Lifted Representation of Relational Causal Models Revisited: Implications for Reasoning and Structure Learning

Sanghack Lee and **Vasant Honavar**
Artificial Intelligence Research Laboratory
College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802

Abstract

Maier et al. (2010) introduced *the relational causal model* (RCM) for representing and inferring causal relationships in relational data. A lifted representation, called *abstract ground graph* (AGG), plays a central role in reasoning with and learning of RCM. The correctness of the algorithm proposed by Maier et al. (2013a) for learning RCM from data relies on the *soundness and completeness* of AGG for *relational d-separation* to reduce the learning of an RCM to learning of an AGG. We revisit the definition of AGG and show that AGG, as defined in Maier et al. (2013b), does *not* correctly abstract all ground graphs. We revise the definition of AGG to ensure that it correctly abstracts all ground graphs. We further show that AGG representation is *not complete* for relational *d*-separation, that is, there can exist conditional independence relations in an RCM that are not entailed by AGG. A careful examination of the relationship between the lack of completeness of AGG for relational *d*-separation and *faithfulness* conditions suggests that weaker notions of completeness, namely *adjacency faithfulness* and *orientation faithfulness* between an RCM and its AGG, can be used to learn an RCM from data.

1 INTRODUCTION

Discovery of causal relationships from observational and experimental data is a central problem with applications across multiple areas of scientific endeavor. There has been considerable progress over the past decades on algorithms for eliciting causal relationships from data under a broad range of assumptions (Pearl, 2000; Spirtes et al., 2000; Shimizu et al., 2006). Most algorithms for causal discovery assume propositional data where instances are independent and identically distributed. However, in many real world applications, these assumptions are violated because the underlying data has a relational structure of the sort that is modeled in practice by an entity-relationship model (Chen, 1976). There has been considerable work on learning predictive models from relational data (Getoor and Taskar, 2007). Furthermore, researchers from different disciplines have studied causal relationships and resulting phenomena on relational world, e.g., peer effects (Sacerdote, 2000; Ogburn and VanderWeele, 2014), social contagion (Christakis and Fowler, 2007; Shalizi and Thomas, 2011), viral marketing (Leskovec et al., 2007), and information diffusion (Gruhl et al., 2004).

Motivated by the limitations of traditional approaches to learning causal relationships from relational data, Maier and his colleagues introduced the relational causal model (RCM) (Maier et al., 2010) and provided a sound and complete causal structure learning algorithm, called the relational causal discovery (RCD) algorithm (Maier et al., 2013a), for inferring causal relationships from relational data. The key idea behind RCM is that a cause and its effects are in a direct or indirect relationship that is reflected in the relational data. Traditional approaches for reasoning on and learning of a causal model cannot be trivially applied for relational causal model (Maier et al., 2013a). Reasoning on an RCM to infer a relational version of conditional independence (CI) makes use of a lifted representation, called *abstract ground graphs* (AGGs), in which traditional graphical criteria can be used to answer relational CI queries. The lifted representation is employed as an internal learning structure in RCD to reflect the inferred CI results among relational version of variables. RCD makes use of a new orientation rule designed specifically for RCM.

Motivation and Contributions RCM (Maier et al., 2010) offer an attractive model for representing, reasoning about, and learning causal relationships implicit in relational data. Arbour et al. (2014) proposed a relational version of propensity score matching method to infer (relational) causal effects from observational data. Marazopoulou et al. (2015) extended RCM to cope with *temporal* relational data. They generalized both RCM and RCD to

Temporal RCM and Temporal RCD, respectively. A lifted representation, called *abstract ground graph* (AGG), plays a central role in reasoning with and learning of RCM. The correctness of the algorithms proposed by Maier et al. (2013a) for learning RCM and Marazopoulou et al. (2015) for Temporal RCM, respectively, from observational data rely on the *soundness and completeness* of AGG for *relational d-separation* to reduce the learning of an RCM to learning of an AGG. The main contributions of this paper are as follows: (i) We show that AGG, as defined in Maier et al. (2013b) does *not* correctly abstract all ground graphs; (ii) We revise the definition of AGG to ensure that it correctly abstracts all ground graphs; (iii) We further show that AGG representation is *not complete* for relational *d*-separation, that is, there can exist conditional independence relations in an RCM that are not entailed by AGG; and (iv) Based on a careful examination of the relationship between the lack of completeness of AGG for relational *d*-separation and *faithfulness* conditions suggests that weaker notions of completeness, namely *adjacency faithfulness* and *orientation faithfulness* between an RCM and its AGG, can be used to learn an RCM from data.

2 PRELIMINARIES

We follow notational conventions introduced in (Maier et al., 2013a; **?**,b; Maier, 2014). An entity-relationship model (Chen, 1976) abstracts the *entities* (e.g., *employee*, *product*) and *relationships* (e.g., *develops*) between entities in a domain using a *relational schema*. The instantiation of the schema is called a *skeleton* where entities form a network of relationships (e.g., *Quinn-develops-Laptop*, *Roger-develops-Laptop*). Entities and relationships have attributes (e.g., *salary* of employees, *success* of products). *Cardinality constraints* specify the cardinality of relationships that an entity can participate in (e.g., *many* employees *can* develop a product.).[1] The following definitions are taken from Maier (2014):

Definition 1. A *relational schema* $\mathcal{S}$ is a tuple $\langle \mathcal{E}, \mathcal{R}, \mathcal{A}, \mathsf{card} \rangle$: a set of entity classes $\mathcal{E}$; a set of relationship classes $\mathcal{R}$ where $R_i = \langle E^i_j \rangle_{j=1}^n$ and $n = |R_i|$ is arity for R_i; attribute classes $\mathcal{A}$ where $\mathcal{A}(I)$ is a set of attribute classes of $I \in \mathcal{E} \cup \mathcal{R}$; and cardinalities $\mathsf{card} : \mathcal{R} \times \mathcal{E} \rightarrow \{\mathsf{one}, \mathsf{many}\}$.

Every relationship class R_i have two or more distinct entity classes.[2] We denote by $\mathcal{I}$ all item classes $\mathcal{E} \cup \mathcal{R}$. We denote by I_X an item class that has an attribute class X assuming, without loss of generality, that the attributes of different item classes are disjoint. Participation of an entity class E_j in a relationship class R_i is denoted by $E_j \in R_i$ if $\exists_{k=1}^{|R_i|} E^i_k = E_j$.

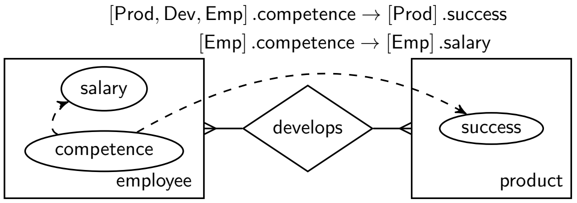
[Prod, Dev, Emp] .competence → [Prod] .success
[Emp] .competence → [Emp] .salary
salary
competence
employee
develops
success
product

Figure 1: A toy example of RCM adopted from Maier (2014) with two relational dependencies: (i) the success of a product depends on the competence of employees who develop it; (ii) employee's salary is affected by his/her competence.

