

Relational Causal Discovery with Cycles

by

Ragib Ahsan

B.S., Bangladesh University of Engineering and Technology, 2012

M.S., University of Illinois at Chicago, 2020

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2022

Chicago, Illinois

Defense Committee:

Professor Elena Zheleva, Chair and Advisor

Professor Barbara Di Eugenio

Professor Brian Ziebart

Professor Theja Tulabandhula, Department of Information and Decision Sciences

Dr. David Arbour, Adobe Research

Copyright by

Ragib Ahsan

2022

*To my parents Hamid and Kumkum for their unconditional support and to my wife
Hridi for her love and understanding.*

ACKNOWLEDGMENTS

I would like to thank my advisor Elena Zheleva for her thoughtful guidance without which I wouldn't have made it this far. The inspiration and support I got from her were invaluable for my research journey, especially as an international student and non-native English speaker. I'm sincerely grateful to her for giving me the opportunity to work with her.

I would also like to thank my committee members — Barbara Di Eugenio, Brian Ziebart, and Theja Tulabandhula for their insightful comments and consideration throughout the dissertation process.

I would like to thank David Arbour who has been a collaborator, committee member, and above all a great mentor for me throughout the journey. I've learned a lot from him which helped me immensely to grow as a researcher. I am grateful for his continued support and mentorship.

I am grateful for my supportive and understanding qualification committee, Xinhua Zhang, Brian Ziebart, and Jon Solworth, who guided me in the early years of my Ph.D.

I am grateful to all the EDGES lab members, Zohreh Ovaisi, Shishir Adhikari, Ahmed Sayeed Faruk, and Gyeongun Lee, for their support and feedback over the years which helped me become a better researcher over the years. Special thanks to Christopher Tran and Zahra Fatemi for their continued support and fantastic sense of humor which worked as a tonic for stress release.

ACKNOWLEDGMENTS (Continued)

In addition, I would like to thank all my collaborators who allowed me to work on different and exciting projects: Chris Kanich, Shubham Singh, Mainack Mondal, Blase Ur.

I want to thank my parents and especially my elder brother, Galib for supporting me in difficult times and pushing me when I almost lost my hope. I want to thank all my friends from home and abroad for sharing quality times and helping me to become a better version of myself.

Finally, I'm grateful to my wife, Hridi for her relentless support and caring without which I would not have accomplished this. I cannot thank her enough.

CONTRIBUTION OF AUTHORS

Chapter 1 provides an introduction and motivation for this work and Chapter 2 provides notations and assumptions used in this work and a brief background on closely related works.

Chapter 3 was previously published in the Proceedings of the first Conference on Causal Learning and Representation (CLearR 22) as “Relational Causal Models with Cycles: Representation and Reasoning” (Ahsan et al., 2022a). David Arbour and my advisor, Elena Zheleva, were involved in the writing and research development process.

Chapter 4 is from a pre-print version of the paper “Learning Relational Causal Models with Cycles through Relational Acyclification”. David Arbour and my advisor, Elena Zheleva, were actively involved in the writing and research development process.

Chapter 5 was previously published in the Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 22) as “Non-parametric Inference of Relational Dependence” Ahsan et al. (2022b). David Arbour and my advisor, Elena Zheleva, were involved in the writing and research development process. Zahra Fatemi contributed to experimental evaluations.

Chapter 6 was previously published in the Proceedings of the 16th International Workshop on Mining and Learning with Graphs (MLG 2020) as “Effectiveness of Sampling Strategies for One-shot Active Learning from Relational Data” Ahsan and Zheleva (2020), and my advisor, Elena Zheleva, was involved in the writing and research development process.

CONTRIBUTION OF AUTHORS (Continued)

Chapter 7 provides a summary of this thesis and reflects on limitations of this work and discusses future research directions.

