-like rules. (2) Encoding their causal queries using the same declarative language. (3) Estimating the results using an algorithm that 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.

As a benchmark for causal inference methods is not yet available [3], we test on two real-life datasets. Overall, this demonstration makes the following contributions:

- It presents CaRL, a first-of-its-kind system for causal inference and policy evaluation from relational data.
- It demonstrates that ignoring the relational structure of data and performing causal inference on the universal table obtained from joining base tables leads to spurious findings and perplexing insights.
- It enables the attendees to experience the challenges of causal inference from real relational data in the academic and medical domains to find answers to the following interesting causal queries:
  – How effective is single- *vs.* double-blinding at reducing bias in academic peer review?
  – What is the effect of hospital size on the affordability of medical care?

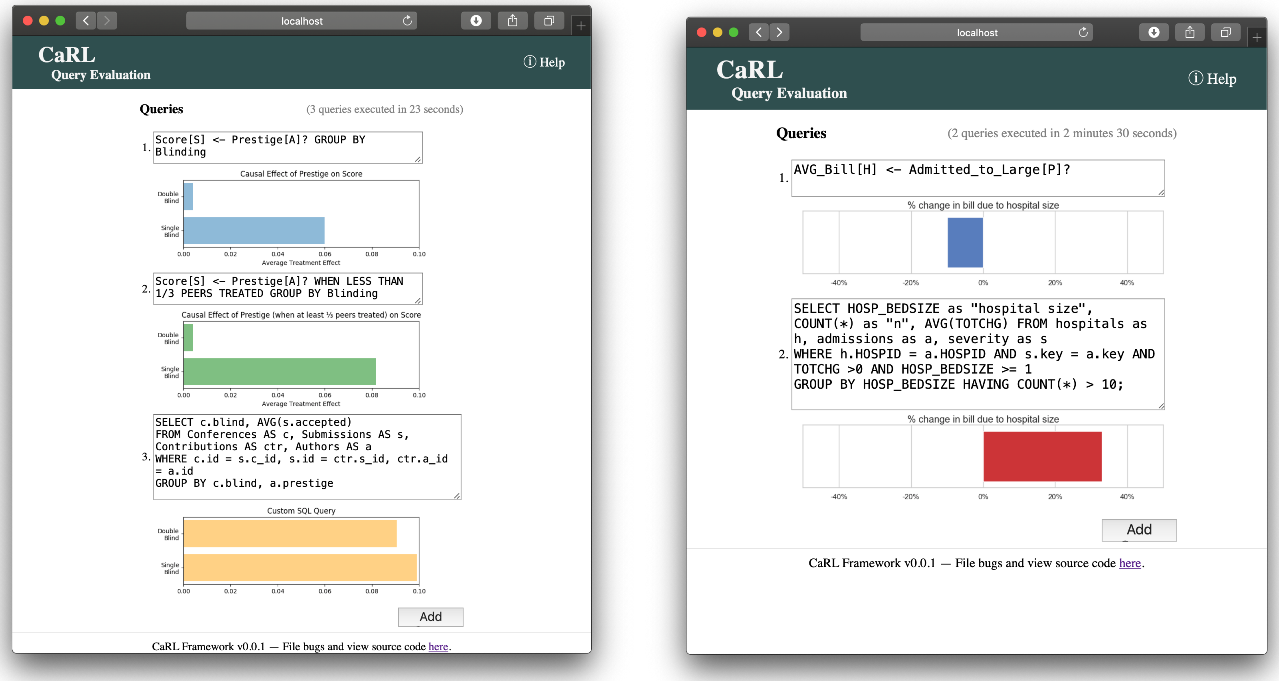
2. ARCHITECTURE

The overall architecture of CaRL is shown in Figure 1. A detailed description of the system can be found on [12]. CaRL allows the users to encode their domain knowledge in terms of *relational causal models*, write *causal queries*, and estimate the query answer based on data stored in a DBMS. Specifically, relational causal models represent (large) causal models in relational domains using a knowledge base that consists of a few number of first order sentences. The actual causal model obtained by *grounding* the formulas in the knowledge base with all constants in a given domain. The ground causal models give semantics to complex external interventions and will be used to answer complex causal queries. CaRL uses techniques such as embedding and aggregation to construct a flat unit-table from which the query answer can be estimated using traditional causal inference methods. A web GUI is provided to facilitate demonstration and exploration, and is shown in Figures 2-3.

3. DEMONSTRATION OUTLINE

Data. In our demonstration, we will operate on two real relational databases. (1) ReviewData which consists of paper submissions to previous academic venues, both accepted and rejected. This is a dataset of 2075 paper submissions by 4490 authors to 10 computer science conferences and workshops, spanning the years 2017–2019. Some venues follow a single-blind review policy, while others follow a double-blind one. In all cases, the identities of the authors have been revealed once review concluded. Each submission is associated with a number of reviews, a numerical score by each of the reviewers, and an acceptance decision. Additionally, scholarly information, such as affiliation and citation count, is associated with each author. (2) NIS which contains medical outcome and insurance data regarding 8 million hospital stays in the US during the year 2006 [5]. This relational dataset contains tables for admissions, hospitals and diagnoses. Columns include type of medical insurance, length of stay, size of hospital and mortality rate. The NIS only contains deidentified data and is about 15GB in size. CaRL uses SQLite and Postgres to store the data in the backend.