Definition 2. A *relational skeleton* σ is an instantiation of relational schema $\mathcal{S}$, represented by a graph of entities and relationships. Let $\sigma(I)$ denote a set of items of item class $I \in \mathcal{I}$ in σ. Let $i_j, i_k \in \sigma$ such that $i_j \in \sigma(I_j)$, $i_k \in \sigma(I_k)$, and $I_j, I_k \in \mathcal{I}$, then we denote $i_j \sim i_k$ if there exists an edge between i_j and i_k in σ.

2.1 RELATIONAL CAUSAL MODEL

Relational causal model (RCM, Maier et al., 2010) is a causal model where causes and their effects are *related* given an underlying relational schema. For example, the success of a product depends on the competence of employees who develop the product (see Figure 1). An RCM models *relational dependencies*; each relational dependency has a cause and its effect, which are represented by *relational variable*s; a relational variable is a pair consisting of a *relational path* and an attribute.

Definition 3. A *relational path* $P = [I_j, \ldots, I_k]$ is an alternating sequence of entity class $E \in \mathcal{E}$ and relationship class $R \in \mathcal{R}$. An item class I_j is called *base class* or *perspective* and I_k is called a *terminal class*. A relational path should satisfy:

1. for every $[E, R]$ or $[R, E]$, $E \in R$;
2. for every $[E, R, E']$, $E \neq E'$; and
3. for every $[R, E, R']$, if $R = R'$, then $\mathsf{card}(R, E) = \mathsf{many}$.

All valid relational paths on the given schema $\mathcal{S}$ are denoted by $\mathbf{P}_{\mathcal{S}}$. We denote the *length* of P by $|P|$, a *subpath* by $P^{i:j} = [P_k]_{k=i}^{j}$ or $P^{i:} = [P_k]_{k=i}^{|P|}$ for $1 \leq i \leq j \leq |P|$, and the *reversed path* by $\tilde{P} = [P_{|P|}, \ldots, P_2, P_1]$. Note that all subpaths of a relational path as well as the corresponding reverse paths are valid. A *relational variable* $P.X$ is a pair of a relational path P and an attribute class X for the terminal class of P. A relational variable is said to be *canonical* if its relational path has a length equal to 1. A *relational dependency* is of the form $[I_j, \ldots, I_k].Y \rightarrow [I_j].X$

[1] The examples are taken from Maier (2014).

[2] In general, the same entity class can participate in a relationship class in two or more different roles. For simplicity, we only consider relationship classes only with distinct entity classes.

such that its cause and effect share the same base class and its effect is canonical.

Given a relational schema $\mathcal{S}$, a *relational (causal) model* $\mathcal{M}_\Theta$ is a pair of a *structure* $\mathcal{M} = \langle \mathcal{S}, \mathbf{D} \rangle$, where $\mathbf{D}$ is the set of *relational dependencies,* and Θ is a set of parameters. We assume acyclicity of the model so that the attribute classes can be partially ordered based on $\mathbf{D}$. The parameters Θ define conditional distributions, $p([I].X|\text{Pa}([I].X))$, for each pair (I, X) where $I \in \mathcal{I}$, $X \in \mathcal{A}(I)$, and $\text{Pa}([I].X)$ is a set of causes of $[I].X$, i.e., $\{P.Y|P.Y \rightarrow [I].X \in \mathbf{D}\}$. This paper focuses on the structure of RCM. Hence we often omit parameters Θ from $\mathcal{M}$.

Terminal Set and Ground Graph Because a skeleton is an instantiation of an underlying schema, a *ground graph* is an instantiation of the underlying RCM given a skeleton translating relational dependencies to every entity and relationship in the skeleton. It is obtained by interpreting the dependencies defined by the RCM on the skeleton using the *terminal sets* of each of the instances in the skeleton.

Given a relational skeleton σ, the *terminal set* of a relational path P given a base $b \in \sigma(P_1)$, denoted by $P|_b$, is the set of terminal items reachable from b when we traverse the skeleton along P. Formally, a terminal set $P|_b$ is defined recursively, $P^{1:1}|_b = \{b\}$ and

$$P^{1:\ell}|_b = \{i \in \sigma(P_\ell) \mid j \in P^{1:\ell-1}|_b,\, i \sim j\} \setminus \textstyle\bigcup_{1 \leq k < \ell} P^{1:k}|_b.$$

This implies that $P^{1:\ell}|_b$ and $P|_b$ will be disjoint for $1 \leq \ell < |P|$. Restricting the traversals so as not to revisit any previously visited items corresponds to the *bridge burning semantics* (hereinafter, BBS) (Maier et al., 2013b). The instantiation of an RCM $\mathcal{M}$ for a skeleton σ yields a ground graph which we denote by $GG_{\mathcal{M}\sigma}$. The vertices of $GG_{\mathcal{M}\sigma}$ are labeled by pairs of items and its attribute, $\{i.X \mid I \in \mathcal{I},\, i \in \sigma(I),\, X \in \mathcal{A}(I)\}$. There exists an edge $i_j.X \rightarrow i_k.Y$ in $GG_{\mathcal{M}\sigma}$ such that $i_j \in \sigma(I_j)$, $i_k \in \sigma(I_k)$, $Y \in \mathcal{A}(I_k)$, and $X \in \mathcal{A}(I_j)$ if and only if there exists a dependency $P.X \rightarrow [I_k].Y \in \mathbf{D}$ such that $i_k \in P|_{i_j}$.

In essence, RCM models dependencies on relational domain as follows: Causal relationships are described from the perspective of each item class; and are interpreted for each items to determine its causes in a skeleton yielding a ground graph. Since an RCM is defined on a given schema, RCM is interpreted on a skeleton so that every ground graph is an instantiation of the RCM.

Throughout this paper, unless specified otherwise, we assume a relational schema $\mathcal{S}$, a set of relational dependencies $\mathbf{D}$, and an RCM $\mathcal{M} = \langle \mathcal{S}, \mathbf{D} \rangle$.

3 REASONING WITH AN RCM

An RCM can be seen as a *meta* causal model or a *template* whose instantiation, a ground graph, corresponds to a *traditional* causal model (e.g., a causal Bayesian network). Reasoning with causal models relies on *conditional independence* (CI) relations among variables. Graphical criteria such as *d*-separation (Pearl, 2000) are often exploited to test CI given a model. Hence, the traditional definitions and methods for reasoning with causal models need to be "lifted" to the relational setting in order to be applicable to *relational* causal models.

Definition 4 (Relational *d*-separation (Maier, 2014))**.** Let $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ be three disjoint sets of relational variables with the same perspective $B \in \mathcal{I}$ defined over relational schema $\mathcal{S}$. Then, for relational model structure $\mathcal{M}$, $\mathbf{U}$ and $\mathbf{V}$ are *d*-separated by $\mathbf{W}$ if and only if, for all skeletons $\sigma \in \Sigma_{\mathcal{S}}$, $\mathbf{U}|_b$ and $\mathbf{V}|_b$ are *d*-separated by $\mathbf{W}|_b$ in ground graph $GG_{\mathcal{M}\sigma}$ for all $b \in \sigma(B)$.