TABLE OF CONTENTS

<u>CHAPTER</u>		<u>PAGE</u>
1	INTRODUCTION	1
2	BACKGROUND	10
2.1	Graphical Causal Models	10
2.1.1	Cyclic Graphical Causal Models	12
2.1.2	Directed Graphs with Hyperedges (HEDGes)	13
2.2	Relational Causal Model (RCM)	15
2.2.1	Relational Variable	15
2.2.2	Ground Graph and Abstract Ground Graph	17
2.2.3	Relational d -separation	19
2.2.4	Counterexample of Completeness of AGG	20
2.3	Causal Discovery from Observational Data	23
2.3.1	Cyclic Causal Discovery	23
2.3.2	Local and Global Causal Structure Learning	24
2.3.3	Relational Causal Discovery	25
2.3.4	Conditional Independence Test for Relational Data	26
2.4	Active Learning for Relational Data	27
2.4.1	Relational Classification	28
2.4.2	Sample Selection from Relational Data	29
3	RELATIONAL CAUSAL MODELS WITH CYCLES: REPRESENTATION AND REASONING	31
3.1	Cyclic Relational Causal Model	32
3.1.1	Example	34
3.2	Relational σ -separation	34
3.3	σ -Abstract Ground Graph	35
3.4	Theoretical Guarantees of Relational σ -separation	36
3.4.1	Completeness of Relational d -separation Under AGG	36
3.4.2	Soundness and Completeness of σ -separation	40
3.4.3	Soundness and Completeness of σ -AGG $_{\mathcal{M}_s}$	41
3.4.4	Soundness and Completeness of relational σ -separation	45
3.4.5	Relational σ -separation Markov Condition	46
3.5	Discussion	47
4	LEARNING RELATIONAL CAUSAL MODELS WITH CYCLES	49
4.1	Relational Causal Discovery with Cycles	50
4.1.1	RCD for Cyclic Relational Causal Models	50

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>		<u>PAGE</u>
4.1.2	A Counterexample	52
4.2	Learning Cyclic Relational Causal Models	52
4.2.1	Relational Acyclification	53
4.2.2	Maximum Hop Threshold for Relational Acyclification	54
4.2.3	Soundness and Completeness of RCD for Cyclic Relational Causal Models	55
4.2.4	Identification of Relational (Non-)cycles	58
4.3	Experiments	60
4.3.1	Experimental Setup	60
4.3.2	Evaluation Criteria	61
4.3.3	Results	61
4.3.4	Demonstration on Real-world Data	64
4.4	Discussion	65
5	RELATIONAL DEPENDENCE TEST	66
5.1	Relational Dependence	66
5.1.1	Example of Relational Dependence	67
5.1.2	General Relational Independence	68
5.2	Relational Dependence Test	69
5.2.1	Non-parametric Aggregate Representations	70
5.2.2	Relational Marginal Independence Test	71
5.2.3	Relational Conditional Independence Test	72
5.2.4	Consistency of Relational Independence Test	73
5.2.4.1	Weak Dependence	74
5.2.4.2	Weak Dependence in Relational Domains	75
5.2.4.3	Relational Marginal Independence Test	76
5.2.5	Large Scale Approximations	78
5.2.6	Extension to Multi-relational Systems	79
5.3	Experimental Evaluations	82
5.3.0.1	Network Datasets	82
5.3.1	Four Cases of Relational (In)dependence	82
5.3.1.1	Synthetic Attribute Generation	83
5.3.2	Experimental Setup	85
5.3.3	Results	87
5.3.3.1	Relational Dependence Sensitivity	87
5.3.3.2	Network Sensitivity	88
5.3.3.3	Scalability	89
5.3.3.4	Diffusion	90
5.3.4	Comparison to Sobolev Independence Criterion (SIC)	92
5.4	Real-world Demonstration	93
5.5	Discussion	97

TABLE OF CONTENTS (Continued)

<u>CHAPTER</u>	<u>PAGE</u>
6 EFFECTIVENESS OF SAMPLING STRATEGIES ON RELATIONAL DATA	98
6.1 Preliminaries	100
6.1.1 Basic Notations for Relational Classification	100
6.1.2 Weisfeiler-Lehman Algorithm	101
6.2 One-shot Active Learning for Relational Classification	102
6.2.1 Problem Definition	102
6.2.2 Sampling for One-shot Active Learning	104
6.2.2.1 Network Sampling Methods	104
6.2.2.2 Non-network Sampling Methods	106
6.2.2.3 Hybrid Sampling Methods	106
6.2.2.4 Weisfeiler-Lehman Sampling	107
6.3 Experimental Evaluation	110
6.3.1 Data	110
6.3.2 Experimental Setup	111
6.3.2.1 The Hash Function for WLS	111
6.3.2.2 Relational Classifiers	112
6.3.2.3 Evaluation Methodology	112
6.3.2.4 Packages and Hardware	113
6.3.3 Results	113
6.4 Discussion	120
7 CONCLUSION	121
7.1 Limitations and Future Directions	122
CITED LITERATURE	127
APPENDICES	136
Appendix A	137
Appendix B	142

LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
I	Real-world demonstration: exploration of the dependence between the habits of students and their first-hop neighbors in 50 Women dataset	95
II	Properties of the datasets used in experimental evaluation.	110
III	Micro-F1 scores of 11 sampling methods across 4 datasets and 4 classifiers for an active learning budget of 224 nodes.	114
IV	Average ranks of sampling methods for different categories of relational classifiers over all datasets.	115

LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Example of relational model with and without a feedback loop. Rectangle, rhombus, and oval shapes represent an entity, relationship, and attributes respectively. Arrows refer to relational dependence. The solid arrows constitute an acyclic relational model. The dashed arrow creates a feedback loop in the model.	3
2	Fragments of a relational skeleton, ground graph, and abstract ground graph corresponding to the relational causal model from Figure 1. The arrows represent relational dependencies.	18
3	A fragment of the ground graph corresponding to the example model \mathcal{M} visually depicted by Lee and Honavar (2015). Only this structure satisfies the cardinality constraints of the model.	22
4	Steps of the general pipeline (LGL) of causal structure learning. The figures are borrowed from (Aliferis et al., 2010a).	24
5	RCD-learned model of MovieLens+ (Maier et al., 2013a).	26
6	A cyclic relational causal model, corresponding ground graph, and σ -AGG.	33
7	General pattern of the counterexample by (Lee and Honavar, 2015). The notations inside the square/rhombus refer to instances of the corresponding entity/relationship classes. The dashed arrows represent realizations of the relational dependencies. The dashed lines can be replaced with arbitrary length valid relational paths. The dotted line represents a hypothetical connection that can nullify the counterexample under assumption 7.	39
8	Counterexample showing RCD produces incorrect output for cyclic RCM under σ -separation.	51
9	Invalid acyclification of σ -AGG from Figure 8b	53
10	An example cyclic relational model and its corresponding DPAG output by RCD under σ -separation.	58
11	Comparison of d -RCD and σ -RCD based on the recall of <i>isPossibleAncestor</i> (top row) and <i>isPossibleCycle</i> (bottom row) queries. The number of entity types increased from left to right.	62
12	Frequency of edge orientation rules for d -RCD (top) and σ -RCD for different numbers of entity types and dependencies.	63
13	A possible cyclic relational model of MovieLens+ based on the output of RCD (Maier et al., 2013a).	64
14	Relational dependence impact on Type I/II errors.	87
15	Relational dependence impact on Type I/II errors while variance of noise varied $\sim \mathcal{N}(1, 0.2)$ over multiple trials.	89

LIST OF FIGURES (Continued)

<u>FIGURE</u>		<u>PAGE</u>
16	Impact of network parameters on Type I/II errors.	90
17	Test scalability.	91
18	Type II error for the Linear Threshold Model on Facebook and Twitter ego-network.	92
19	Type I/II errors with polynomial dependency model on synthetic networks for all three cases.	93
20	Illustration of classic Weisfeiler-Lehman algorithm: 1) same initial label to all nodes, 2) first relabeling after sorting signature strings, 3) final stable labels.	101
21	Macro-F1 scores by different sampling methods using different relational classifiers.	116
22	Top three sampling methods vs ALFNET using ICA classifier on Citeseer dataset.	119

LIST OF ABBREVIATIONS

i.i.d	Independently and Identically Distributed
CSL	Causal Structure Learning
LGL	Local-To-Global
WLS	Weisfeiler-Lehman Sampling
CI	Conditional Independence
NIRD	Non-parametric Inference of Relational Dependence.
AGG	Abstract Ground Graph

SUMMARY

Most of the tools and methods developed for causal discovery rely on a graphical representation based on Bayesian networks which assumes independent and identically distributed (i.i.d) instances. Probabilistic relational models have been developed to relax this assumption. The key advantage of these relational models is that they can represent systems involving multiple types of entities interacting with each other with some probabilistic dependence. Causal reasoning over such relational systems is key to understanding many real-world phenomena, such as social influence. Influence in complex dynamic systems is often mutual and represented by a feedback loop or cycle in the relational model. Identifying mutual influence in relational models is of great interest in the research community. For example, social scientists and marketing experts are interested in studying the social dynamics between people and products in social networks. However, there is a lack of available methods for discovering mutual influence or cycles in complex relational systems. Most of the works on causal structure learning assume that the observational data samples are identically and independently distributed (i.i.d) and as a result are not directly applicable to relational models. At the same time, existing causal discovery algorithms for relational causal models assume acyclicity and rely on prior domain knowledge for conditional independence tests.

