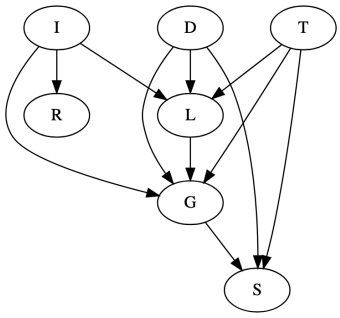
Relational Causal Models: A declarative framework for causal inference from relational data

Contents



(a)

Figure 1: (a) A causal DAG in a simple university domain S = student's satisfaction, G = student's grade, L = lecture attendance, I = student's intelligence, D = course difficulties and T = professor teaching skills (cf. Ex. 1.1);

1 INTRODUCTION

Non-parametric structural equations and their graphical representation, causal DAGs, developed by Pearl [9] are simple and powerful abstraction of the functional relationship between variables in a domain. However, they have a rigid structure and therefore incapable of modeling variable number of objects and relations between them in relational domains. We illustrate with an example.

Example 1.1. The causal DAG in Fig 1 represents a small fraction of a causal model in a university domain. The DAG postulates the following assumptions: student grades in a course are a function of their intelligence, lecture attendance, the professor's teaching skills and some latent exogenous factors such as their health condition; lecture attendance of students is a function of their intelligence, the professor's teaching skills and other exogenous variables; a student's satisfaction is a function of their intelligence, grade, their professor's teaching skills, and other exogenous factors. However, the causal DAG fails to capture confounding due to causal interaction of objects from related entities. For instance, to asses the effect of tutoring sessions on student performance, student's social connections' such as their teammates or friends affect both their treatment assignment and their outcomes. As shown in Fig 1(b), in this case, there is an inference between student, because the treatment of one students affects it peer's outcomes. Moreover, the outcome exhibits contagion, because related students are likely to study together. Expressing causal interaction between the related objects in a network using NSEs is either impossible or tedious.

In this paper, we introduce *relational causal models* that leads to a richer language and inference methods which can capture complex causal dependencies that are present in real-world scenarios. The goal is to provide the theoretical foundation and tools for answering the following types of causal queries in relational settings:

- What is the effect of tutoring sections on the average grade in a class? is the effect of having one highly intelligent student in a course on the average grade in class? What is the effect of the popularity of the instructor on the average grade in a class?
- what is the effect of Bob's lecture attendance in a class on this grade in that class?
- What is the effect of having team projects on the average grade in the class?

2 BACKGROUND IN CAUSAL INFERENCE

Causal Models and Structural Equations. A probabilistic causal model (PCM) is a tuple $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}, \mathtt{Pr}_{\mathbf{U}} \rangle$, where $\mathbf{U}$ is a set of background or exogenous variables that cannot be observed but which can influence the rest of the

model; $\mathbf{V}$ is a set of observable or endogenous variables; $\mathbf{F} = (F_X)_{X \in \mathbf{V}}$ is a set of *non-parametric structural equations (NSE)* $F_X : Dom(\mathbf{Pa_U}(X)) \times Dom(\mathbf{Pa_V}(X)) \rightarrow Dom(X)$, where $\mathbf{Pa_U}(X) \subseteq \mathbf{U}$ and $\mathbf{Pa_V}(X) \subseteq \mathbf{V} - \{X\}$ are called the exogenous parents and endogenous parents of X respectively; and $\mathtt{Pr_U}$ 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). [SR: why completely determined? should not it be probabilistic given the valUes of parents like BN?]

Causal DAG. To each PCM M we associate a causal graph G with nodes consisting of the endogenous variables $\mathbf{V}$, and edges consisting of all pairs (Z, X) such that $Z \in \mathbf{Pa_V}(X)$; we write $Z \rightarrow X$ for an edge. G is always assumed to be acyclic, and called Causal DAG. One can show that the probability distribution on the exogenous variables uniquely determined a distribution $\mathtt{Pr_V}$ on the endogenous variables and, under the *causal sufficiency* assumption[1], $\mathtt{Pr_V}$ forms a Bayesian network, whose graph is exactly G:

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

Thus justifies omitting the exogenous variables from the causal DAG, and capturing their effect through the probability distribution Eq.(1). We will only refer to endogenous variables, and drop the subscript $\mathbf{V}$ from $\mathbf{Pa_V}$ and $\mathtt{Pr_V}$. A path in G means an undirected path, i.e. we may traverse edges either forwards or backwards; a directed path is one where we traverse edges only forwards.

Rule-based representation: A NSEs F_X can be viewed as a proportional rule $\mathcal{R}(X)$ of the following form that represent a probabilistic mappings with conditional probability distribution $\mathtt{Pr}(X \mid \mathbf{Pa}(X))$:

$$X \Leftarrow \bigotimes_{V \in \mathbf{Pa}(X)} V \tag{2}$$

where, the $\otimes$ symbol splits the inputs of F_X. For example, the following rules correspond to the NSEs underling the causal DAG in Fig 1. Note that rules with empty body denote that the value of the variable in the head entirely determined by exogenous variables.

$$\mathrm{T} \Leftarrow \tag{3}$$
$$\mathrm{D} \Leftarrow \tag{4}$$
$$\mathrm{I} \Leftarrow \tag{5}$$
$$\mathrm{R} \Leftarrow \mathrm{I} \tag{6}$$
$$\mathrm{S} \Leftarrow \mathrm{T} \otimes \mathrm{G} \otimes \mathrm{D} \tag{7}$$
$$\mathrm{L} \Leftarrow \mathrm{I} \otimes \mathrm{T} \tag{8}$$
$$\mathrm{G} \Leftarrow \mathrm{I} \otimes \mathrm{D} \otimes \mathrm{T} \otimes \mathrm{L} \tag{9}$$

d-Separation. We review the notion of d-separation, which is the graph-theoretic characterization of conditional independence. A *path* $\mathbf{P}$ from X to Y is a sequence of nodes $X = V_1, \ldots, V_\ell = Y$ such that $V_i \rightarrow V_{i+1}$ or $V_i \leftarrow V_{i+1}$ forall i. $\mathbf{P}$ is *directed* if $V_i \rightarrow V_{i+1}$ forall i, and in that case we write $X \stackrel{*}{\rightarrow} Y$, and say that X is an *ancestor*, or a *cause* of Y, and Y is a *descendant* or an *effect* of X. If the path contains a subsequence $V_{k-1} \rightarrow V_k \leftarrow V_{k+1}$ then V_k is called a *collider*. A path with a collider is *closed*; otherwise it is *open*; an open path has the form $X \stackrel{*}{\leftarrow\!\!\rightarrow} Y$, i.e. X causes Y or Y causes X or they have a common cause. Given two sets of nodes $\mathbf{X}, \mathbf{Y}$ we say that a set $\mathbf{Z}$ *d-separates*[2] $\mathbf{X}$ and $\mathbf{Y}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$, if for any all paths P from $\mathbf{X}$ to $\mathbf{Y}$ one of the followings hold: (1) $\mathbf{P}$ is closed at a collider node V such that neither V nor any of its descendants are in $\mathbf{Z}$; (2) $\mathbf{P}$ contains a non-collider node V' such that $V' \in \mathbf{P}$. We say that a set $\mathbf{Z}$. d-Separation is a sufficient condition for conditional independence.

Counterfactuals and do Operator. A *counterfactual* is an intervention where we actively modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. Pearl [7] described the *do* operator that allows this effect to be computed on a causal DAG, denoted $\mathtt{Pr}(Y|do(X = x))$. To compute this value, we assume that X is determined by a constant function $X = x$ instead of a function provided by the causal DAG. This assumption corresponds to a modified graph with all edges into $\mathbf{X}$ removed, and values of these variables are set to $\mathbf{x}$.

do-Calculus. A set of sound and complete axioms known as do-calculus can be used to decide whether the effect of intervention can be identified from the observational data [7]. Specifically, it has been shown that the probability $\mathtt{Pr}(y|\mathtt{do}(\mathbf{X} = \mathbf{x}))$ with $\mathbf{X} = X_1 \ldots X_n$ is identifiable iff by repeatedly applying the following rules, one can obtain a probabilistic formula free from the do operator and in terms of observed probabilities.

- Ignoring observations:

$$\mathtt{Pr}(y|\mathtt{do}(x), z, w) = \mathtt{Pr}(y|\mathtt{do}(x), w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)G_{\overline{X}} \tag{10}$$

[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]d stands for "directional".

- Action/Observation exchange

$$\mathtt{Pr}(y|\mathtt{do}(x),\mathtt{do}(z),w) = \mathtt{Pr}(y|\mathtt{do}(x),z,w) \text{ if } (Y \perp\!\!\!\perp Z|X,W)G_{\overline{X},\underline{Z}} \tag{11}$$

- Ignoring actions/interventions

$$\mathtt{Pr}(y|\mathtt{do}(x),\mathtt{do}(z),w) = \mathtt{Pr}(y|\mathtt{do}(x),w) \text{ if } (Y \perp\!\!\!\perp Z|X,W)G_{\overline{X},\overline{Z(W)}} \tag{12}$$

where, $G_{\overline{X}}$ denotes the perturbed graph in which all edges pointing to X have been deleted; $G_{\underline{X}}$ denotes the perturbed graph in which all edges pointing from X have been deleted; $Z(W)$ denote the set of nodes in Z which are not ancestors of W.

3 PRELIMINARIES

We denote variables by uppercase letters, X, Y, Z, V; their values with lower case letters, x, y, z, v; and denote sets of variables or values using boldface ($\mathbf{X}$ or $\mathbf{x}$). A *first order atom* has the form $P(X_1, \ldots, X_k)$, where P is either a *function* or *predicate* symbol and the $X_1, \ldots, X_k$ is a sequences of *terms*, i.e., variables or constants. Each variable X is typed with a *domain* Δ_X and each atom P has a set of values called the *range* of $P(X)$denoted $Ran(P)$. A predicate is an atom with range {True, False}. [SR: to be simplified later.]