Relational Causal Models. Our demonstration will start by specifying relational causal models (cf. Section 2) that encode intuitive background knowledge about the data sets used throughout the demonstration. Relational causal models consist of a set of Datalog-like rules where each rule is composed of a head, body and Boolean conjunctive query (BCQ) and has the format *head* ⇐ *body* WHERE BCQ. The head and body are composed of atomic attributes from the schema, along with free variables and/or constants. For every set of constants that satisfy the BCQ, the rule encodes a causal link between the head and body of the corresponding ground rules obtained by replacing the free variables in the head and body of the rules with these constants.



localhost

CaRL
Query Evaluation

ⓘ Help

Queries (3 queries executed in 23 seconds)

1.
```
Score[S] <- Prestige[A]? GROUP BY
Blinding
```

2.
```
Score[S] <- Prestige[A]? WHEN LESS THAN
1/3 PEERS TREATED GROUP BY Blinding
```

3.
```
SELECT c.blind, AVG(s.accepted)
FROM Conferences AS c, Submissions AS s,
Contributions AS ctr, Authors AS a
WHERE c.id = s.c_id, s.id = ctr.s_id, ctr.a_id
= a.id
GROUP BY c.blind, a.prestige
```

CaRL Framework v0.0.1 — File bugs and view source code here.

localhost

CaRL
Query Evaluation

ⓘ Help

Queries (2 queries executed in 2 minutes 30 seconds)

1.
```
AVG_Bill[H] <- Admitted_to_Large[P]?
```

2.
```
SELECT HOSP_BEDSIZE as "hospital size",
COUNT(*) as "n", AVG(TOTCHG) FROM hospitals as
h, admissions as a, severity as s
WHERE h.HOSPID = a.HOSPID AND s.key = a.key AND
TOTCHG >0 AND HOSP_BEDSIZE >= 1
GROUP BY HOSP_BEDSIZE HAVING COUNT(*) > 10;
```

CaRL Framework v0.0.1 — File bugs and view source code here.

Figure 3: Causal queries, their answers and for comparison, the corresponding naive SQL query. Queries are input and executed one-by-one, similar to cells in a Jupyter notebook.

An an example in our demonstration we use the following simple relational causal model for ReviewData:

$$\text{Prestige}[A] \Leftarrow \text{Qualification}[A] \text{ WHERE Person}(A) \tag{1}$$

$$\text{Quality}[S] \Leftarrow \text{Qualification}[A], \text{Prestige}[A] \text{ WHERE Author}(A, S) \tag{2}$$

$$\text{Score}[S] \Leftarrow \text{Prestige}[A] \text{ WHERE Author}(A, S) \tag{3}$$

$$\text{Score}[S] \Leftarrow \text{Quality}[S] \text{ WHERE Submission}(S) \tag{4}$$

For instance rule (3) encodes the assumption that for every submission S and author A that satisfies the predicate Author(A, S) (*i.e.* "A is an author on S"), the attributes in the body (prestige of the author A) causally affect the attribute in the head (review score P receives). Specifically, this rule is a template for generating ground causal rules obtained by grounding, i.e., replacing the free variables of the rule with constants. For example, given the instance People = $\{Bob, Carlos, Eva\}$, Submissions = $\{s1, s2, s3\}$, Authors = $\{(Bob, s1), (Eva, s1), (Eva, s2), (Eva, s3), (Carlos, s3)\}$ The following ground causal rules are obtained from (3):

$$\text{Score}[s1] \Leftarrow \text{Prestige}[Bob],\ \text{Prestige}[Eva]$$
$$\text{Score}[s2] \Leftarrow \text{Prestige}[Eva]$$
$$\text{Score}[s3] \Leftarrow \text{Prestige}[Carlos],\ \text{Prestige}[Eva]$$

These ground rules can be represented graphically using the causal diagram in Figure 2. Hence, relational causal models can be seen as a template for generating causal diagrams. Similarly, (1) specifies the assumption that the qualification of an author causally affects their prestige, (2) encodes that quality of a paper is causally affected by the prestige (and thus access to resources) and qualification of its authors, and (4) encodes the assumption that the score of a submission causally depends on its quality.

In our demonstration we use the following relational causal model for the medical insurance dataset, NIS:

$$\text{Bill}[P] \Leftarrow \text{Illness\_Severity}[P] \tag{5}$$

$$\text{Bill}[P] \Leftarrow \text{Private\_Ownership}[H] \text{ WHERE Admitted}(P, H) \tag{6}$$

$$\text{Bill}[P] \Leftarrow \text{Surgery\_Performed}[P] \tag{7}$$

$$\text{Admitted\_to\_large}[P] \Leftarrow \text{Illness\_Severity}[P] \tag{8}$$

The rule (5) above encodes the belief that more severe illnesses cause higher bills, (6) encodes that privately owned hospitals bill patients differently from public hospitals, (7) holds that surgical admissions cause greater medical expenses than diagnostic admissions, and (8) indicates that a more severe class of illness causes admission to a large hospital rather than a small hospital.