The primary objective of this thesis is to address the deficiencies of existing relational causal discovery approaches by developing tools and methods for practical application of causal discovery on complex relational systems with feedback loops or cycles. In my thesis, I develop

SUMMARY (Continued)

sigma-abstract ground graph, a sound and complete representation for cyclic relational causal models which can capture conditional independence relationships consistent across all possible instantiations of the model by an operator called relational sigma-separation. Based on this new representation and theoretical guarantees, I define relational acyclification- an important property of cyclic relational causal models that helps identifying such models from observational data. I prove that under certain assumptions and conditions, the existing relational causal discovery algorithm (RCD) is sound and complete for cyclic relational causal models. Experimental evaluations conducted on publicly available datasets and synthetically generated datasets support my theoretical contributions. The state-of-the-art conditional independence (CI) test for relational data depends on domain knowledge about the type of dependence and lacks convergence guarantees which makes them unsuitable for large-scale real-world applications. In my thesis, I formalize a general notion of relational dependence and develop a consistent CI test method based on kernel mean embeddings that can capture complex dependence functions over node neighborhoods. Causal structure learning is NP-hard and constraint-based algorithms reduce this complexity by learning the local structure of each variable first. I conducted a comprehensive study of sampling strategies for relational classification which can help identify the important variables in local structures. I introduced a novel sampling method based on Weisfeiler-Lehman isomorphism that provides competitive predictive accuracy for one-shot active learning in relational classification.

CHAPTER 1

INTRODUCTION

Recent advances in machine learning have introduced numerous tools and methods for informed decision-making from observational data. The utility of such methods is often limited because of their associational nature and failure to capture causal knowledge. Researchers in practice often look for causal answers through randomized controlled trials (RCT) which can be costly or involve unethical manipulation (Pearl, 2009). For example, one cannot force a person to smoke to observe how their social connections get influenced by the action. As a result, causal inference from observational data is much desired and often the only feasible solution for practitioners. The goal of causal structure learning (CSL) is to estimate the causal model of the true data generating process which is typically represented with graphical models (Pearl, 2009). The discovery of such graphical models from observational samples enables causal effect identification and estimation through the do-calculus (Pearl, 2009; Tucci, 2013).

Most of the tools and methods developed for causal discovery rely on a graphical representation based on Bayesian networks which assume independent and identically distributed (i.i.d) instances. Probabilistic relational models (Getoor et al., 2007) have been developed that relax this assumption. The key advantage of these relational models is that they can represent systems involving multiple types of entities interacting with each other with some probabilistic dependence. Causal reasoning over such relational systems is key to understanding many real-world phenomena, such as social influence. Influence in complex dynamic systems is often

mutual and represented by a feedback loop or cycle in the relational model. For example, the study by Christakis and Fowler (2007) shows the challenges in estimating the contribution of environmental factors and peer effect in the spread of obesity in social network where a peer effect can be represented as a feedback loop between an ego (focal node) and its alters (direct ties) in the network. Intuitively, the true causal structure is acyclic over time since a cause always precedes its effect— friends’ influence on a person followed by their influence on friends. Even though the granular cause and effect happen in minutes or hours, measures of these effects are typically cumulative averages over a much longer time period. For example, in biological systems, interactions occur in fractions of seconds whereas the measurements are typically taken in minutes or hours. In such cases, a cyclic representation provides a natural way of reasoning about the causes and effects. Effective methods for identifying and estimating such phenomena can have implications for clinical and public health interventions. For this reason, social scientists and marketing experts are interested to study the social dynamics between people and products in social networks (Bakshy et al., 2011, 2015; Ogburn et al., 2020). However, there is a lack of available methods for discovering mutual influence or cycles in complex relational systems.