There are two things implicit in this definition: (i) *all-ground-graphs semantics* which implies that *d*-separation must be hold over *all* instantiations of the model; (ii) the terminal set items of two *different* relational variables may overlap (which we refer to as *intersectability*). In other words, two relational variables $U = P.X$ and $V = P'.X$ of the same perspective B and the same attribute, are said to be *intersectable* if and only if:

$$\exists_{\sigma \in \Sigma_{\mathcal{S}}} \exists_{b \in \sigma(B)} P|_b \cap P'|_b \neq \emptyset. \tag{1}$$

In order to allow testing of conditional independence on all ground graphs, Maier et al. (2013a) introduced an *abstract ground graph* (AGG), which abstracts *all ground graphs* and is able to cope with the *intersectability* of relational variables. We first recapitulate the original definition of AGGs.

3.1 ORIGINAL ABSTRACT GROUND GRAPHS

An abstract ground graph $AGG_{\mathcal{M}B}$ is defined for a given relational model $\mathcal{M}$ and a perspective $B \in \mathcal{I}$ (Maier et al., 2013a), Since we fix the model, we omit the subscript $\mathcal{M}$ and denote the abstract ground graph for perspective B by AGG_B. The resulting graph consists of two types of vertices: $\mathbf{RV}_B$ and $\mathbf{IV}_B$; and two types of edges: $\mathbf{RVE}_B$ and $\mathbf{IVE}_B$.

We denote by $\mathbf{RV}_B$ the set of *all* relational variables (RV) whose paths originate in B. We denote by $\mathbf{RVE}_B$ the set of all edges between the relational variables in $\mathbf{RV}_B$. A relational variable edge (RVE) implies *direct* influence arising from one or more dependencies in $\mathbf{D}$. There is an RVE $P.X \rightarrow Q.Y$ if there exists a dependency $R.X \rightarrow [I_Y].Y \in \mathbf{D}$ that can be interpreted as a direct influence from $P.X$ to $Q.Y$ from perspective B. Such an interpretation is implemented by an extend function, which takes two relational paths and produces a set of relational paths: If $P \in \mathsf{extend}(Q, R)$, then there exists an RVE $P.X \rightarrow Q.Y$

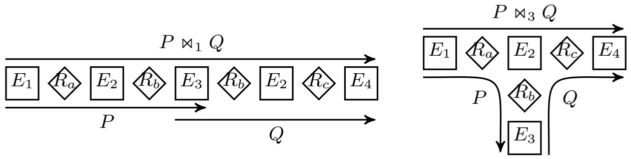

Figure 2: A schematic example of how extend is computed showing two relational paths in $P \bowtie Q$ where $\mathsf{card}(R_b, E_3) = \mathsf{many}$. If $\mathsf{card}(R_b, E_3)$ is one, then $P \bowtie_1 Q$ is not valid due to rule 3 of relational path. A path $P \bowtie_2 Q$ is invalid due to the violation of rule 2, i.e., $[\ldots, E_2, R_b, E_2, \ldots]$.

where

$$\mathsf{extend}(Q, R) = \{Q^{1:|Q|-i} + R^{i:} | i \in \mathsf{pivots}(\tilde{Q}, R)\} \cap \mathbf{P}_{\mathcal{S}}, \quad (2)$$

$\mathsf{pivots}(S, T) = \{i | S^{1:i} = T^{1:i}\}$, and '+' is a concatenation operator. We will use a binary *join* operator '$\bowtie$' for extend and denote $Q^{1:|Q|-i} + R^{i:}$ by $Q \bowtie_i R$ for a pivot i. A schematic overview of extend is shown in Figure 2.

We denote by $\mathbf{IV}_B$ the set of *intersection variables* (IVs), i.e., unordered pairs of *intersectable* relational variables in $\mathbf{RV}_B$. Given two RVs $P.X$ and $P'.X$ that are intersectable with each other, we denote the resulting intersection variable by $P.X \cap P'.X$ (Here, the intersection symbol '$\cap$' denotes *intersectability* of the two relational variables). By the definition (Maier et al., 2013b), if there exists an RVE $P.X \to Q.Y$, then there exist edges $P.X \cap P'.X \to Q.Y$ and $P.X \to Q.Y \cap Q'.Y$ for every P' and Q' intersectable with P and Q, respectively. The IVs and the edges that connect them with RVs (IVEs) correspond to *indirect influences* (arising from intersectability) as opposed to *direct* influence due to dependencies (which are covered by RVs and RVEs). We denote by $\mathbf{IVE}_B$ the set of all such edges that connect RVs with IVs.

Two AGGs with different perspectives share no vertices nor edges. Hence, we view all AGGs, $\{AGG_B\}_{B \in \mathcal{I}}$, as a collection or a single multi-component graph $\mathbf{AGG} = \bigcup_{B \in \mathcal{I}} AGG_B$. We similarly define $\mathbf{RV}$, $\mathbf{IV}$, $\mathbf{RVE}$, and $\mathbf{IVE}$ as the unions of their perspective-based counterparts.

For any mutually disjoint sets of relational variables $\mathbf{U}, \mathbf{V}$, and $\mathbf{W}$, one can test $\mathbf{U} \perp \mathbf{V} \mid \mathbf{W}$, conditional independence admitted by the underlying probability distribution, by checking $\bar{\mathbf{U}} \perp\!\!\!\perp \bar{\mathbf{V}} \mid \bar{\mathbf{W}}$ (traditional) d-separation on an AGG[3], where $\bar{\mathbf{V}}$ includes $\mathbf{V}$ and their related IVs, $\bar{\mathbf{V}} = \mathbf{V} \cup \{V \cap T \in \mathbf{IV} \mid V \in \mathbf{V}\}$. Figure 3 illustrates relational d-separation on an AGG.

We later show that the preceding definition of AGG does *not* properly abstract all ground graphs; nor does it guarantee the correctness of reasoning about relational d-separation in an RCM. We revise the definition of AGG (Section 3.2) so as to ensure that the resulting AGG abstracts all ground graphs. However, we find that even with the revised definition of AGG, the AGG representation is *not complete* for relational d-separation, that is, there can exist conditional independence relations in an RCM that are not entailed by AGG (Section 4.1). A careful examination of the lack of completeness of AGG for relational d-separation with respect to causal faithfulness yields useful insights that allow us to make use of weaker notions of faithfulness to learn RCM from data (Section 4.2).