The interface that CaRL provides for communicating this background knowledge is illustrated in Figure 2, along with the partial previews of the causal diagrams that the system provides. Once we perform the initial demo, we will allow the audience to experiment with their own causal rules.

Causal Queries. CaRL supports three types of causal queries: (1) Average Treatment Effect (ATE) queries which estimate the difference in outcome between all units receiving the treatment and none receiving it. For example the following computes the ATE of Prestige of authors on Score of papers across double-blind and single blind conferences (dictated by the GROUP BY operator), i.e., it compares papers' scores in two hypothetical worlds in which all authors are and are not affiliated with prestigious institutions.

$$\text{Score}[S] \Leftarrow \text{Prestige}[A]?\ \text{GROUP BY Blinding} \tag{9}$$

(2) Aggregated Response queries, which allow estimating the effect of a treatment on an aggregation of response variables. The following is an example of this type of query,

indicated by the use of the prefix AVG_, which compares the average hospital bill in the two hypothetical worlds in which all patients are admitted to all large hospitals vs all small hospitals (as measured by number of beds):

$$AVG\_Bill[H] \Leftarrow Admitted\_to\_Large[P]? \qquad (10)$$

(3) Isolated/Relational Effect queries, which allow decomposing treatment effects into those caused by one's own treatment and those caused by peers' treatments. This allows for studying network effects in the data, for example comparing the effects of the coauthors' prestige with the effect of an author's own prestige on a paper's acceptance. The following query of this type, indicated by the usage of WHEN LESS THAN to define the threshold used for peer treatments, performs two-way comparisons of the hypotheticals: no prestigious authors and no prestigious coauthors, all prestigious authors and no prestigious coauthors, no prestigious authors and all prestigious coauthors:

$$\begin{aligned}\text{Score}[S] \Leftarrow \text{Prestige}[A]?\ &\text{WHEN LESS THAN } 1/3\\ &\text{PEERS TREATED GROUP BY Blinding} \qquad (11)\end{aligned}$$

The interface for querying, as well as sample results, can be seen in Figure 3. For convenience and comparison, the interface also allows the user to execute simple SQL queries and visualize them.

As with the relational causal model, the audience can interact with CaRL and pose customized queries of their own choosing after the initial demonstration. For example, the audience may be interested in the effect of different treatments (e.g. public vs. private ownership) or different outcomes (e.g. mortality rate).

Query Answering. The query answering component of CaRL accepts as input a causal query, a relational database and a ground causal model. It then, 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. We refer the reader to [12] for more details.

Result interpretation. For query (9), CaRL estimates that in single-blind conferences, prestigious authors are 6% more likely to be accepted, while in double-blind conferences they are 0.1% more likely. The corresponding SQL query obtained by joining all tables finds the reduction in bias is only 11% vs 9%, a much smaller change. We note that experimental studies favor our result in finding a large debiasing effect due to double-blinding [13].

Query (11), which decomposes query (9) into the effect of an author's own prestige and his or her coauthors' prestige estimates that 5% is attributable to the author, while 1% is attributable to the coauthors. This is in accordance with intuition.

On Query (10), we find that admittance to a large hospital reduces costs by 11%. The corresponding SQL query run on the universal table obtained by joining all tables finds that patients who entered a large hospital had bills increase by 33%. We argue that our result is more accurate: large hospitals benefit from economies of scale (which reduce the actual cost of care) while also treating more severe cases (which increase their correlation with high cost care). In fact, insurance literature supports the existence of healthcare economies of scale and it is the policy of several national governments to consolidate small hospitals to increase care efficiency [4].

In both cases, CaRL identifies the correct trend while naive SQL queries that ignore the relationality of data lead to incorrect and perplexing insights.

4. CONCLUSION

We have outlined CaRL, a framework for causal inference in relational domains. Key components are the domain-specific language, which captures essential information about the mechanisms of the system under study and allows for expressive queries; an interpreter that creates a causal model, summarizes and embeds it such that it can be computed, and provides estimates of the average treatment effect, along with other causal quantities. We demonstrate CaRL on two real-life datasets from academia and medicine, demonstrating that it avoids the false discoveries made by existing methods and that its predictions are in line with experts' expectations.

Acknowledgements

This research was supported by NSF III 1703281 and NSF IIS 1907997.

5. REFERENCES

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

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

[3] Celi et al. Tan open benchmark for causal inference using the mimic-iii and philips datasets.

[4] 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.

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

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

[7] 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.

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

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

[10] 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.

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

[12] Babak Salimi, Moe Kayali, Harsh Parikh, Lise Getoor, Sudeepa Roy, and Dan Suciu. Causal relational learning. *ACM SIGMOD International Conference on Management of Data*, 2020.

[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, November 2017.