The development of *relational causal models*, which generalize over structural causal models, is an important step towards capturing interactions between non-i.i.d instances (Maier et al., 2013b,a; Lee and Honavar, 2015; Bhattacharya et al., 2020). Relational models involve multiple types of interacting entities with probabilistic dependencies among their attributes. Maier et al. (2013b) developed a lifted causal representation named *abstract ground graph (AGG)* that

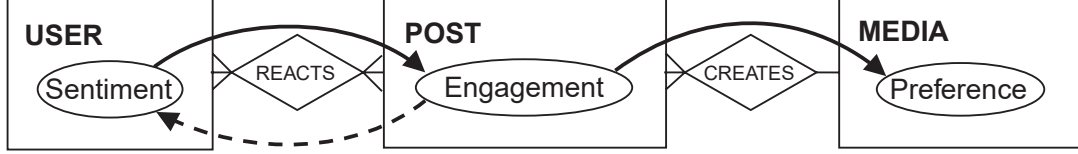


Figure 1: Example of relational model with and without a feedback loop. Rectangle, rhombus, and oval shapes represent an entity, relationship, and attributes respectively. Arrows refer to relational dependence. The solid arrows constitute an acyclic relational model. The dashed arrow creates a feedback loop in the model.

abstracts over all instantiations of a relational model. In addition, they introduced relational d -separation criteria to answer relational queries through AGG. These tools and methods enable relational causal discovery from observational data (Maier et al., 2013a). However, existing studies in relational causal models assume acyclicity and do not allow for reasoning about identification in the presence of feedback loops.

Figure 1 shows an example of a relational model where users react to news articles or posts on a specific topic (e.g., vaccines) generated by media agencies. The preference of the media regarding a topic (e.g., pro- vs anti-vaccination) is influenced by the engagement or feedback (e.g., positive/negative comments) it receives on its existing posts. The sentiment of a user towards a given post directly impacts their engagement in the post. The relational model representation tempts one to conclude that the sentiment of users regarding vaccination is

independent of the preference of media agencies given the engagements of the posts the users react to. However, as I show in section 2.2.2, this is not necessarily true.

One of the fundamental challenges in learning **cyclic relational causal models** is the lack of a theoretically sound abstract representation that can capture conditional independence relationships consistent across all possible realizations of the given model. In Figure 1, users' sentiment can also be impacted by the engagements in posts they interact with. The dashed arrow in Figure 1 represents such a dependency which makes the model a cyclic one. Unfortunately, even though the model seems simple and realistic, the abstract ground graph (AGG) representation and relational d -separation no longer apply in the general setting due to the presence of a feedback loop (i.e., a cycle). I propose relevant representations in Chapter 3.

To the best of my knowledge, no existing study focuses on discovering relational causal models with cycles from observational data. The closest works on cyclic causal discovery assume i.i.d data or do not consider relational data. Richardson (1996) develop a cyclic causal discovery (CCD) algorithm which is shown to be sound but not complete. In recent work, Mooij and Claassen (2020) provides necessary conditions for constraint-based causal discovery algorithms developed for acyclic causal models, such as PC (Pearl et al., 2000) and FCI (Spirtes et al., 2000), to be sound and complete for cyclic causal models under σ -separation criteria. There are several other algorithms for cyclic causal discovery from i.i.d samples (Rothenhäusler et al., 2015; Strobl, 2019a) but no such algorithm exists for cyclic relational causal models. Sound and complete algorithms have been proposed for learning relational causal models from observational data (Maier et al., 2013a; Lee and Honavar, 2016a,b). However, they assume

acyclicity and thus cannot reason about mutual influence or cycles. A big obstacle in adopting these existing discovery methods is reasoning about the equivalence class and identifiability of cycles in relational causal models. I address this obstacle in Chapter 4.

Apart from theoretical aspects, relational causal discovery faces challenges for application in real-world scenarios. One of the building blocks of constraint-based structure learning algorithms is conditional independence tests. There are some conditional independence tests proposed over the years but they are mostly intended for i.i.d propositional variables (Gretton et al., 2005; Strobl et al., 2019; Zhang et al., 2011). The few CI test methods targeted for non-i.i.d data make strong assumptions about the nature of dependence among the interconnected entities (Flaxman et al., 2015; Lee and Honavar, 2017). Flaxman et al. (2015) develop a test between propositional variables accounting for latent homophily in a grid network but do not consider relational variables. The state-of-the-art method for relational conditional independence test, KRCIT, considers a flattened representation of the relational data and eventually utilizes an existing CI test designed for i.i.d data. However, their approach has three key limitations. First, it requires practitioners to make explicit assumptions about the data generating processes and to specify an aggregation function over the relational variable a priori. Second, existing tests rely on propositionalization, which refers to the process of projecting connected data to a single, propositional table, which raises statistical concerns (Maier et al., 2013c). Third, it is computationally expensive and inapplicable to large relational datasets. Because of these limitations of the state-of-the-art relational conditional independence methods, the application of relational causal discovery faces two major challenges. First, it’s not generalizable

and requires prior domain knowledge, and second, it's not scalable to the size of real-world networks. I address these limitations in Chapter 5.