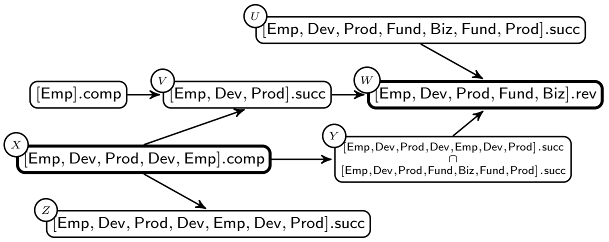
U
[Emp, Dev, Prod, Fund, Biz, Fund, Prod].succ
[Emp].comp
V
[Emp, Dev, Prod].succ
W
[Emp, Dev, Prod, Fund, Biz].rev
X
[Emp, Dev, Prod, Dev, Emp].comp
Y
[Emp,Dev,Prod,Dev,Emp,Dev,Prod].succ
∩
[Emp,Dev,Prod,Fund,Biz,Fund,Prod].succ
Z
[Emp, Dev, Prod, Dev, Emp, Dev, Prod].succ

Figure 3: An AGG example excerpted from Maier (2014) with *business unit* (Biz) which *funds* (Fund) its *products* from its revenue (rev). The revenue of business units that fund the products developed by an employee (W) is affected by the employee's co-workers' competence (X), i.e., $\bar{W} \not\perp\!\!\!\perp \bar{X}$. Two are conditionally independent by blocking both V and Y. Since IV Y is in $\bar{U}$ and $\bar{Z}$, both $\bar{W} \perp\!\!\!\perp \bar{X} | \overline{\{V, U\}}$ and $\bar{W} \perp\!\!\!\perp \bar{X} | \overline{\{V, Z\}}$ hold, which are equivalent to $(W \perp X | \{V, U\})_{\mathcal{M}}$ and $(W \perp X | \{V, Z\})_{\mathcal{M}}$, respectively.

3.2 ABSTRACT GROUND GRAPHS - A REVISED DEFINITION

Because of the importance of $\mathbf{IV}$ and $\mathbf{IVE}$ in $\mathbf{AGG}$ in reasoning about relational d-separation, it is possible that errors in abstracting all ground graphs could lead to errors in CI relations inferred from an $\mathbf{AGG}$. We proceed to show that 1) the criteria for determining *intersectability* (Maier, 2014) are *not sufficient,* and 2) the definition of $\mathbf{IVE}$, as it stands, does not guarantee the *soundness* of AGG as an abstract representation of the all ground graphs of an RCM. We provide the *necessary and sufficient* criteria for determining IVs and a *sound* definition for IVEs.

3.2.1 Intersectability and IV

The declarative characterization of *intersectability* (Eq. 1) does not offer practical procedural criteria to determine *intersectability*. Based on the criteria (Maier, 2014), two different relational paths P and Q are *intersectable* if and only if 1) they share the same perspective, say $B \in \mathcal{I}$, and 2) they share the common terminal class, and 3) one path is

[3] We denote conditional independence by '$\perp$' in general. We use '$\perp\!\!\!\perp$' to represent (traditional) d-separation on a directed acyclic graph., e.g., $\mathbf{AGG}_{\mathcal{M}}$ or $GG_{\mathcal{M}\sigma}$. Furthermore, we parenthesize conditional independence and use a subscript to specify the scope of the conditional independence, if necessary.

$$\begin{array}{lr} 1)\ \exists_{\sigma\in\Sigma_{\mathcal{S}}}\exists_{b\in\sigma(B)}\exists_{i_j\in Q|_b} & R|_{i_j}\cap P|_b \qquad \neq\emptyset \\ 2)\ \exists_{\sigma\in\Sigma_{\mathcal{S}}}\exists_{b\in\sigma(B)} & P|_b\cap P'|_b\neq\emptyset \\ 3)\ \exists_{\sigma\in\Sigma_{\mathcal{S}}}\exists_{b\in\sigma(B)}\exists_{i_j\in Q|_b} & R|_{i_j}\cap P|_b\cap P'|_b\neq\emptyset \end{array}$$

Figure 4: Comparison of 1) the necessary condition of the existence of an RVE $P \to Q$ through R, the cause path of a dependency (attributes are omitted), 2) intersectability between P and P', and 3) co-intersectability of $\langle Q, R, P, P' \rangle$.

not a prefix of the other. We will prove that the preceding criteria are *not sufficient*. In essence, we will show the conditions under which non-emptiness of $P|_b \cap Q|_b$ for any $b \in \sigma(B)$ in any skeleton σ always contradicts the BBS. For the proof, we define $\mathsf{LLRSP}(P, Q)$ (*the length* of *the longest required shared path*) for two relational paths P and Q of the common perspective as

$$\max\{\ell \mid P^{1:\ell} = Q^{1:\ell},\ \forall_{\sigma\in\Sigma_{\mathcal{S}}}\forall_{b\in\sigma(B)} \left|P^{1:\ell}|_b\right| = 1\}.$$

$\mathsf{LLRSP}(P, Q)$ is computed as follows. Initially set $\ell = 1$ since $P_1 = Q_1$. Repeat incrementing ℓ by 1 if $P_{\ell+1} = Q_{\ell+1}$ and either $P_\ell \in \mathcal{R}$ or $P_\ell \in \mathcal{E}$ with $\mathsf{card}(P_\ell, P_{\ell+1}) = \mathsf{one}$.

Lemma 1. *Given a relational schema $\mathcal{S}$, let P and Q be two different relational paths satisfying the (necessary) criteria of Maier (2014) and $|Q| \leq |P|$. Let m and n be $\mathsf{LLRSP}(P, Q)$ and $\mathsf{LLRSP}(\tilde{P}, \tilde{Q})$, respectively. Then, P and Q are intersectable if and only if $m + n \leq |Q|$.*

Proof. See Appendix. □

The lemma demonstrates the criteria by Maier (2014) do not rule out the case of $m+n > |Q|$ where P and Q cannot be intersectable.

3.2.2 Co-intersectability and IVE

Based on the definition (Maier, 2014), an IVE exists between an IV, $U \cap V$, and an RV, W, if and only if there exists an RVE between U and W or V and W. It would indeed be appealing to define IV, $U \cap V$, such that it inherits properties of the corresponding RVs, U and V. However, the abstract ground graph resulting from such a definition turns out to be not a sound representation of the underlying ground graphs. We proceed to prove this result.

Definition 5 (Co-intersectability)**.** Given a relational schema $\mathcal{S}$, let Q, R, P, and P' be valid relational paths of the same perspective B where $P \in Q \bowtie R$ and P and P' are intersectable. Then, a tuple $\langle Q, R, P, P' \rangle$ is said to be *co-intersectable* if and only if

$$\exists_{\sigma\in\Sigma_{\mathcal{S}}}\exists_{b\in\sigma(B)}\exists_{i_j\in Q|_b} R|_{i_j} \cap P|_b \cap P'|_b \neq \emptyset. \qquad (3)$$

We relate co-intersectability with the definition of IVE. Let an RVE $P.X \to Q.Y$ is due to some dependencies $R.X \to [I_Y].Y \in \mathbf{D}$ where $P \in Q \bowtie R$. This implies

$$\exists_{\sigma\in\Sigma_{\mathcal{S}}}\exists_{b\in\sigma(B)}\exists_{i_j\in Q|_b} R|_{i_j} \cap P|_b \neq \emptyset, \qquad (4)$$

and there are edges from X of $R|_{i_j} \cap P|_b$ to Y of $Q|_b$ in GG_σ. In order for the intersectability of P' with P translates into an influence between P and Q, it is necessary that there exists a skeleton that admits such influence. However, we can construct a counterexample that satisfies the necessary conditions for the existence of an RVE and the conditions for intersectability but does *not* satisfy the conditions for co-intersectability (see Figure 4 for a comparison of Eq. 4, 1, and 3).