A *relational schema* is a tuple $\mathbf{S} = (\mathbf{E}, \mathbf{R}, \mathbf{A})$, where $\mathbf{E} = \{E_1, \ldots, E_n\}$ is a set of unary *entity predicates* representing entity classes in a domain; $\mathbf{R} = \{R_1, \ldots, R_m\}$ is set of binary *relationship predicates* that describe the relationship between pairs of entity classes; $\mathbf{A} = \{A_1, \ldots, A_k\}$ is a set of *functional symbols* that describe the *attributes* of entity classes or the relationship classes, A *relational structure* over an schema $\mathbf{S}$ is a tuple $\mathbf{\Delta} = (\Delta, \mathbf{E}^\Delta, \mathbf{R}^\Delta)$ consists of a a set of objects Δ together with an interpretation of each k-ary predicate symbol $P \in \mathbf{E} \cup \mathbf{R}$ as a k-ary relation on Δ, i.e., $P^\Delta \in \Delta^n$. Given a each relation $R(X, X') \in \mathbf{R}$ that represents a self relationship between objects of an entity class, we capture higher order relationship between the objects with predicate $R(X, X', k) \leftrightarrow \exists X_1, X_2, X_{k-1} \;\; R(X, X_1), \ldots, R(X_k, X')$ that represents objects that are in a $k \geq 1$ order relationship. $R(X, X', k)$ coincide with $R(X, X')$ for $k = 1$. For example, $\textsc{Friend}(S, S', 2)$ captures the second order friendship between students, i.e., the friends of a friend relationship. We capture with $R(X, X', k_{\leq}) \leftrightarrow \bigvee_{i=2}^{k} R(X, X', i)$ objects that are in a relationship of order less than or equal to k.

A relational structure $\mathbf{\Delta}$ can be represented with an edge-labeled undirected graph $G_\Delta = (\mathbf{V}, \mathbf{E})$, where the vertices are correspond to objects in $\mathbf{\Delta}$ and $(V, U) \in E$ with label R iff $R \in \mathbf{R}$ and $R(V, V') \in R^\Delta$. A *relational instance* $\mathbf{I}$ from $\mathbf{S}$ consists of a relational structure $\mathbf{\Delta}$ together with an interpretation of each k-ary function symbol $A \in \mathbf{A}$ as a k-ary function function $A^\Delta : \Delta^k \rightarrow Ran(A)$.

Example 3.1. The following schema $\mathbf{S}$ represents a simplified university domain consisting of entity classes students, courses, professors and teams. Students can register for several courses that are thought by professors. Classes can have several teams consist of multiple students. Students are related together by teammate and friends relations.

$\mathbf{E}$ = {Professor(X), Student(X), Course(X), Team(X)}

$\mathbf{R}$ = {Friend(X, Y), Teammates(X, Y), Has_Team(X, Y), Teaches(X, Y), Registered(X, Y), Member(X, Y)}

$\mathbf{A}$ = {Teaching_Skills(X), Popularity(X), Difficulty(X), Grade(X, Y), Lecture_attend(X), Satisfaction(X, Y), Intelligence(X), Ranking(X)}

Fig 2 shows an instance from $\mathbf{S}$ with an underling relational structure represented in Fig 3.

4 CAUSAL INFERENCE FROM RELATIONAL DATA

This sections proposes relational causal models, a relational extension of Pearl's causal models that can reason about causal interaction of objects in complex relational settings.

4.1 Relational Causal Model

This section introduces relational causal models. First, we introduce a first-order extension of non-parametric structural equations.

Definition 4.1. Given an schema $\mathbf{S} = (\mathbf{E}, \mathbf{R}, \mathbf{A})$ *and an atom* $P \in \mathbf{R} \cup \mathbf{A}$, *a* relational non-parametric structural equation (RNSE) $\mathfrak{R}(P(\mathbf{X}))$ *is a rule of the following form:*

$$P(\mathbf{X}) \Leftarrow \bigotimes_{i \in \mathbf{s}_p} \; \bigotimes_{\{\mathbf{X}_i \,:\, \varphi_i(\mathbf{Y})\}} P_i(\mathbf{X}_i)$$

where, $\mathbf{s}_p \in 2^{\{1..|\mathbf{P}|\}}$, $P_i \in \mathbf{A} \cup \mathbf{R}$ *and* $\varphi(\mathbf{Y})$ *is a conjunctive query possibly with negation over the atoms in* $\mathbf{A} \cup \mathbf{R}$ *that is safe w.r.t.* $\mathbf{X} \cup \mathbf{X}_i$.

RNSEs can be viewed as a first order extension of the rule-based representation of NSEs in Eq 2. Indeed, a RNSE $\mathfrak{R}(P(\mathbf{X}))$ together with a relational structure Δ defines a set of NSEs obtained by grounding $\mathfrak{R}(P(\mathbf{X}))$, as defined next.

Definition 4.2. Given an schema $\mathbf{S} = (\mathbf{E}, \mathbf{R}, \mathbf{A})$, *a relational structure* $\mathbf{\Delta}$ *and a RNSE* $\mathfrak{R}(P(\mathbf{X}))$, *a* grounding *of* $\mathfrak{R}(P(\mathbf{X}))$ *w.r.t.* $\mathbf{\Delta}$ *is a NSE* $\mathcal{R}(P(\mathbf{x}))$ *with* $P(\mathbf{x}) \in P^\Delta$ *that obtained from* $\mathfrak{R}(P(\mathbf{X}))$ *in following steps:*

- *replace any occurrences of* $\mathbf{X}$ *with* $\mathbf{x}$,
- *for each* $i \in \mathbf{s}_p$, *replace* $\bigotimes_{\{\mathbf{X}_i \,:\, \varphi_i(\mathbf{Y})\}} P_i(\mathbf{X}_i)$ *with* $P_i(\mathbf{x}_{i1}) \otimes, \ldots, \otimes P_i(\mathbf{x}_{ik})$ *such that for each* $\mathbf{x}_{ij}, j = 1, k, \Delta \models \varphi_i([\mathbf{Y}/\mathbf{x}_{ij}, \mathbf{x}])$, *where* $[\mathbf{Y}/\mathbf{x}_{ij}, \mathbf{x}]$ *substitutes any occurrences of* $\mathbf{X}$ *and* $\mathbf{X}_i$ *in* φ_i, *respectively with* $\mathbf{x}$ *and* $\mathbf{x}_{ij}$.

Professor(X)

X
Susan
David
Alon

Student(X)

X
Bob
Jack
Alex

Course(X)

X
cse344
cse403

Team(X)

X
#11101
#11820

Friend(X,Y)

X	Y
Bob	Alex

Teammate(X,Y)

X	Y
Bob	Jack
Bob	Alex

Has_Team(X,Y)

X	Y
cse344	#11101
cse403	#11820

Teaches(X,Y)

X	Y
Susan	cse344
David	cse344
Alon	cse403

Registered(X,Y)

X	Y
Bob	cse344
Bob	cse403
Jack	cse403
Alex	cse344

Member(X,Y)

X	Y
Bob	#11101
Jack	#11101
Bob	#11820
Alex	#11820

X	Diff(X)	Rating(X)	Tutor(X)
cse344	High	High	Yes
cse344	Low	Low	No

X	Intelligence(X)
Bob	Low
Jack	High
Alex	Low

X	Y	Grade(X,Y)	Lec_Att(X,Y)	Sat(X,Y)
Bob	cse344	A	Yes	Yes
Bob	cse403	B	No	No
Jack	cse403	B	Yes	Yes
Alex	cse344	A	No	Yes

X	Teach_Skills(X)	Popularity(X)
Susan	Low	High
David	High	High
Alon	Low	Low

Figure 2: An instance from the university schema in Ex. 3.1.

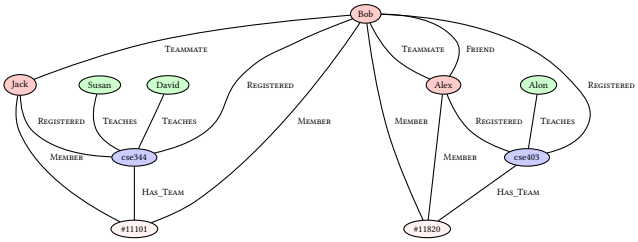


Figure 3: The graphical representation of the relational structure in Fig 2.

Let us denote the set of all groundings of a NRSEs $\mathfrak{R}(P(\mathbf{X}))$ with $\mathcal{G}(\mathfrak{R}(P(\mathbf{X})))$.

Example 4.1. Consider the following RNSEs concern with the university schema in Ex. 1.1:

$$\textsc{Intel}(S) \Leftarrow \tag{13}$$

$$\textsc{Diff}(C) \Leftarrow \tag{14}$$

$$\textsc{Skilled}(P) \Leftarrow \tag{15}$$

$$\textsc{Grade}(S,C) \Leftarrow \bigotimes_{\{S':\textsc{Teammates}(S,S',C)\}} \textsc{Lec\_Att}(S',C) \bigotimes_{\{S':\textsc{Friends}(S,S',k_{\leq})\wedge\textsc{Reg}(S',C)\}} \textsc{Lec\_Att}(S',C) \tag{16}$$

$$\bigotimes_{\{S':\textsc{Teammates}(S,S',C)\}} \textsc{Intel}(S') \bigotimes_{\{S':\textsc{Friends}(S,S',k_{\leq})\wedge\textsc{Reg}(S',C)\}} \textsc{Intel}(S') \tag{17}$$

$$\bigotimes_{\{P:\textsc{Teaches}(P,C)\}} \textsc{Skilled}(P) \otimes \textsc{Diff}(C) \otimes \textsc{Intel}(S) \otimes \textsc{Tutor}(C) \tag{18}$$

$$\textsc{Rating}(C) \Leftarrow \bigotimes_{\{P:\textsc{Teaches}(P,C)\}} \textsc{Skilled}(P) \bigotimes_{\{P:\textsc{Reg}(S,C)\}} \textsc{Intel}(S) \otimes \textsc{Tutor}(C) \tag{19}$$

$$\bigotimes_{\{S:\textsc{Reg}(S,C)\}} \textsc{Sat}(S,C) \otimes \textsc{Diff}(C) \tag{20}$$

$$\textsc{Lec\_Att}(S,C) \Leftarrow \bigotimes_{\{P:\textsc{Teaches}(P,C)\}} \textsc{Skilled}(P) \bigotimes_{\{S':\textsc{Teammates}(S,S',C)\}} \textsc{Lec\_Att}(S',C) \otimes \textsc{Tutor}(C) \tag{21}$$

$$\bigotimes_{\{S':\textsc{Friends}(S,S',k_{\leq})\wedge\textsc{Reg}(S',C)\}} \textsc{Lec\_Att}(S') \otimes \textsc{Diff}(C) \otimes \textsc{Intel}(S) \tag{22}$$

$$\textsc{Sat}(S,C) \Leftarrow \bigotimes_{\{P:\textsc{Teaches}(P,C)\}} \textsc{Skilled}(P) \otimes \textsc{Grade}(S,C) \otimes \textsc{Intel}(S) \tag{23}$$

$$\bigotimes_{\{S':\textsc{Friend}(S,S',k_{\leq})\wedge\textsc{Reg}(S',C)\}} \textsc{Sat}(S',C) \bigotimes_{\{S':\textsc{Teammates}(S',S,C)\}} \textsc{Sat}(S',C) \otimes \textsc{Tutor}(C) \tag{24}$$

$$\textsc{Tutor}(C) \Leftarrow \bigotimes_{\{P:\textsc{Teaches}(P,C)\}} \textsc{Skilled}(P) \bigotimes_{\{P:\textsc{Reg}(S,C)\}} \textsc{Intel}(S) \otimes \textsc{Diff}(C) \tag{25}$$

$$\textsc{Pop}(P) \Leftarrow \textsc{Skilled}(P) \bigotimes_{\{S:\textsc{Reg}(S,C)\wedge\textsc{Teaches}(P,C)\}} \textsc{Sat}(S,C) \bigotimes_{S,C:\textsc{Teaches}(P,C)} \textsc{Diff}(C) \tag{26}$$