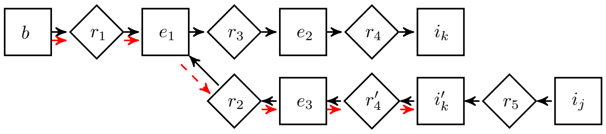

Figure 5: A schematic illustration of Example 1 superimposing a skeleton and relational paths. The items for P' starting with b should follow a dashed red line, and, hence, P' cannot be related to a ground graph edge between i_k and i_j, i.e., an RVE between P and Q (attributes and connections between entities and relationships are omitted).

Example 1. Let $\mathcal{S}$ be a relational schema where $\mathbf{E} = \{I_j, I_k, B, E_1, E_2, E_3\}$, $\mathbf{R} = \{R_i\}_{i=1}^{5}$ such that $R_1 = \langle B, E_1 \rangle$, $R_2 = \langle E_1, E_3 \rangle$, $R_3 = \langle E_1, E_2 \rangle$, $R_4 = \langle E_2, E_3, I_k \rangle$, $R_5 = \langle I_k, I_j \rangle$ with the cardinality of each relationship and each entity in the relationship being one. Let

- $Q = [B, R_1, E_1, R_2, E_3, R_4, I_k, R_5, I_j]$,
- $R = [I_j, R_5, I_k, R_4, E_3, R_2, E_1, R_3, E_2, R_4, I_k]$,
- $P = [B, R_1, E_1, R_3, E_2, R_4, I_k]$, and
- $P' = [B, R_1, E_1, R_2, E_3, R_4, I_k]$.

Observe that

1. $P \in \mathsf{extend}\,(Q, R)$;
2. P' and P are intersectable; and
3. P' is a subpath of Q.

This example satisfies Eq. 1 and Eq. 4. Assume for contradiction that there exists a skeleton σ satisfying Eq. 3. Since, in this example, the cardinality of each relationship and each entity in the relationship is one, for each $b \in \sigma(B)$, there exists only one $i_j \in Q|_b$ and only one $i_k \in P|_b$. By the assumption, $P'|_b = \{i_k\}$. Since P' is a subpath of Q, $P'|_b$ will end at $i'_k = R^{1:3}|_{i_j}$ (see Figure 5). Due to BBS, $R|_{i_j} \cap R^{1:3}|_{i_j} = \emptyset$, that is, $\{i_k\} \cap \{i'_k\} = \emptyset$. This contradicts the assumption that $i_k = i'_k$.

This counterexample clearly represents there is an interdependency between intersection variables and RVEs. Therefore, we revise the definition of **IVE** accompanying co-intersectability.

Definition 6 (IVE)**.** There exists an IVE edge, $P.X \cap P'.X \to Q.Y$ (or $P.X \to Q.Y \cap Q'.Y$), if and only if there exists a relational path R such that $R.X \to [I_Y].Y \in \mathbf{D}$, $P \in Q \bowtie R$, and $\langle Q, R, P, P' \rangle$ (or $\langle P, \tilde{R}, Q, Q' \rangle$) is *co-intersectable*.

To determine IVEs, *co-intersectability* of a tuple can be computed by solving a constraint satisfaction problem involving four paths in the tuple.

Implications of Co-intersectability We investigated the necessary and sufficient criteria for intersectability and revised the definition of IVE so as to guarantee that AGG correctly abstracts all ground graphs as asserted (although incorrectly) by Theorem 4.5.2 (Maier, 2014). The new criterion, called *co-intersectability,* is especially interesting since it describes the interdependency between intersection variables and related relational variable edges. Several of the key results (e.g., soundness and completeness of AGG for relational *d*-separation, Theorem 4.5.2) and concepts (e.g., (B,h)-reachability) of Maier (2014) are based on *independence* between intersection variables and related relational variable edges. Hence, it is useful to carefully scrutinize the relationship between AGG and relational *d*-separation.

4 NON-COMPLETENESS OF AGG FOR RELATIONAL D-SEPARATION

We first revisit the definition of relational *d*-separation. Given three disjoint sets of relational variables $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$ of a common perspective $B \in \mathcal{I}$, $\mathbf{U}$ and $\mathbf{V}$ are relational *d*-separated given $\mathbf{W}$, denoted by $(\mathbf{U} \perp \mathbf{V} \mid \mathbf{W})_{\mathcal{M}}$, if and only if

$$\forall_{\sigma \in \Sigma_{\mathcal{S}}} \forall_{b \in \sigma(B)} \left(\mathbf{U}|_b \perp\!\!\!\perp \mathbf{V}|_b \mid \mathbf{W}|_b\right)_{GG_{\mathcal{M}\sigma}}.$$

From Theorem 4.5.4 of (Maier, 2014), the lifted representation $\mathbf{AGG}_{\mathcal{M}}$ is said to be sound (or complete) for relational *d*-separation of $\mathcal{M}$ if (traditional) *d*-separation holds on the $\mathbf{AGG}_{\mathcal{M}}$ with a modified CI query only when (or whenever) relational *d*-separation holds true. Then, the completeness of AGG for relational *d*-separation can be represented as

$$(\mathbf{U} \perp \mathbf{V} \mid \mathbf{W})_{\mathcal{M}} \Rightarrow \left(\bar{\mathbf{U}} \perp\!\!\!\perp \bar{\mathbf{V}} \mid \bar{\mathbf{W}}\right)_{\mathbf{AGG}_{\mathcal{M}}}.$$

The completeness can be proved by the construction of a skeleton $\sigma \in \Sigma_{\mathcal{S}}$ demonstrating *d*-connection $\left(\mathbf{U}|_b \not\perp\!\!\!\perp \mathbf{V}|_b \mid \mathbf{W}|_b\right)_{GG_{\mathcal{M}\sigma}}$ for some $b \in \sigma(B)$ if $(\bar{\mathbf{U}} \not\perp\!\!\!\perp \bar{\mathbf{V}} \mid \bar{\mathbf{W}})_{\mathbf{AGG}_{\mathcal{M}}}$. In other words, we might disprove the completeness by showing

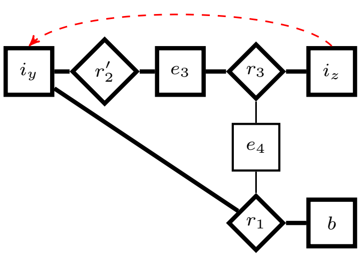

Figure 6: Co-intersectability of $\langle Q, D_2, S, S' \rangle$ where $i_y \in Q|_b$, $i_z \in D_2|_{i_y}$, $i_z \in S|_b$, and $i_z \in S'|_b$. The thick line highlights items for S' from b to i_z. The red dashed line represents the instantiation of an RVE $S.Z \to Q.Y$ as $i_z.Z \to i_y.Y$ in a ground graph (attributes are omitted).

$$\left(\bar{\mathbf{U}} \not\perp\!\!\!\perp \bar{\mathbf{V}} \mid \bar{\mathbf{W}}\right)_{\mathbf{AGG}_{\mathcal{M}}} \wedge \\ \forall_{\sigma \in \Sigma_{\mathcal{S}}} \forall_{b \in \sigma(B)} \left(\mathbf{U}|_b \perp\!\!\!\perp \mathbf{V}|_b \mid \mathbf{W}|_b\right)_{GG_{\mathcal{M}\sigma}}.$$

4.1 A COUNTEREXAMPLE

The following counterexample shows that AGG is not complete for relational *d*-separation.

Example. Let $\mathcal{S} = \langle \mathcal{E}, \mathcal{R}, \mathcal{A}, \mathsf{card} \rangle$ be a relational schema such that: $\mathcal{E} = \{E_i\}_{i=1}^{5}$; $\mathcal{R} = \{R_j\}_{j=1}^{3}$ with $R_1 = \langle E_1, E_2, E_4 \rangle$, $R_2 = \langle E_2, E_3 \rangle$, and $R_3 = \langle E_3, E_4, E_5 \rangle$; $\mathcal{A} = \{E_2 : \{Y\}, E_3 : \{X\}, E_5 : \{Z\}\}$; and $\forall_{R \in \mathcal{R}} \forall_{E \in R} \mathsf{card}\,(R, E) = \mathsf{one}$. Let $\mathcal{M} = \langle \mathcal{S}, \mathbf{D} \rangle$ be a relational model with

$$\mathbf{D} = \{D_1.X \to [I_Y].Y,\ D_2.Z \to [I_Y].Y\}$$

such that $D_1 = [E_2, R_2, E_3, R_3, E_4, R_1, E_2, R_2, E_3]$ and $D_2 = [E_2, R_2, E_3, R_3, E_5]$. Let $P.X$, $Q.Y$, $S.Z$, and $S'.Z$ be four relational variables of the same perspective $B = E_1$ where their relational paths are distinct where

- $P = [E_1, R_1, E_2, R_2, E_3]$,
- $Q = [E_1, R_1, E_4, R_3, E_3, R_2, E_2]$,
- $S = [E_1, R_1, E_4, R_3, E_5]$, and
- $S' = [E_1, R_1, E_2, R_2, E_3, R_3, E_5]$.

Given the above example, we can make two claims.

Claim 1. $\left(\overline{P.X} \not\perp\!\!\!\perp \overline{S'.Z} \mid \overline{Q.Y}\right)_{\mathbf{AGG}_{\mathcal{M}}}$.

Proof. See Appendix. □

Assuming that AGG is complete for relational *d*-separation, we can infer $(P.X \not\perp S'.Z \mid Q.Y)_{\mathcal{M}}$ and there must exist a pair of a skeleton σ and a base $b \in \sigma(B)$ that satisfies $(P.X|_b \not\perp\!\!\!\perp S'.Z|_b \mid Q.Y|_b)_{GG_{\mathcal{M}\sigma}}$. However, we claim that such a skeleton and base may not exist.

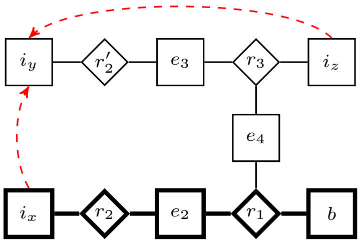

Figure 7: A subgraph of ground graphs to represent $i_x \rightarrow i_y \leftarrow i_z$. Only this substructure satisfies BBS assumption and cardinality constraints.

Claim 2. There is no $\sigma \in \Sigma_{\mathcal{S}}$ and $b \in \sigma(B)$ such that

$$(P.X|_b \not\perp\!\!\!\perp S'.Z|_b \mid Q.Y|_b)_{GG_{\mathcal{M}\sigma}}\,.$$

Proof. See Appendix. □

The counterexample demonstrates that a *d*-connection path captured in an $\mathbf{AGG}_{\mathcal{M}}$ might not have a corresponding *d*-connection path in *any* ground graph.

Corollary 1. *The revised (as well as the original) abstract ground graph for an RCM is not complete for relational d-separation.*

It is possible that an additional test can be utilized to check whether there *exists* such a ground graph that can represent a *d*-connection path captured in $\mathbf{AGG}_{\mathcal{M}}$. However, the efficiency of such an additional test is unknown and designing such a test is beyond the scope of this paper.

4.2 RELATING NON-COMPLETENESS WITH FAITHFULNESS

In light of the preceding result that AGG is not complete for relational *d*-separation, we proceed to examine the relationship between an RCM and its lifted representation in terms of the sets of conditional independence relationships that they admit. In RCM, there are several levels of relationship regarding the sets of conditional independence: between the underlying probability distributions and the ground graphs, between the ground graphs of an RCM and the RCM, and between the RCM and its AGG:

$$\{p \leftrightarrow GG_{\mathcal{M}_\Theta\sigma}\}_{\sigma\in\Sigma_{\mathcal{S}}} \leftrightarrow \mathcal{M}_\Theta \leftrightarrow \mathbf{AGG}_{\mathcal{M}}$$

In RCM, the *causal Markov condition* and *causal faithfulness condition* (see below) can be applied between a ground graph $GG_{\mathcal{M}\sigma}$ and its underlying probability distribution p. Both conditions are assumed for learning an RCM from relational data. Relational *d*-separation requires a set of conditional independence of $\mathcal{M}_\Theta$ using those deduced from every ground graph $GG_{\mathcal{M}_\Theta\sigma}$ for every $\sigma \in \Sigma_{\mathcal{S}}$. In light of the lack of completeness of AGG for relational *d*-separation, the set of conditional independence relations admitted by $\mathcal{M}_\Theta$ and its lifted representation $\mathbf{AGG}_{\mathcal{M}}$ are *not* necessarily equivalent (see Corollary 1).

We will relate $\mathcal{M}$ and $\mathbf{AGG}_{\mathcal{M}}$ using an *analogy* of *causal Markov condition* and *faithfulness* (Spirtes et al., 2000; Ramsey and Spirtes, 2006) interpreting $\mathbf{AGG}_{\mathcal{M}}$ and $\mathcal{M}$ as a DAG G and a distribution p, respectively. We first recapitulate the definitions for causal Markov condition and faithfulness.

Definition 7 (Causal Markov Condition (Ramsey and Spirtes, 2006))**.** Given a set of variables whose causal structure can be represented by a DAG G, every variable is probabilistically independent of its non-effects (non-descendants in G) conditional on its direct causes (parents in G).

The causal Markov condition (i.e., local Markov condition) is not directly translated into the relationship between $\mathbf{AGG}_{\mathcal{M}}$ and $\mathcal{M}$ since they refer to different variables. However, the soundness of $\mathbf{AGG}_{\mathcal{M}}$ for relational *d*-separation of $\mathcal{M}$ (i.e., global Markov condition) would be sufficient to interpret causal Markov condition between $\mathbf{AGG}_{\mathcal{M}}$ and $\mathcal{M}$. That is,

$$\forall_{U,V,W\in\mathbf{RV}} \left(\bar{U} \perp\!\!\!\perp \bar{V} \mid \bar{W}\right)_{\mathbf{AGG}_{\mathcal{M}}} \Rightarrow (U \perp V \mid W)_{\mathcal{M}}$$

where U, V, and W are distinct relational variables sharing a common perspective.

Definition 8 (Causal Faithfulness Condition (Ramsey and Spirtes, 2006))**.** Given a set of variables whose causal structure can be represented by a DAG, no conditional independence holds unless entailed by the causal Markov condition.