Babak: [discuss choice of k] The RNSEs in Eq 13-15 assert that intelligence of students, the difficulty of courses and teaching skill of professors determined by latent exogenous factors. The RNSE in Eq47 asserts that grade of a student in a course is function of the teaching skills of the professors that teach the course, lecture attendance of the student's friends and teammates, the difficulty of the course, their intelligence of student and whether the course has tutoring sessions. The followings are ground instantiations of the RNSE in Eq 48-47 w.r.t the relational structure in Fig. 2:

$$\textsc{Rating}(\text{cse344}) \Leftarrow \textsc{Skilled}(\text{Susan}) \otimes \textsc{Skilled}(\text{Alex}) \otimes \textsc{Pop}(\text{Susan}) \otimes \tag{27}$$

$$\textsc{Tutor}(\text{cse344}) \otimes \textsc{Sat}(\text{Bob, Alex}) \otimes \textsc{Diff}(\text{cse344}) \otimes \textsc{Pop}(\text{Alex}) \otimes \textsc{Sat}(\text{Susan, cse344}) \tag{28}$$

$$\textsc{Rating}(\text{cse403}) \Leftarrow \textsc{Skilled}(\text{Alon}) \otimes \textsc{Pop}(\text{Alon}) \otimes \textsc{Sat}(\text{Bob, cse403}) \tag{29}$$

$$\otimes \textsc{Sat}(\text{Alex, cse403}) \otimes \textsc{Diff}(\text{cse403}) \otimes \textsc{Tutor}(\text{cse403}) \tag{30}$$

As oppose to the previous cases, the RNSE in Eq. **??** exhibits cyclic causal relationships that capture capture social influences and dynamics in relational settings. Consider for instance the following two groundings of Eq. **??**, in which the lecture attendance of Bob in cse344 is affected by and in turn affect the lecture attendance of Alec in cse344.

$$\textsc{Lec\_Att}(\text{Bob, cse344}) \Leftarrow \textsc{Skilled}(\text{Susan}) \otimes \textsc{Skilled}(\text{David}) \otimes \textsc{Tutor}(\text{cse344}) \tag{31}$$

$$\textsc{Lec\_Att}(\text{Jack, cse344}) \otimes \textsc{Difficulty}(\text{cse344}) \otimes \textsc{Intel}(\text{Bob}) \tag{32}$$

$$\textsc{Lec\_Att}(\text{Jack, cse344}) \Leftarrow \textsc{Skilled}(\text{Susan}) \otimes \textsc{Skilled}(\text{David}) \otimes \textsc{Tutor}(\text{cse344}) \tag{33}$$

$$\textsc{Lec\_Att}(\text{Jack, cse344}) \otimes \textsc{Difficulty}(\text{cse344}) \otimes \textsc{Intel}(\text{Jack}) \tag{34}$$

As shown in Ex 4.1, RNSEs are enable the ability to capture causal relationships in complex relational domains that involve, variable number of objects and cyclic causal effects. Now, we give our definition of relational causal models.

Definition 4.3. A probabilistic relational causal model (PRCM) *is a tuple* $\mathfrak{M} = \langle \mathbf{S}, \mathbf{\Delta}, \mathbf{U}, \mathfrak{R}(\mathbf{S}), \mathtt{Pr}_{\mathbf{U}|\mathbf{\Delta}} \rangle$*, where* $\mathbf{S} = (\mathbf{E}, \mathbf{R}, \mathbf{A})$ *is relational schema,* $\mathbf{\Delta}$ *is a relational structure over* $\mathbf{S}$*,* $\mathfrak{R}(\mathbf{S})$ *is a set of RNSEs associated to each* $A \in \mathbf{A}$*;* $\mathbf{U}$ *is a set of latent exogenous variables associated to* $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$*, the grounding of* $\mathfrak{R}(\mathbf{S})$ *w.r.t.* $\mathbf{\Delta}$*;* $\mathtt{Pr}_{\mathbf{U}|\mathbf{\Delta}}$ *is a conditional probability distribution on exogenous variables given the relational structure* $\mathbf{\Delta}$*.*

A RCM $\mathfrak{M}$ defines a set of ground NSEs $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$. Let us denote the set of all groundings of an atom $A(\mathbf{X})$ in $\mathfrak{R}(\mathbf{S})$ as $\mathcal{G}(A(\mathbf{X}))$. Denote $\mathcal{G}(\mathbf{A})$ the set of all ground atoms appear in $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$. $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$ can be represented with a *relational causal graph (RCG)* $G_{\mathfrak{M}} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ corresponds to $\mathcal{G}(A)$ and there is directed edge from V' to V if there is a NSE $\mathcal{R}(V) \in \mathcal{G}(\mathfrak{R}(\mathbf{S}))$ such that V' appear in the body of $\mathcal{R}(V)$. Fig 4 shows a RCG associated to a RCM formed by the RNSEs in Ex 4.1. Throughout this paper we assume the following *relational causal sufficiency*: for any $\mathbf{u}_i, \mathbf{u}_j \in \mathbf{U}$, $\mathbf{u}_j \perp\!\!\!\perp \mathbf{u}_i | \mathbf{\Delta}$.

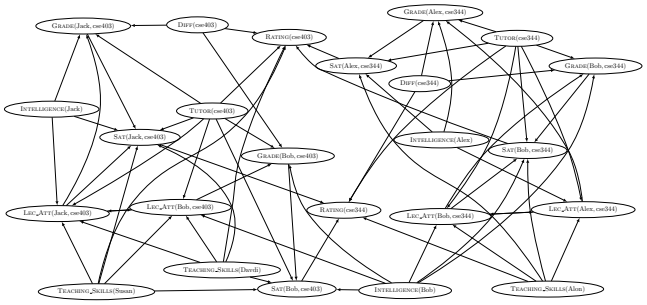


Figure 4: A dependency graph associated to the groundings relational causal model in Ex 4.1. Babak: [The graph needs to be updated]

The distribution $\mathtt{Pr}_{\mathbf{U}|\mathbf{\Delta}}$ over the exogenous variables induces a probability distribution $\mathtt{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta})$ over the ground variables $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$. Hence, a first order atom $A(\mathbf{X})$ in a RNSE can be viewed as a first order random variable that defines ground random variables $A(\mathbf{x}) \in \mathcal{G}(A(\mathbf{X}))$. Moreover, similar to conventional causal models (cf. Sec 8) the NSEs in $\mathcal{G}(\mathfrak{R}(\mathbf{S}))$ can be viewed as probabilistic mappings. While in principle it is possible to associated distinguished conditional probability distributions to these NSEs, with finite data and computational resources some assumptions are needed in order to estimate these distribution. In this paper, we make the following *homogeneity assumption*. Given a relational causal model $\mathfrak{M}$, we assume the underling generative process for all ground instantiations of a RNEM $\mathfrak{R}(P)$ are homogeneous For instance, in Ex 4.1, we assume the the underling process that determines the ranking of a course given the teaching skills of its instructors, the satisfaction of its students and etc. is the same for all courses. Under this assumption, the same conditional probability distributions can be assigned to each grounding of $\mathfrak{R}(P)$. However, to define valid conditional probability distributions that factorize the joint probability distribution $\mathtt{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta})$ the following issues must be addressed:

- **Cyclic causal dependencies:** The underling relational causal graphs associated to a relational causal model may include cycles, that capture dynamic process in relational domains that involve with feed back loop. [3] For instance, in Ex 4.1, lecture attendance of friends or teammates determined in a dynamic process in which students influences one another over a course of time. Cycles prohibit the factorization of the joint distribution as the product of the conditional probability distributions corresponds to the underling NSEs. For instance, the RCG shown

[3] Note that in causality literature cycles also used to represent associations due to missing confounders. To simplify the exposition we do not consider such cases.

Fig 5(a) defines two NSEs to which we associate $\mathtt{Pr}(T_1|T_2)$ and $\mathtt{Pr}(T_2|T_1)$. However, the joint probability distribution $\mathtt{Pr}(T_1, T_2)$ can not be factorized as the product of $\mathtt{Pr}(T_2|T_1)$ and $\mathtt{Pr}(T_1|T_2)$. Note such factorization is crucial for identifying the effect of interventions.

- **Variable sizes of ground instantiations:** Under the homogeneity assumption, a RNES $\mathfrak{R}(P)$ captures a probabilistic mapping that is generic to all its groundings. However, these groundings may have variable number of atoms in their bodies. For instance, in Ex 4.1, due to different number of instructors and students of a course, the groundings of the RNSE in Eq 48can have difference sizes. Therefore, to associate the same conditional probability distribution to the groundings of a RNSE $\mathfrak{R}(P)$, it needs to be normalize such that the grounding of the normalized $\mathfrak{R}(P)$ are of the same size.

Unfolding Relational Data with Embedding. To address the issue of variable sizes of the groundings of a RNSE we introduce relational embedding. The core idea of relational embedding is to unfold relational structures by projecting the feature vectors of objects from an entity class into a latent feature of objects from the related entity classes. For instance, in Ex 4.1, intelligence of students registered for a course can be project into a latent feature of course that represents "the propensity that a course is registered by intelligent students".

Babak: [How to justify the use of embeddings?]

Harsh: [We can use an example to justify why real world situations like learning/performance of students in a course is a function of professors teaching the course. Assume there are two professors teaching a course and have different contributions to course. One professor is teaching a hard concept and spends 0.8 time with them while other professor might teach an easy concept and spend only 0.2 time with them. In this case the students performance is highly determined by the first professor's attribute. This is some sort of weighted average of professors attributes.]