By the counterexample above, $\mathcal{M}$ is not strictly *faithful* to $\mathbf{AGG}_{\mathcal{M}}$, because more conditional independences hold in $\mathcal{M}$ than those entailed by $\mathbf{AGG}_{\mathcal{M}}$.

4.2.1 Weaker Faithfulness Conditions

Ramsey and Spirtes (2006) showed that the two weaker types of faithfulness – *adjacency-faithfulness* and *orientation-faithfulness* – are sufficient to retrieve a maximally-oriented causal structure from a data under the causal Markov condition. What we have showed is that there are more conditional independence hold in $\mathcal{M}$ than those entailed by its corresponding $\mathbf{AGG}_{\mathcal{M}}$. However, the two weaker faithfulness conditions hold true (if they are appropriately interpreted in an RCM and its lifted representation).

Adjacency-Faithfulness

Definition 9 (Adjacency-Faithfulness (Ramsey and Spirtes, 2006))**.** Given a set of variables $\mathbf{V}$ whose causal structure can be represented by a DAG G, if two variables X, Y are adjacent in G, then they are dependent conditional on any subset of $\mathbf{V} \setminus \{X, Y\}$.

Let U, V be two distinct relational variables of the same perspective B. We limit U and V to be non-intersectable to each other. Otherwise, they must not be adjacent to each other by the definition of RCM since an edge between intersectable relational variables yields a feedback in a ground graph. If there is an edge $U \rightarrow V$ in $\mathbf{AGG}_{\mathcal{M}}$,

$$\forall_{\mathbf{W}\subseteq\mathbf{RV}_B\setminus\{U,V\}}\,(U \not\perp V \mid \mathbf{W})_{\mathcal{M}}$$

We can construct a skeleton $\sigma \in \Sigma_{\mathcal{S}}$ where its corresponding ground graph $GG_{\mathcal{M}\sigma}$ satisfies that $U|_b$ and $V|_b$ are singletons and $U|_b \cup V|_b$ are disjoint to $(\mathbf{RV}_B \setminus \{U, V\})\,|_b$ for $b \in \sigma(B)$. Lemma 4.4.1 by Maier (2014) describes a method to construct a *minimal* skeleton to represent U and V with a single $b \in \sigma(B)$. It guarantees that $U|_b$ and $V|_b$ are singletons and every relational variable $W \in \mathbf{RV}_B \setminus \{U, V\}$ satisfies $W|_b \cap U|_b = \emptyset$ and $W|_b \cap V|_b = \emptyset$.

Orientation-Faithfulness

Definition 10 (Orientation-Faithfulness (Ramsey and Spirtes, 2006))**.** Given a set of variables $\mathbf{V}$ whose causal structure can be represented by a DAG G, let $\langle X, Y, Z\rangle$ be any unshielded triple in G.

(O1) if $X \rightarrow Y \leftarrow Z$, then X and Z are dependent given any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y;

(O2) otherwise, X and Z are dependent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y.

Let U, V, and W be three distinct relational variables of the same perspective B forming an unshielded triple in $\mathbf{AGG}_{\mathcal{M}}$. Similarly, V is not intersectable to both U and W. The condition (O1) can be written as

$$\forall_{\mathbf{T}\subseteq\mathbf{RV}_B\setminus\{U,W\}}(U \not\perp W \mid \mathbf{T} \cup \{V\})_{\mathcal{M}}$$

if edges are oriented as $U \rightarrow V \leftarrow W$ in $\mathbf{AGG}_{\mathcal{M}}$. Otherwise,

$$\forall_{\mathbf{T}\subseteq\mathbf{RV}_B\setminus\{U,W\}}(U \not\perp W \mid \mathbf{T} \setminus \{V\})_{\mathcal{M}}$$

for the condition (O2). Again, constructing a minimal skeleton for U, V, and W guarantees that no $T \in \mathbf{RV}_B \setminus \{U, V, W\}$ can represent any item in $\{U|_b, V|_b, W|_b\}$. Thus, the existence of V in the conditional determines (in)dependence in the ground graph induced from the minimal skeleton.

Learning RCM with Non-complete AGG RCD (Relational Causal Discovery, Maier et al. (2013a)) is an algorithm for learning the structure of an RCM from relational data. In learning RCM, AGG plays a key role: AGG is constructed using CI tests to obtain the relational dependencies of an RCM. The lack of completeness of AGG for relational *d*-separation in RCM raises questions about the correctness of RCD. A careful examination of AGG through the lens of faithfulness suggests that *adjacency-faithful* and *orientation-faithful* conditions can be applied to $\mathbf{AGG}_{\mathcal{M}}$ to recover correct partially-oriented dependencies for an RCM. However, it is still unclear whether RCD recovers maximally-oriented dependencies with the acyclicity of AGG (i.e., relational variables) not the acyclicity of RCM (i.e., attribute classes). This raises the possibility of an algorithm for learning the structure of an RCM from relational data that does not require the intermediate step of constructing a lifted representation.

5 CONCLUDING REMARKS

There is a growing interest in relational causal models (Maier et al., 2010, 2013a; **?**; Maier, 2014; Arbour et al., 2014; Marazopoulou et al., 2015). A lifted representation, called *abstract ground graph* (AGG), plays a central role in reasoning with and learning of RCM. The correctness of the algorithm proposed by Maier et al. (2013a) for learning RCM from data relies on the *soundness and completeness* of AGG for *relational d-separation* to reduce the learning of an RCM to learning of an AGG. We showed that AGG, as defined in (Maier et al., 2013a), does *not* correctly abstract all ground graphs. We revised the definition of AGG to ensure that it correctly abstracts all ground graphs. We further showed that AGG representation is *not complete* for relational *d*-separation, that is, there can exist conditional independence relations in an RCM that are not entailed by AGG. Our examination of the relationship between the lack of completeness of AGG for relational *d*-separation and *faithfulness* suggests that weaker notions of completeness, namely *adjacency faithfulness* and *orientation faithfulness* between an RCM and its AGG can be used to learn an RCM from data. Work in progress is aimed at: 1) identifying the necessary and sufficient criteria for guaranteeing the completeness of AGG for relational *d*-separation; 2) establishing whether the RCD algorithm outputs a maximally-oriented RCM even when the completeness of AGG for relational *d*-separation does not hold; and 3) devising a structure learning algorithm that does not rely on a lifted representation.

APPENDIX

We first prove Lemma 1 in Section 3.2.1.

Lemma. *Given a relational schema $\mathcal{S}$, let P and Q be two different relational paths satisfying the (necessary) criteria of Maier (2014) and $|Q| \leq |P|$. Let m and n be $\mathsf{LLRSP}(P, Q)$ and $\mathsf{LLRSP}(\tilde{P}, \tilde{Q})$, respectively. Then, P and Q are intersectable if and only if $m + n \leq |Q|$.*

Proof. (If part) If $m + n \leq |Q|$, then we can construct a skeleton σ such that $P|_b \cap Q|_b \neq \emptyset$ for some $b \in \sigma(P_1)$ by adding unique items for Q and for $P^{m+1:|P|-n}$ and complete the skeleton in the same manner as shown in Lemma 3.4.1 (Maier, 2014). Note that if $m + n = |Q|$, then $|P| \geq |Q| + 2$ since $P \neq Q$ and a relational path is an alternating sequence. This guarantees that there are at least two items for $P^{m+1:|P|-n}$.