Given a relational schema $\mathbf{S} = (\mathbf{E}, \mathbf{R}, \mathbf{A})$, we extend the set of function symbols $\mathbf{A}$ with a set of *embedding function symbols* $\Phi = \{\phi_1, \ldots, \phi_m\}$, where each ϕ_i, $i = 0, m$ is defined by a RNSE $\mathfrak{R}(\phi_i)$ that is a deterministic function. For instance, consider the following RMRs concern with the university schema in Ex 4.1.

$$\phi_I(C) \Leftarrow \underset{\{S:\textsc{Reg}(S,C)\}}{\otimes} \textsc{Intel}(S) \tag{35}$$

$$\phi_S(C) \Leftarrow \underset{\{S:\textsc{Reg}(S,C)\}}{\otimes} \textsc{Sat}(S, C) \tag{36}$$

$$\phi_{SK}(C) \Leftarrow \underset{\{P:\textsc{Teaches}(P,C)\}}{\otimes} \textsc{Skilled}(P) \tag{37}$$

$$\phi_{\textsc{Fla}}(S, C) \Leftarrow \underset{\{S':\textsc{Friends}(S,S',\leq k)\wedge\textsc{Reg}(S',C)\}}{\otimes} \textsc{Lec\_Att}(S', C) \tag{38}$$

$$\phi_{TI}(S, C) \Leftarrow \underset{\{S':\textsc{Teammates}(S,S',C)\}}{\otimes} \textsc{Intel}(S') \tag{39}$$

$$\phi_{FI}(S, C) \Leftarrow \underset{\{S':\textsc{Friends}(S,S',\leq k)\wedge\textsc{Reg}(S',C)\}}{\otimes} \textsc{Intel}(S') \tag{40}$$

$$\phi_{FS}(S, C) \Leftarrow \underset{\{S':\textsc{Friends}(S,S')\wedge\textsc{Reg}(S',C)\}}{\otimes} \textsc{Sat}(S', C) \tag{41}$$

$$\phi_{TS}(S, C) \Leftarrow \underset{\{S':\textsc{Teammates}(S',S,C)\}}{\otimes} \textsc{Sat}(S', C) \tag{42}$$

$$\phi_D(P) \Leftarrow \underset{\{C:\textsc{Teaches}(P,C)\}}{\otimes} \textsc{Diff}(C) \tag{43}$$

$$\phi_{TS}(P) \Leftarrow \underset{\{C:\textsc{Teaches}(P,C)\}}{\otimes} \phi_S(C) \tag{44}$$

$$\tag{45}$$

The relational embedding functions $\phi_I(C)$ and $\phi_S(C)$ embed students' intelligence and satisfaction: $\phi_I(C)$ measures the propensity that a course is registered by intelligent students; $\phi_S(C)$ measures the propensity that students satisfied by the course. $\phi_{SK}(C)$ embeds professors teaching skill as a course feature. As oppose to the previous embedding that related the descriptive attributes of objects from different entity sets, the functions in Eq **??**-42 embed the descriptive attributes of objects from the same entity sets. Specifically, they measure the propensity that the friends or teammates of a student attend lectures or satisfies in a course. Finally, $\phi_{TS}(P)$ is a nested embedding which captures the propensity that that a professor teaches courses in which students are satisfied.

In this section we assume the embedding functions are given. In section Sec 5, we develop a set of techniques to learn them from data. Using the embedding functions in Eq 35-44 the NRSE in Ex. 4.1 can be reformulated as follows:

$$\textsc{Grade}(S, C) \Leftarrow \phi_{SK}(C) \otimes \phi_{TL}(S, C), \otimes\phi_{FL}(S, C) \otimes \phi_{TI}(S, C), \otimes\phi_{FI}(S, C) \otimes \textsc{Intel}(S) \tag{46}$$

$$\otimes\, \textsc{Tutor}(C) \otimes \textsc{Diff}(C) \tag{47}$$

$$\textsc{Rating}(C) \Leftarrow \phi_{SK}(C) \otimes \phi_I(C) \otimes \phi_S(C) \otimes \textsc{Diff}(C) \otimes \textsc{Tutor}(C) \otimes \textsc{Diff}(C) \tag{48}$$

$$\textsc{Lec\_Att}(S, C) \Leftarrow \phi_{SK}(C) \otimes \otimes\phi_{TL}(S, C), \otimes\phi_{FL}(S, C) \otimes \textsc{Diff}(C) \otimes \textsc{Intel}(S) \otimes \textsc{Tutor}(C) \tag{49}$$

$$\textsc{Sat}(S, C) \Leftarrow \phi_{SK}(C) \otimes \otimes\phi_{FS}(C) \otimes \phi_{TS}(C) \otimes \textsc{Grade}(S, C) \otimes \textsc{Intel}(S) \tag{50}$$

$$\textsc{Tutor}(C) \Leftarrow \phi_{SK}(C) \otimes \phi_I(C) \otimes \textsc{Diff}(C) \tag{51}$$

$$\textsc{Pop}(P) \Leftarrow \textsc{Skilled}(P) \otimes \phi_{TS}(P) \otimes \phi_D(P) \tag{52}$$

Now, it is easy to verify that the groundings of above RNSE have the same size. We say a RCM $\mathfrak{M}$ is *normalized* if all the groundings of a RNSE $\mathfrak{R} \in \mathfrak{R}(\mathbf{S})$ are of the same size. Any RCM can be normalized by introducing appropriate relational embedding functions.

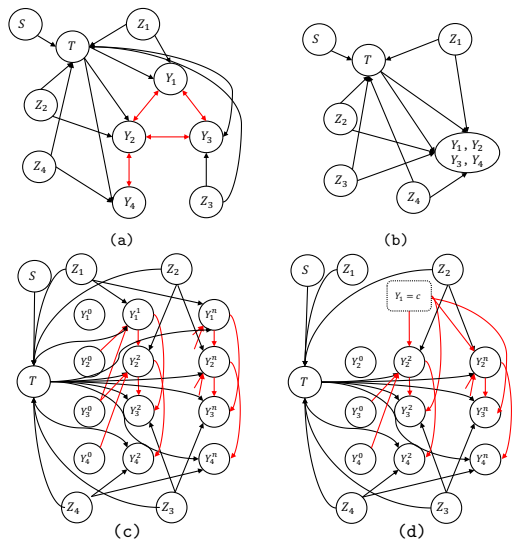


Figure 5: Unrolling cycles in a relational causal graph.

Unrolling Cycles: Cyclic causal relationship or feed-back loops arise naturally in relational domains where individuals with relational ties, such as employee and supervisor, or teammates and friends, exhibit strong behavioral contingency. For instance, the cycle between the lecture attendance of Bob and Jack in Fig 4 suggests that the decisions regarding attending lectures were made collectively in a dynamic process in which Bob and Jack influence each other over time. There have been a number of works modeling feedback relationships, e.g., [4, 8, 10–12]. In this paper, we use discrete Markov chains to model feedback loops, i.e., we assume the interaction between the objects, e.g., friends or teammates, occurs at discrete time steps, e.g., their social encounters.

Denote $SCC(G_{\mathfrak{M}}) = \mathbf{C}_1 \dots \mathbf{C}_k$ the *Strongly Connected Components(SCC)* of a RCG $G_{\mathfrak{M}}$. Recall that an SCC $\mathbf{C}$ of a directed graph is a maximal strongly connected subgraph, i.e., there is a path between all pairs of vertices in $\mathbf{C}$. Denote $\hat{G}_{\mathfrak{M}}$ the DAG of the SCCs of $G_{\mathfrak{M}}$, i.e., the DAG obtained by associating a node to each SCC $\mathbf{C}_i$, for $i = [1, k]$, and a set of edges $(\mathbf{C}_i, \mathbf{C}_j)$ if there is an edge in $G_{\mathfrak{M}}$ from any of the nodes in $\mathbf{C}_i$ to any of the nodes $\mathbf{C}_j$. For instance, Fig. 5(b) shows the DAG of the SCCs of the RCM in Fig. 5(a).

Given an SCC $\mathbf{C} = \{X_1, \dots, X_k\} \in SCC(G_{\mathfrak{M}}))$, suppose $X_1, \dots, X_k$ is a total order that presents the temporal dependence between X_is. Denote $\mathbf{Pa}_{\mathbf{C}}(X) \subseteq \mathbf{Pa}(X)$ a subset of the parents of X in $G_{\mathfrak{M}}$ such that $\mathbf{Pa}_{\mathbf{C}}(X) \subseteq \mathbf{C}$ and $\mathbf{Pa}_{\overline{\mathbf{C}}}(X) = \mathbf{Pa}(X) \setminus \mathbf{Pa}_{\mathbf{C}}(X)$. Furthermore, let $\mathbf{Pa}^{+}_{\mathbf{C}}(X)$ and $\mathbf{Pa}^{-}_{\mathbf{C}}(X)$ respectively, denote subsets of $\mathbf{Pa}_{\mathbf{C}}(X)$ that are greater than and less than X in the total order. We unroll $\mathbf{C}$ by assuming the underling generative processes starts with a random values for X_is and updates the values of X_i according to its associated COP given the current value of its parents. Let $\mathbf{C}^t = \{X_1^t, \dots, X_k^t\}$ denotes the values of X_is at time-step t. The process can be modeled using a causal DAG $G_{\mathbf{C}}$ with vertices $\mathbf{Pa}(\mathbf{C})$ and $\mathbf{C}^i$, for $i = 0, n$, and directed edges from variables in $\mathbf{Pa}^{-}_{\mathbf{C}}{}^{t+1}(X) \cup \mathbf{Pa}^{+}_{\mathbf{C}}{}^{t}(X) \cup \mathbf{Pa}_{\overline{\mathbf{C}}}(X)$ to X^{t+1} for $t = 0, n$. Fig. 5(c) shows the causal DAG obtained by unrolling $\{Y_1, Y_2, Y_3, Y_4\}$ in the RCM in Fig. 5(b).

It is easy to see that $G_{\mathbf{C}}$ encodes the following CIs:

$$X_i^{t+1} \perp\!\!\!\perp \mathbf{C}^{t+1}_{<i} \cup \mathbf{C}^t \cup \mathbf{Pa}(\mathbf{C}) \mid \mathbf{Pa}^{-}_{\mathbf{C}}{}^{t+1}(X) \cup \mathbf{Pa}^{+}_{\mathbf{C}}{}^{t}(X) \cup \mathbf{Pa}_{\overline{\mathbf{C}}}(X) \tag{53}$$

For instance, the following CIs encoded by the causal DAG in Fig. 5(c).

$$Y_1^{t+1} \perp\!\!\!\perp Z_2, Z_3, Z_4, \mathbf{Y}^t, Y_4^{t+1} \mid Y_2^t, Y_3^t, T, Z_1 \tag{54}$$

$$Y_2^{t+1} \perp\!\!\!\perp Z_1, Z_3, Z_4, \mathbf{Y}^t \mid Y_1^{t+1}, Y_3^t, T, Z_2 \tag{55}$$

$$Y_3^{t+1} \perp\!\!\!\perp Z_1, Z_2, Z_4, \mathbf{Y}^t, Y_4^{t+1} \mid Y_1^{t+1}, Y_2^{t+1}, T, Z_3 \tag{56}$$

$$Y_4^{t+1} \perp\!\!\!\perp Z_1, Z_2, Z_3, \mathbf{Y}^t, Y_1^{t+1}, Y_3^{t+1} \mid Y_2^{t+1}, T, Z_1 \tag{57}$$

Suppose $\mathbf{C}_1 \dots \mathbf{C}_k$ is in a topological order w.r.t. $\hat{G}_{\mathfrak{M}}$, it is easy to see that the DAG obtained by unrolling each $\mathbf{C}$ in $\hat{G}_{\mathfrak{M}}$ encodes the following CIs:

$$\mathbf{C}_i \perp\!\!\!\perp \mathbf{C}_j | \mathbf{Pa}(\mathbf{C}_i) \text{ for all } j < i \tag{58}$$

In order to use $G_{\mathbf{C}}$ for modeling and identifying external interventions, we need to have access to the detailed temporal data. In practice data is collected at a particular time-step, hence, $G_{\mathbf{C}}$ can not be used for inference. However, $G_{\mathbf{C}}$ forms a Markov chain that can be used to define a semantic for external interventions. Specifically, $G_{\mathbf{C}}$ defines a Markov chain $\{\mathbf{C}^n\}_{n>0}$ with state space $\mathbf{S} = Dom(\mathbf{C})$ and transition probabilities:

$$T(X_1^t, \dots, X_k^t \to X_1^{t+1}, \dots, X_k^{t+1}) = \prod_{i=1}^{k} \Pr(X_i \mid \mathbf{Pa}^{-}_{\mathbf{C}}{}^{t+1}(X) \cup \mathbf{Pa}^{+}_{\mathbf{C}}{}^{t}(X) \cup \mathbf{Pa}_{\overline{\mathbf{C}}}(X)) \tag{59}$$

For instance, the causal DAG in Fig. 5(c) define a Markov chain with the following transition probabilities:

$$T(Y_1^t, Y_2^t, Y_3^t, Y_4^t \to Y_1^{t+1}, Y_2^{t+1}, Y_3^{t+1}, Y_4^{t+1}) = \Pr(Y_1^{t+1}|Y_2^t, Y_3^t, Z_1, T)\Pr(Y_2^{t+1} \tag{60}$$

$$\Pr(Y_3^{t+1}|Y_1^{t+1}, Y_2^{t+1}, Z_3, T)\Pr(Y_4^{t+} \tag{61}$$

The following proved in Appendix.

Proposition 4.4. $\Pr(\mathbf{C}|\mathbf{Pa}(\mathbf{C}))$ *is the stationary distribution of the Markov chain* $\{\mathbf{C}^n\}_{n>0}$.

Given a variable $X \in \mathbf{C}$, the semantic of an external intervention $\mathsf{do}(X = x^*)$ can be defined in terms of a modified Markov chain $\{\mathbf{C}^n_{-X}\}_{n>0}$ with state space $\mathbf{S}' = Dom(\mathbf{C}_{-X})$ and transition probabilities obtained from Eq. 59 by further conditioning on $X = x^*$, whenever $X \in P(X_i)$. In practice, we are typically interested in the effect of an intervention in the limit distribution of $\{\mathbf{C}^n_{-X}\}_{n>0}$ as $t \rightarrow \infty$. It is easy to show that $\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C}) \setminus \mathbf{P}|X = x^*)$ is the unique stationary distribution of $\{\mathbf{C}^n_{-X}\}_{n>0}$, where $\mathbf{P} \subseteq \mathbf{Pa}(X)$ such $\mathbf{P} \cap \mathbf{Pa}(X') = \emptyset$, for all $X' \in \mathbf{C} \setminus \{X\}$. We show the following:

Proposition 4.5. *The modified Markov chain* $\{\mathbf{C}^n_{-X}\}_{n>0}$ *obtained from* $\{\mathbf{C}^n\}_{n>0}$ *by performing an external intervention* $\mathsf{do}(X = x^*)$ *has a unique stationary distribution* $\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C}) \setminus \mathbf{P}|X = x^*)$*, where* $\mathbf{P} \subseteq \mathbf{Pa}(X)$ *such* $\mathbf{P} \cap \mathbf{Pa}(X') = \emptyset$.

Notice that if the Markov chain observed in random time-steps, the transition probabilities can be estimated from observational data [2]. Given these transition probabilities the stationary distribution of $\{\mathbf{C}^n_{-X}\}_{n>0}$ can be computed using Markov Chain Monte Carlo (MCMC) sampling method. In MCMC methods, each X_i is sampled according to CPD associated to X_i according to a predefined sampling order. The samples generated after some burn-in period can be considered as random samples from the stationary distribution of $\{\mathbf{C}^n_{-X}\}_{n>0}$. However, it has been shown both theoretically and empirically that in the conditional probability distributions are inconsistent the the resulting equilibrium is a function of the sample ordering [4, 10].

However, if the observed data collected at equilibrium of $\{\mathbf{C}^n\}_{n>0}$, the parameters of the stationary distribution of $\{\mathbf{C}^n_{-X}\}_{n>0}$ can be directly estimated from observed data. First, we show that the stationary distribution of $\{\mathbf{C}^n\}_{n>0}$ has a parsimonious factorization that can be exploited for efficient estimation. The factorization implied from the fact that in the stationary distribution of $\{\mathbf{C}^n\}_{n>0}$ the following CIs hold:

$$X_i \perp\!\!\!\perp \mathbf{C}_{-i} \cup \mathbf{Pa}(\mathbf{C}) \setminus \mathbf{N}(X_i) \mid \mathbf{N}(X_i) \tag{62}$$

where, $\mathbf{N}(X_i)$ denotes the neighbors of X_i in moralized sub graph induced by $\mathbf{C} \cup \mathbf{Pa}(C)$, denoted $M(G_{\mathbf{C}\cup\mathbf{Pa}(C)})$. Notice that Eq 62 obtained immediately from Eq 53 and the assumption of stationary distribution. The CIs in Eq. 62 imply that $M(G_{\mathbf{C}\cup\mathbf{Pa}(C)})$ forms a Markov network for $\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C})|\Delta)$, hence it can be factorizes according to $CL(M(G_{\mathbf{C}\cup\mathbf{Pa}(C)}))$, the cliques of $\mathbf{M}(G_{\mathbf{C}})$, i.e,

$$\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C})) = \prod_{C \in CL(CL(M(G_{\mathbf{C}\cup\mathbf{Pa}(C)}))} \phi(C) \tag{63}$$

The generative process of the RCM $\mathfrak{M}$, given a set of CPDs $\mathsf{Pr}(X \in \mathbf{Pa}(X))$, for each $X \in \mathcal{G}(\mathfrak{R}(\mathbf{S}))$ and a topological order $\mathbf{C}_1, \ldots \mathbf{C}_k$ of the SCCs in $\hat{G}_{\mathfrak{M}}$, proceed as follows: the process iteratively runs the underling Markov chain of $\mathbf{C}_i$ until the unique stationary distribution $\mathsf{Pr}(\mathbf{C}|\mathbf{Pa}(\mathbf{C}), \Delta)$ is reached. Also note that from the CIs in Eq 58, it is implied that $\mathsf{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\Delta)$ admits the following factorization:

$$\mathsf{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\Delta) = \prod_{\mathbf{C} \in SCC(G_{\mathfrak{M}})} \mathsf{Pr}(\mathbf{C}|\mathbf{Pa}(\mathbf{C}), \Delta) \tag{64}$$

Now, the following equation implied from Prop. 4.4 and 4.5 relates the equilibrium distribution of a RCM before and after an external intervention $\mathsf{do}(X = x^*)$, where $X \in \mathbf{C}$:

$$\mathsf{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\Delta, \mathsf{do}(X = \mathbf{c})) = \mathsf{Pr}(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\Delta) \frac{\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C}) \setminus \mathbf{P}, X = x^*)}{\mathsf{Pr}(\mathbf{C} \mid \mathbf{Pa}(\mathbf{C}))} \tag{65}$$

If SCCs are singletons, Eq. 122 becomes similar to Eq ... for DAGs. Eq 122 can be naturally extended to capture interventions on a set of variables.

Babak: [Once we figure out how to estimate the above probabilistic formula we need to think about taking the average among all individual for which we should use embedding. In this example we need to embed the outcome of student's friends a feature of the student. Meaning that the embeddings are recursive (or should be defined in a recursive fashion).]

4.2 Causal Estimands and Covariate Adjustment

This section gives a sufficient conditions for identifying the effect of interventions in RCMs. We are given a RCM $\mathfrak{M}$ consists of a treatment atom $T(\mathbf{W})$ with the set of groundings $\mathcal{G}(T(\mathbf{W})) = \{T(\mathbf{w}_1) \ldots T(\mathbf{w}_m)\}$ and an outcome atom $Y(\mathbf{X})$ with a set of groundings $\mathcal{G}(Y(\mathbf{X})) = \{Y(\mathbf{x}_1) \ldots Y(\mathbf{x}_n)\}$. Given a vector of treatment assignments $\vec{t} = (t_1, \ldots, t_m)$, denote $\Pr(Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t}))$ the probability that $Y(\mathbf{x}) = y$ observed after an external intervention that sets $T(\mathbf{w}_i)$ to t_i, for $i = 1, m$. Our goal is to estimate the *average causal effect*(ACE) of a treatment $\vec{t}$ in compare to a treatment $\vec{t}'$ defined as follows:

$$\text{ACE} \stackrel{\text{def}}{=} \sum_{\mathbf{x}} \frac{1}{n} (\mathbb{E}[Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t})] - \mathbb{E}[Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t}')]) \tag{66}$$

Babak: [How can we capture other causal estimates in the literature, e.g., spill over effect, intervening on the network]

Next, we prove a graphical criterion for identifying $\Pr(Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t}))$ that can be checked against $\hat{G}_{\mathfrak{M}}$, the DAG of the SCCs of the RCG $G_{\mathfrak{M}}$ associated to $\mathfrak{M}$. In this paper we assume the treatment variables $\mathcal{G}(T(\mathbf{W})) = \{T(\mathbf{w}_1) \ldots T(\mathbf{w}_m)\}$ partitioned into k SCCs in $\hat{G}_{\mathfrak{M}}$, denoted $\mathbf{C}_1, \ldots, \mathbf{C}_k$. That is we assume a treatment variable can only have feed-back loop to the other treatment variables. Given a ground outcome variable $Y(\mathbf{x}) \in \mathcal{G}(Y(\mathbf{X}))$, denote $T_{\mathbf{x}} \subseteq \mathcal{G}(T(\mathbf{W}))$ such that for each $T(\mathbf{w}) \in T_{\mathbf{x}}$ there exists a directed path from $T(\mathbf{w})$ to

$Y(\mathbf{x})$ in $\hat{G}_{\mathfrak{M}}$. Clearly $Y(\mathbf{x})$ is causally independent any variable in $\mathcal{G}(T(\mathbf{W})) \setminus T_{\mathbf{x}}$. Furthermore, let $\vec{t}_{\mathbf{x}}$ be the vector of treatment assignment associated to $T_{\mathbf{x}}$.