(Only if part) Let c be in $P|_b \cap Q|_b$ for some arbitrary skeleton $\sigma \in \Sigma_{\mathcal{S}}$ and $b \in \sigma(P_1)$. Then, there should be two lists of items corresponding to P and Q sharing the first m and the last n. The condition $m + n > |Q|$ implies $Q|_b$ is a singleton set. We define

$$\mathbf{p} = \langle p_1, \ldots, p_m, p_{|P|-n+1}, \ldots, p_{|P|} \rangle$$

and

$$\mathbf{q} = \langle q_1, \ldots, q_{|Q|} \rangle,$$

where $\{q_\ell\} = Q^{1:\ell}|_b$ and $\{p_\ell\} = P^{1:\ell}|_b$ for $1 \leq \ell \leq m$, and $p_{|P|-l+1} \in \tilde{P}^{1:l}|_c$ for $1 \leq l \leq n$. We can see that $p_1 = q_1 = b$ and $p_{|P|} = q_{|Q|} = c$. Moreover,

$$p_m = q_m = q_{|Q|-(|Q|-m)} = p_{|P|-(|Q|-m)}$$

by the definition of LLRSP. If $|Q| < |P|$, then $m \neq |P| - |Q| + m$ and mth item for P is repeated at $|P| - (|Q| - m)$th, which violates the BBS. Otherwise, it is not the case, since $|P| = |Q|$ implies $\mathbf{p} = \mathbf{q}$ and, hence, $P = Q$ by the definition of LLRSP, which contradicts the assumption that P and Q are different relational paths. □

We provide proofs for two claims regarding the counterexample in Section 4.1.

Claim. $\left(\overline{P.X} \not\perp\!\!\!\perp \overline{S'.Z} \mid \overline{Q.Y}\right)_{\mathbf{AGG}_{\mathcal{M}}}$.

Proof. By the definition of RVE, there are RVEs $P.X \to Q.Y$ and $Q.Y \leftarrow S.Z$ in $\mathbf{AGG}_{\mathcal{M}}$ since $P = Q \bowtie_6 D_1$ and $S \in Q \bowtie_4 D_2$. Moreover, there is an IVE $Q.Y \leftarrow S.Z \cap S'.Z$ in $\mathbf{AGG}_{\mathcal{M}}$ since 1) S and S' are *intersectable,* 2) there is an RVE $Q.Y \leftarrow S.Z$, and 3) $\langle Q, D_2, S, S' \rangle$ is *co-intersectable* (see Figure 6).[4] Since $P.X \to Q.Y \leftarrow S.Z \cap S'.Z$ and $S.Z \cap S'.Z \in \overline{S'.Z}$, we derive $\left(P.X \not\perp\!\!\!\perp \overline{S'.Z} \mid Q.Y\right)_{\mathbf{AGG}_{\mathcal{M}}}$, which implies $\left(\overline{P.X} \not\perp\!\!\!\perp \overline{S'.Z} \mid Q.Y\right)_{\mathbf{AGG}_{\mathcal{M}}}$. Furthermore, conditioning on $\overline{Q.Y}$, compared to $Q.Y$, does not block any possible d-connection paths between $\overline{P.X}$ to $\overline{S'.Z}$ since there are only incoming edges to $\overline{Q.Y}$. Finally, $\left(\overline{P.X} \not\perp\!\!\!\perp \overline{S'.Z} \mid \overline{Q.Y}\right)_{\mathbf{AGG}_{\mathcal{M}}}$ holds. □

Claim. There is no $\sigma \in \Sigma_{\mathcal{S}}$ and $b \in \sigma(B)$ such that

$$\left(P.X|_b \not\perp\!\!\!\perp S'.Z|_b \mid Q.Y|_b\right)_{GG_{\mathcal{M}\sigma}}.$$

Proof. Suppose that there exist such a skeleton σ and base $b \in \sigma(B)$ satisfying $\left(P.X|_b \not\perp\!\!\!\perp S'.Z|_b \mid Q.Y|_b\right)_{GG_{\mathcal{M}\sigma}}$. Every terminal set for P, Q, and S' given the base must not be empty because of the definition of d-separation and the fact that attribute classes X and Z are connected only through Y (i.e., Y is a collider). Since every cardinality is one, terminal sets must be singletons. Let $\{i_x\} = P.X|_b$, $\{i_y\} = Q.Y|_b$, and $\{i_z\} = S'.Y|_b$. Furthermore, since i_x and i_z must be d-connected given i_y, $GG_{\mathcal{M}\sigma}$ must have two edges $i_x \to i_y \leftarrow i_z$, which requires $i_x \in D_1|_{i_y}$ and $i_z \in D_2|_{i_y}$. However, due to BBS and cardinality constraints (i.e., one), there exists only one possible structure (see Figure 7) where i_x and i_z are the cause of i_y while satisfying all previously mentioned conditions except $\{i_z\} = S'.Y|_b$. In other words, the constraint $\{i_z\} = S'.Y|_b$ violates with the set of the rest of conditions. Hence, there exists no such skeleton and base. □

Acknowledgments

This research was supported in part by the Edward Frymoyer Endowed Professorship held by Vasant Honavar and in part by the Center for Big Data Analytics and Discovery Informatics which is co-sponsored by the Institute for Cyberscience, the Huck Institutes of the Life Sciences, and the Social Science Research Institute at the Pennsylvania State University.

References

Arbour, D., Marazopoulou, K., Garant, D., and Jensen, D. (2014). Propensity score matching for causal inference with relational data. In *Proceedings of the UAI 2014 Workshop Causal Inference: Learning and Prediction*, pages 25–34.

Chen, P. P.-S. (1976). The entity-relationship model – toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379.

Getoor, L. and Taskar, B. (2007). *Introduction to statistical relational learning*. MIT press.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM.

Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5.

Maier, M. (2014). *Causal Discovery for Relational Domains: Representation, Reasoning, and Learning*. PhD thesis, University of Massachusetts Amherst.

[4] Note that the original definition of $\mathbf{AGG}_{\mathcal{M}}$ does not check *co-intersectability* and $Q.Y \leftarrow S.Z \cap S'.Z$ is granted.

Maier, M., Marazopoulou, K., Arbour, D., and Jensen, D. (2013a). A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*, pages 371–380, Bellevue, WA. AUAI Press.

Maier, M., Marazopoulou, K., and Jensen, D. (2013b). Reasoning about independence in probabilistic models of relational data. *Approaches to Causal Structure Learning Workshop, UAI 2013*.

Maier, M., Taylor, B., Oktay, H., and Jensen, D. (2010). Learning causal models of relational domains. In *Proceedings of the Twenty-Fourth National Conference on Artificial Intelligence*, pages 531–538.

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

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

Pearl, J. (2000). *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press.

Ramsey, J. and Spirtes, P. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 401–408. AUAI Press.

Sacerdote, B. (2000). Peer effects with random assignment: Results for Dartmouth roommates. Technical report, National bureau of economic research.

Shalizi, C. R. and Thomas, A. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods Research*, 40(2):211–239.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*, volume 81. MIT press.