Theorem 4.6 (Relational Adjustment). *Given $\hat{G}_{\mathfrak{M}}$, the DAG of the SCCs of a RCG $G_{\mathfrak{M}}$, the effect of an intervention* $\mathsf{do}(T(\mathbf{W}) = \vec{t})$ *that sets the ground variables $T(\mathbf{z}_i) \in \mathcal{G}(T(\mathbf{W}))$ to t_i on a ground variable $Y(\mathbf{x}) \in \mathcal{G}(Y(\mathbf{X}))$ is given by the following* relational adjustment formula*:*

$$\mathsf{Pr}(Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t}), \boldsymbol{\Delta}) = \sum_{\mathbf{c},\mathbf{z}} \mathsf{Pr}\big(Y(\mathbf{x}) = y \mid \mathbf{Z} = \mathbf{z}, T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \boldsymbol{\Delta}\big)\mathsf{Pr}\big(\mathbf{Z} = \mathbf{z}\big) \tag{67}$$

For any $\mathbf{Z} \subseteq \mathcal{G}(\mathfrak{R}(\mathbf{S})) \setminus \big\{\mathcal{G}(T(\mathbf{W})) \cup Y(\mathbf{x})\big\}$ *such that:*

$$Y(\mathbf{x}) \perp\!\!\!\perp \mathbf{Pa}(T_{\mathbf{x}}) \mid_{\hat{G}_{\mathfrak{M}}} T_{\mathbf{x}}, \mathbf{Z} \tag{68}$$

Notice that $\mathbf{Z} = \mathbf{Pa}(T_{\mathbf{x}}))$ always satisfies the independence assumption in Eq. 68, hence it is a sufficient set of covariates for adjustment. However, $\mathbf{Pa}(T_{\mathbf{x}})$ is not necessarily minimal. Identifying a minimal set of covariate for adjustment is important for building efficient estimators for Eq. 68. For example in the RCG in Fig 5(a), clearly Eq. 68 is satisfies by choosing $\mathbf{Z} = \{Z_1, Z_2, Z_3, Z_4\}$. By plugging $\mathbf{Z}$ into the relational adjustment formula in Eq 67, Eq. **??** is recovered.

The existing parametric and non-parametric density estimation can be used to estimate Eq 67. Indeed, any consistent estimators for $\mathsf{Pr}\big(Y(\mathbf{x}) = z \mid \mathbf{Z} = \mathbf{z}, T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \boldsymbol{\Delta}\big)$ and $\mathsf{Pr}\big(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\Delta}\big)$ becomes a consistent estimator for $\mathsf{Pr}(Y(\mathbf{x}) = y \mid \mathsf{do}(T(\mathbf{W}) = \vec{t}), \boldsymbol{\Delta})$. For continuous outcomes, assuming $Y(x) \in \mathbf{C}$, $\mathsf{Pr}\big(Y(\mathbf{x}) = z \mid \mathbf{Z} = \mathbf{z}, T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \boldsymbol{\Delta}\big)$ can be reformulated as $\sum_{\mathbf{C}\setminus\{Y(x)\}} \mathsf{Pr}\big(\mathbf{C} \mid \mathbf{pa}(\mathbf{C}), \boldsymbol{\Delta}\big)\mathsf{Pr}(\mathbf{pa}(\mathbf{C} \mid \mathbf{Z} = \mathbf{z}, T_{\mathbf{x}} = \vec{t}_{\mathbf{x}})$, where $\big(\mathbf{C} \mid \mathbf{pa}(\mathbf{C}), \boldsymbol{\Delta}\big)$ can be parametric according to the underling Markov Network.

However, notice that in principle an outcome variable can have feed-back loop to all outcome variables, which results in a saturated model for $\mathsf{Pr}\big(\mathbf{C} \mid \mathbf{pa}(\mathbf{C}), \boldsymbol{\Delta}\big)$. In addition, each outcome variable $Y(\mathbf{x})$ can be causally affected by all treated objects, i.e., $T_{\mathbf{x}} = \mathcal{G}(T(\mathbf{Z})$ which makes the summation in Eq 67 intractable. In practice, it is reasonable to assume $Y(\mathbf{x})$ only affected by treated objects and outcome of objects that are in some sort of *relational proximity* of $\mathbf{x}$. Therefore, we expect $|T_{\mathbf{x}}| \ll m$ and $\mathsf{Pr}\big(\mathbf{C} \mid \mathbf{pa}(\mathbf{C}), \boldsymbol{\Delta}\big)$ has a parsimonious factorization that make the estimation of Eq 67 feasible. For example, the grade of a student in a course is independent of whether other courses has tutoring sessions, hence, $\Pr(\textsc{Grade}(s, c) = g \mid \mathsf{do}(\textsc{Tutor}(C) = \vec{1})) = \Pr(\textsc{Grade}(s, c) = g \mid \mathsf{do}(\textsc{Tutor}(c) = 1))$, where

$$\mathbb{E}\left[\textsc{Grade}(s, c) \mid \mathsf{do}(\textsc{Tutor}(c) = 1)\right] = \sum_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3} \mathbb{E}\left[\textsc{Grade}(s, c) \mid \textsc{Tutor}(c) = 1, \phi_{SK}(c) = \mathbf{z}_1, \phi_I(c) = \mathbf{z}_2, \textsc{Diff}(c) = \mathbf{z}_3\right]\mathsf{Pr}\big(\phi_{SK}(c) = \mathbf{z}_1, \phi_I(c) = \mathbf{z}_2, \textsc{Diff}(c) = \mathbf{z}_3\big) \tag{69}$$

Now it is implied from the homogeneity assumption that the ACE of tutoring sessions on student grade is given by

$$ACE = \sum_{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3} \big(\mathbb{E}\left[\textsc{Grade}(s, c) \mid \textsc{Tutor}(c) = 1, \phi_{SK}(c) = \mathbf{z}_1, \phi_I(c) = \mathbf{z}_2, \textsc{Diff}(c) = \mathbf{z}_3\right]\mathbb{E}\left[\textsc{Grade}(s, c) \mid \textsc{Tutor}(c) = 0, \phi_{SK}(c) = \mathbf{z}_1, \phi_I(c) = \mathbf{z}_2, \textsc{Diff}(c) = \mathbf{z}_3\right] \mathsf{Pr}\big(\phi_{SK}(c) = \mathbf{z}_1, \phi_I(c) = \mathbf{z}_2, \textsc{Diff}(c) = \mathbf{z}_3\big) \tag{70}$$

Note that Eq 70 can be seen as comparing the expected grade of students in classes that are similar terms of the teaching skill of the instructors, intelligence of students and difficulty but different in terms of having tutoring sessions.

Estimating ACE using Eq. 67 in settings where $T_{\mathbf{x}}$ has variable sizes for different outcome variable $Y(x)$ becomes challenging. To see this suppose we are interested in the effect of teaching skill on students satisfaction. In this case $\textsc{Sat}(s, c)$ depends only on teaching skills of the instructors of the course c and has no parent in $\hat{G}_{\mathfrak{M}}$, hence

$$\Pr(\textsc{SAT}(s, c) = g \mid \mathsf{do}(\textsc{Skill}(P) = \vec{1})) = \Pr(\textsc{SAT}(s, c) = g \mid \textsc{Skill}(P_1) = 1, \ldots, \textsc{Skill}(P_m = 1)) \tag{71}$$

where, $P_1, \ldots P_m$ are the instructors of c. Now, since the number of instructors varies across different classes, estimating ATE using Eg. 71 is challenging. An immediate solution is to partition the outcomes to groups that are homogeneous on $|T_{\mathbf{x}}|$, computer ATE inside each group and take the expected value. This approach is effective when partitioning creates few number of groups. An alternative is to adjust for on some function of the treatment vector $f(\vec{t})$. However, this condition does not uniquely identifies interventions in general. In this paper, we focus on functions and treatment vectors that uniquely determine an intervention, i.e, the treatment vectors $\vec{1}$ and $\vec{0}$ and conditions $\textsc{mean}(\vec{t}) = 1$ or $\textsc{mean}(\vec{t}) = 0$ define unique interventions in which all objects receive and not receive the treatment. Furthermore, to answer queries such as what would be the expected grade of an student in a course if either 20%, exactly k, etc. receives a treatments, we assume all interventions represented by the condition that satisfy these effects has the same effect on the outcome. Under this assumption Eq. 72 can be reformulate as

$$\Pr(\textsc{SAT}(s, c) = g \mid \mathsf{do}(\textsc{Skill}(P) = \vec{1})) = \Pr(\textsc{SAT}(s, c) = g \mid f\big(\textsc{Skill}(P_1), \ldots, \textsc{Skill}(P_m)\big) = f(\vec{1})) \tag{72}$$

Now, the ACE of teaching skill on students grade can be computed using Eq. 72 interpreted as comparing the average of grade of students in courses in which all instructors have and do not have teaching skill. For a more complex example suppose we are interested in the effect of tutoring session on students GPA. It hods:

$$\mathbb{E}[\text{GPA}(s) \mid \mathsf{do}(\textsc{Tutor}(c) = 1)] = \sum_{\mathbf{z}_{1i},\mathbf{z}_{2i},\mathbf{z}_{3i}} \mathbb{E}[\text{GPA}(s) \mid \textsc{Tutor}(c_i) = t_i, \phi_{SK}(c_i) = \mathbf{z}_{11}, \phi_I(c_i) = \mathbf{z}_{2i}, \textsc{Diff}(c_i) = \mathbf{z}_{3i}]\mathsf{Pr}(\textsc{Tutor}(c_i) = 1, \phi_{SK}(z_i) = \mathbf{z}_{11}, \phi_I(c_i) = \mathbf{z}_{2i}, \textsc{Diff}(z_i) = \mathbf{z}_{3i}) \quad \text{for } i = 1, k \tag{73}$$

where, k is the number of course registered by the student s. Notice that, as oppose to Eq 69 that adjusts for the covariates of a course, Eq 78 adjusts for the joint distribution of covariates of the courses registered by a student. Now,

since each student may take different number of students then estimating ACE becomes challenging.

To circumvent the issue we use embedding. Specifically, we project the joint distribution of the covariates $C(\mathbf{z}_i)$ for $i = 1, m$ into a latent space of fixed size that represent a latent feature of $\mathbf{x}_i$. For instance, consider the following embeddings:

$$\phi_{RD}(S) \Leftarrow \bigotimes_{\{C:\text{Reg}(S,C)\}} \text{Diff}(C) \tag{74}$$

$$\phi_{RTS}(C) \Leftarrow \bigotimes_{\{C:\text{Reg}(S,C)\}} \phi_{TS}(C) \tag{75}$$

$$\phi_{RI}(S) \Leftarrow \bigotimes_{\{C:\text{Reg}(S,C)\}} \phi_{I}(C) \tag{76}$$

$$\tag{77}$$

These functions embed joint distribution of course features registered by a student into a student attribute. Using the embeddings, one can reformulate the probabilistic formula in Eq 78.

$$\mathbb{E}\left[\text{GPA}(s) \mid \text{do}(\text{Tutor}(c) = 1)\right] = \text{GPA}_s(\vec{t}, \Delta) = \sum_{z'_1, z'_2, z'_3} \mathbb{E}[\text{GPA}(s) \mid f(\text{Tutor}(c_1), \ldots, \text{Tutor}(c_{|\mathcal{K}_S|})) = y(\vec{t}),$$
$$\phi_{RD}(c) = \mathbf{z}'_1, \phi_{RTC}(c) = \mathbf{z}'_2, \phi_{RI}(c) = \mathbf{z}'_3]\Pr(\phi_{RD}(c) = \mathbf{z}'_1, \phi_{RTS}(c) = \mathbf{z}'_2, \phi_{RI}(c) = \mathbf{z}'_3) \tag{78}$$

Now the ACE can be computed using 78 and interpreted as comparing the expected GPA of students that are similar wrt. the embedded features, i.e., the tend to register for courses with similar covariate features but different wrt. having tutoring sessions.

Intervening on Relational Structure: ...

5 LEARNING RELATIONAL EMBEDDING

This section formally defines the problem of learning relational embeddings from data. We are given a relational instance I from a relational schema $\mathbf{S}$ that is generated with the RCM $\mathfrak{M} = \langle \mathbf{S}, \mathbf{\Delta}, \mathbf{U}, \mathfrak{R}(\mathbf{S}), \Pr_{\mathbf{U}|\Delta} \rangle$. Suppose $\mathfrak{M}$ consists of a set of latent embedding functions Φ. Our goal is to learn the latent embedding functions that are relevant to an outcome of interest $Y(\mathbf{X})$ from the instance I.

Let $\Phi^Y \subseteq \Phi$ be a set of embeddings such that for each $\phi(\mathbf{X}_i) \in \Phi^Y$ there exists a $\phi(\mathbf{x}_i) \in \mathcal{G}(\phi(\mathbf{X}_i))$ and a $Y(\mathbf{x}) \in \mathcal{G}(Y(\mathbf{x}))$ such that $\phi(\mathbf{x}_i)$ has a directed path to $Y(\mathbf{x})$ in the underling RCG $G_{\mathfrak{M}}$. Φ^Y consists of embeddings that are indeed relevant to $Y(\mathbf{X})$. Let $\Phi^Y_d = \phi_1(\mathbf{X}_1) \ldots \phi_k(\mathbf{X}_k)$ be a subset of Φ^Y such that: $\mathbf{x}_i \subseteq \mathbf{x}$ and the directed path from $\phi_1(\mathbf{x}_i)$ to $Y(\mathbf{x})$ is not intermediated by any embedding function. For a given set of embedding functions Φ^Y, let D be a table with attributes $Y, A_1, \ldots A_k$ that consists of tuples $(Y(\mathbf{x}), \phi_1(\mathbf{x}_i), \ldots \phi_1(\mathbf{x}_k))$, where $\mathbf{x}_i \subseteq \mathbf{x}$, for all $Y(\mathbf{x}) \in \mathcal{G}(Y(\mathbf{x}))$.

Suppose $f(.)$ is an ML model trained on D and attributes $A_1, \ldots A_k$ to predict Y, i.e, $f(A_1, \ldots A_n) = \hat{Y}$. Denote $\mathbf{L}(Y, \hat{Y}|\Phi^Y, f)$ a loss function defined over an ML model f and a given set of embeddings Φ^Y. The problem of learning relational embeddings is the following optimization problem:

$$\underset{\Phi^Y, \mathrm{C}}{\text{argmin}}\ \mathcal{L}(Y, \hat{Y}|\Phi^Y, f) \tag{79}$$

The above optimization problem is NP-complete. In the subsequent, we propose several heuristics that restrict the search space to a particular set of functions. Without loss of generality we restrict to two stage embedding functions.

Predefined Aggregates.

Hierarchical Linear Regression.

Kernel Estimation.

Neural Networks.

6 OPTIMIZATIONS

7 EXPERIMENTS

We consider the university set-up we have been discussing so far. Here, there are three entities courses, students and teachers. Multiple students enroll for multiple courses and multiple teachers can co-teach a course, however, a teacher cannot teach two courses in one semester. A course has multiple attributes including courses level, course department, number of students enrolled etc. Each student with either associated with some of the courses by the relation 'enrolled' or have not enrolled for any courses. Each students have multiple attributes too including age, gender, cumulative grade point average, ethinicity, etc. Similarly we consider the attributes of the teachers too like age, gender, students' rating, number of courses taught, department affiliated, etc. Courses and teachers are related to each other by the relation 'teaches'. We treat courses with the tutoring and are interesting in understanding if tutoring helps students' performance for e.g. GPA. Thus given the pretreatment setup we are interested in understanding what is the effect of tutoring option in a course on students' performance. We elaborate this set-up and it's mathematical formulation in the next subsection. We try to answer the following questions using this synthetic data experiments setup to answer the following questions:

- How effective is HumeL in predicting the effect of external interventions ?
- Can the summarizing methodology and embedding in latent space recover the actual mappings in the true data generative process?

7.1 Synthetic Dataset

Experiment Setup: We generated synthetic relational data according to a university schema introduced above . To simulate the effect of tutoring sessions on student grades, we generated a relational structure consists of $N_1 = 10\ 000$

students S, $N_2 = 700$ courses C and $N_3 = 500$ professors P using the following data generating process.

We consider one time step process, pre-treatment time-point 0 and post-treatment time-point 1, where the tutoring is if a course had tutoring option for a given semester. We are interested in estimating the causal effect of tutoring on student's grade point average. For each student s in the set of students $\mathcal{S}$, we generate pre-treatment attributes using following model:

$$\text{GPA(s)} \sim Normal(\mu = 3, \sigma^2 = 1, min = 0, max = 4)$$

Similarly, for any teacher p $\in \mathcal{P}$ the pre-treatment covariates are generated using the following process:

$$\textsc{Rating}(p) \sim Normal(\mu = 4, \sigma^2 = 1, min = 0, max = 5)$$

Finally, the features for any course $c \in C$ were generated as follows:

$$\textsc{Level}(c) \sim Uniform[1, 5]$$

$$\textsc{Department}(c) \sim Multinomial(p_{eng} = \frac{1}{2}, p_{sci} = \frac{1}{3},$$

$$\text{p}_{socsci} = \tfrac{1}{12}, p_{art} = \tfrac{1}{12})$$

The relations $\textsc{Registered}(\mathcal{S}, C)$ and $\textsc{Teaches}(\mathcal{P}, C)$ were generated using two random bipartite graphs and $\textsc{Friendship}(\mathcal{S}, \mathcal{S})$ were generated using a Erdös-Rényi random graph ER(N_1, 1/20).

To simulate the treatment $\textsc{Tutoring}(C)$, we sampled from a binomial distribution with success probability that is a logistic function of the embedded attributes $\phi_I(c)$ (cf. Eq. 35) and $\phi_{SK}(c)$ (cf. Eq. 37), the difficulty of course and some exogenous factors.

$$\begin{aligned}\textsc{Tutoring}(c) \sim \textsc{Binomial}(&\textsc{Logistic}(0.1\textsc{size}(\text{c}) - 2\textsc{Level}(\text{c}) \\ &+ \mathbf{1}_{[\textsc{Department}(c)=eng \text{ or } sci]} - \mathbf{1}_{[\textsc{Department}(c)=art]} \\ &- 2))\end{aligned}$$

$$\textsc{Attendance}(s, c) \sim \text{exponential}(\lambda = \sum_{s' \in \textsc{Friends}(s, s')} \textsc{Attendance}(s', c) + 10, max = 100)$$

$$\begin{aligned}\textsc{Grade}(\text{s,c}) \sim \textsc{Round}(4 \times \textsc{Logistic}(&10\text{GPA}(s) \\ &+ 20 \sum_p \mathbf{1}_{[\textsc{Teaches}(p,c)=1]} \textsc{Rating}(p) \\ &+ \textsc{Attendance}(\text{s,c}) - 5\textsc{Level}(\text{c}) + 10\textsc{Tutoring}(\text{c}) - 10))\end{aligned}$$

Given the above setups, we assume that each course is of one credit. Thus the grade point average will be a simple average of their grades for all the courses they have taken.

Experiment Results:

7.2 Real Dataset

Moe: [A few options I've come across: MovieLens, Slashdot Zoo, Ames Mutagenicity Benchmark]

8 RELATED WORKS

The importance of social network effects was brought to the foreground by Christakis and Fowler, who demonstrated strong impacts on the spread of obesity (2007), smoking (2008), alcohol consumption (2010). However, this work received strong criticism for using statistical techniques that assumed no interference between units (Lyons 2011). Early work on this topic attempted to address this issue by using graph clustering techniques on the social network and then applying treatments on the cluster level rather than the individual level [13].

One attempt to extend Pearl's causality model to the relational domain is due to Arbour et al. [1]. Another is due to Ogburn et al. [6].

There is also something similar to cycle unrolling against time steps in [6] (pages 6, 7).

Mention summarization has been used in the ML community but not causality community. Mention something about Sofrygin. Lam et al. approach the embedding problem in relational databases by using recurrent neural networks [3]. However, as they note, such networks are not able to handle unordered input data.

REFERENCES

[1] David Arbour, Dan Garant, and David Jensen. Inferring network effects from observational data. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 715–724, New York, NY, USA, 2016. ACM.

[2] Flavia Barsotti, Yohann De Castro, Thibault Espinasse, and Paul Rochet. Estimating the transition matrix of a markov chain observed at random times. *Statistics & Probability Letters*, 94:98–105, 2014.

[3] Hoang Thanh Lam, Tran Ngoc Minh, Mathieu Sinn, Beat Buesser, and Martin Wistuba. Learning features for relational data. *CoRR*, abs/1801.05372, 2018.

[4] Steffen L Lauritzen and Thomas S Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002.

[5] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

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

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

[8] Judea Pearl and Rina Dechter. Identifying independencies in causal graphs with feedback. *arXiv preprint arXiv:1302.3595*, 2013.

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

[10] David Poole and Mark Crowley. Cyclic causal models with discrete variables: Markov chain equilibrium semantics and sample ordering. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1060–1068, 2013.

[11] Mark Schmidt and Kevin Murphy. Modeling discrete interventional data using directed cyclic graphical models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 487–495. AUAI Press, 2009.

[12] Robert H Strotz and Herman OA Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, pages 417–427, 1960.

[13] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 329–337, New York, NY, USA, 2013. ACM.

9 APPENDIX

Proof. Denote $\pi^t(\mathbf{x}_1, \dots x_k)$ be the probability of being in state $x_1, \dots x_k$ after t transitions. It holds that:

$$\pi^{t+1}(x_1, \dots x_k) = \sum_{x'_1, \dots x'_k} \pi^t(x'_1, \dots x'_k) T(x'_1, \dots, x'_k \to x_1, \dots, x_k) \tag{80}$$

A distribution π is stationary if $\pi^i = \pi^{i+1}$, i.e.,

$$\pi(x_1, \dots x_k) = \sum_{x'_1, \dots x'_k} \pi(x'_1, \dots x'_k) T(x'_1, \dots, x'_k \to x_1, \dots, x_k) \tag{81}$$

We need to show the following, which is implied from the CIs in Eq 58:

$$\Pr(x_1, \dots x_k | \mathbf{Pa(C)}) = \sum_{x'_1, \dots x'_k} \Pr(x'_1, \dots x'_k | \mathbf{Pa(C)}) T(X_1^t, \dots, X_k^t \to X_1^{t+1}, \dots \tag{82}$$

$$RHS = \sum_{x'_1, \dots x'_k} \Pr(x'_1, \dots x'_k | \mathbf{Pa(C)}) \prod_{i=1}^{k} \Pr(X_i \mid \mathbf{Pa}_{\mathbf{C}}^{-t+1}(X) \cup \mathbf{Pa}_{\mathbf{C}}^{+t}(X) \cup \mathbf{Pa}_{\bar{\mathbf{C}}} \tag{83}$$

$$= \sum_{x'_1, \dots x'_k} \Pr(x'_1, \dots x'_k | \mathbf{Pa(C)}) \prod_{i=1}^{k} \Pr(x_i \mid \mathbf{x}_{<i} \cup \mathbf{x'}_{>i}, \mathbf{Pa(C)}) \tag{84}$$

$$= \sum_{x'_1, \dots x'_k} \Pr(x'_1, \dots x'_k | \mathbf{Pa(C)}) \Pr(x_1 \mid x'_2, \dots, x'_k, \mathbf{Pa(C)}), \dots, \Pr(x_k \mid x \tag{85}$$

$$= \sum_{x'_2, \dots x'_k} \sum_{x'_1} \Pr(x'_1, \dots x'_k | \mathbf{Pa(C)}) \Pr(x_1 \mid x'_2, \dots, x'_k, \mathbf{Pa(C)}), \dots, \Pr(x_ \tag{86}$$

$$= \sum_{x'_2, \dots x'_k} \Pr(x'_2, \dots x'_k | \mathbf{Pa(C)}) \Pr(x_1 \mid x'_2, \dots, x'_k, \mathbf{Pa(C)}), \dots, \Pr(x_k \mid x \tag{87}$$

$$= \sum_{x'_2, \dots x'_k} \Pr(x_1, x'_2, \dots, x'_k, \mathbf{Pa(C)}), \Pr(x_2 \mid x_1, x'_3, \dots, x'_k, \mathbf{Pa(C)}), \dots, \tag{88}$$

$$= \sum_{x'_3, \dots x'_k} \Pr(x_1, x_2, x'_3 \dots, x'_k, \mathbf{Pa(C)}) \Pr(x_3 \mid x_1, x_2, x'_4, \dots, x'_k, \mathbf{Pa(C)}), \tag{89}$$

$$\vdots \tag{90}$$

$$= \Pr(x_1, x_2, x_4, \dots, x_k \mid \mathbf{Pa(C)}) \tag{91}$$

$$= \Pr(\mathbf{C}, \mathbf{Pa(C)}) \tag{92}$$

$$= LHS \tag{93}$$

For example, for the Markov Chain in Fig. 5(c), we need to show:

$$\Pr(y_1, y_2, y_3, y_4 | z_0, z_1, z_2, z_3, t) = \sum_{y'_1, \ldots, y'_4} \Pr(y'_1, y'_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{94}$$

$$\Pr(y_1 | y'_2, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{95}$$

$$\Pr(y_2 | y_1, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{96}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{97}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{98}$$

$$RHS = \sum_{y'_2, \ldots, y'_4} \sum_{y'_1} \Pr(y'_1, y'_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{99}$$

$$\Pr(y_1 | y'_2, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{100}$$

$$\Pr(y_2 | y_1, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{101}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{102}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{103}$$

$$= \sum_{y'_2, \ldots, y'_4} \Pr(y'_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{104}$$

$$\Pr(y_1 | y'_2, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{105}$$

$$\Pr(y_2 | y_1, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{106}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{107}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{108}$$

$$= \sum_{y'_2, \ldots, y'_4} \Pr(y_1, y'_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{109}$$

$$\Pr(y_2 | y_1, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{110}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{111}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{112}$$

$$= \sum_{y'_3, \ldots, y'_4} \sum_{y'_2} \Pr(y_1, y'_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{113}$$

$$\Pr(y_2 | y_1, y'_3, y'_4, z_0, z_1, z_2, z_3, t) \tag{114}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{115}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{116}$$

$$= \sum_{y'_3, \ldots, y'_4} \Pr(y_1, y_2, y'_3, y'_4 | z_0, z_1, z_2, z_3, t) \tag{117}$$

$$\Pr(y_3 | y_2, y_1, y'_4, z_0, z_1, z_2, z_3, t) \tag{118}$$

$$\Pr(y_4 | y_1, y_2, y_3, z_0, z_1, z_2, z_3, t) \tag{119}$$

$$\vdots \tag{120}$$

$$= \Pr(y_1, y_2, y_3, y_4 | z_0, z_1, z_2, z_3, t) \tag{121}$$

The uniqueness of the stationary distribution implied from the fact that the Markov Chain is irreducible and aperiodic, which in turn implies from the assumption of strict positively of the distribution. It is known that irreducible and aperiodic Markov chains has a unique stationary distribution [5].

□

Proof. It is clear that

$$\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta}, \mathrm{do}(T(\mathbf{W}) = \vec{t}), \mathbf{\Delta}) = \Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathrm{do}(T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}), \mathbf{\Delta})$$

Since the intervention $\mathrm{do}(T(\mathbf{W}) = \vec{t})$ manipulates all elements of an SCC $\mathbf{C}_i$, for $i = 1, m$, by generalizing Eq 122 we obtain:

$$\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathrm{do}(T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \mathbf{\Delta}) = \frac{\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta})}{\prod_{i=1}^{k} \Pr(\mathbf{C}_i \mid \mathbf{Pa}(\mathbf{C}_i))} \tag{122}$$

The following holds according to the chain rule of probability:

$$\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta}) = \Pr(\mathbf{z})\Pr(\mathbf{pa}(T_{\mathbf{x}}) \mid \mathbf{z}) \prod_{i=1}^{k} \left(\Pr(\mathbf{C}_i \mid \bigcup_{j=1}^{i-1} \mathbf{C}_j), \mathbf{pa}(T_{\mathbf{x}}), \mathbf{z} \right) \Pr(\mathbf{s} \tag{123}$$

where, $\mathbf{S} = \mathcal{G}(\mathfrak{R}(\mathbf{S})) \setminus \{\mathbf{Z} \cup \mathbf{Pa}(T_{\mathbf{x}}), T_{\mathbf{x}}\}$.

The following immediately implied from Eq. 124, the independence assumption in Eq. 68 and Markov compatibility of $\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathbf{\Delta})$ with $\hat{G}_{\mathfrak{M}}$ and the fact that $\mathbf{pa}(T_{\mathbf{x}})$ is disjoint from $T_{\mathbf{x}}$:

$$\Pr(\mathcal{G}(\mathfrak{R}(\mathbf{S}))|\mathrm{do}(T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}), \mathbf{\Delta}) = \Pr(\mathbf{z})\Pr(\mathbf{pa}(T_{\mathbf{x}}) \mid \mathbf{z})\Pr(\mathbf{s}|T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \mathbf{z}, \mathbf{\Delta}) \tag{124}$$

By marginalizing over all variables except for $Y(\mathbf{x})$ we obtain:

$$\Pr(Y(\mathbf{x})|\mathrm{do}(T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}), \mathbf{\Delta}) = \sum_{s', z, \mathbf{pa}(T_{\mathbf{x}})} \Pr(\mathbf{z})\Pr(\mathbf{Pa}(T_{\mathbf{x}}) \mid \mathbf{z})\Pr(Y(\mathbf{x}), \mathbf{s}'|T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \tag{125}$$

$$= \sum_{z} \Pr(\mathbf{z})\Pr(Y(\mathbf{x})|T_{\mathbf{x}} = \vec{t}_{\mathbf{x}}, \mathbf{z}, \mathbf{\Delta}) \tag{126}$$

□