Real-world relational data are often quite large and any discovery or prediction tasks on it suffers from it. A natural solution for this is to consider a sample from the original data. However, sampling from relational data is not a trivial task since it needs to take the relational structure in consideration. There are several sampling methods proposed for relational data over the years but there is a lack of understanding about the effectiveness of these sampling methods. The closest works on such a comparative study relies on relational classification tasks for the comparison of sampling methods Berton et al. (2016); Ahmed et al. (2013). These studies are relatively old and doesn't consider modern deep-learning-based relational classification methods. In this study, I conduct a comprehensive empirical evaluation of existing sampling methods for relational data and proposed a new sampling technique based on latent structural properties of the data. The results and discussions on the empirical study in provided in Chapter 6.

The goal of this thesis is to develop theoretically sound methods and necessary tools to enable relational causal discovery with cycles from real-world observational samples. Here, I summarize the contributions of each chapter:

I Relational Causal Models with Cycles: Representation and Reasoning. In this work, I focus on the representation of and reasoning about cyclic RCMs. I define a new abstract representation, σ -abstract ground graph (σ -AGG) Ahsan et al. (2022a) which generalizes over cyclic relational models. In order to reason about relational queries in σ -AGG, I introduce *relational σ -separation* and provide proof for its soundness and com-

pleteness for all instantiations of a relational model. I show the sufficient conditions for the completeness of σ -AGG by first resolving an open problem on the completeness of AGG. Finally, I discuss the Markov condition of relational σ -separation and its implications. This work lays the foundation for developing algorithms to discover cyclic relational causal models from observational data.

II Learning Relational Causal Models with Cycles. Based on the development of necessary representation and reasoning methods for cyclic relational causal models, I focus on causal discovery from observational samples. In order to reason about the equivalence class and identifiability of cycles in relational models, I introduce *relational acyclification*, an operation that helps to reason over the scope of cyclic relational models which are identifiable with constraint-based causal discovery algorithms. I characterize the necessary conditions for the existence of valid relational acyclification. Following this criterion, I show that RCD (Maier et al., 2013a), a pioneering relational causal discovery algorithm for acyclic relational models, is sound and complete for cyclic relational models under σ -separation and causal sufficiency assumption. I provide experimental results on synthetic relational models in support of my claims. I also demonstrate the effectiveness of the algorithm on a real-world dataset.

III Nonparametric Inference of Relational Dependence. In this thesis, I focus on developing tools and techniques to make relational causal discovery methods applicable to large-scale real-world use cases. I develop a general definition for relational dependence and an accompanying statistical test, NIRD Ahsan et al. (2022b) for marginal and

conditional relational dependence that is able to capture a family of aggregate functions for characterizing relational dependence. Specifically, I propose using kernel mean embeddings as aggregations that lend themselves to a non-parametric inference of relational dependence with standard statistical tests. Since kernel tests can be notoriously slow and impractical, I introduce an approximation to my test which makes it scalable to larger networks. I evaluate the proposed method by comparing it to KRCIT on a variety of synthetic and semi-synthetic networks and simulate several social network characteristics, such as structure, density, and size.

IV Effectiveness of Sampling Strategies on Relational Data. I conduct a comprehensive study of sampling strategies for relational classification in a one-shot active learning setup over four real-world datasets and four state-of-the-art relational classifiers which I describe in Chapter 6. I consider both graph sampling algorithms as well as sampling strategies specifically designed for semi-supervised node classification. I also propose a sampling approach based on the Weisfeiler-Lehman algorithm which shows promising results in the empirical evaluation. My proposed sampling method, Weisfeiler-Lehman Sampling (WLS) Ahsan and Zheleva (2020) relies solely on the structural role of nodes for label acquisition decisions. One of its main advantages is that it is computationally efficient and yet harnesses structural information effectively. My empirical evaluation shows that even though there isn't one sampling method that performs the best consistently across datasets and classifiers, Weisfeiler-Lehman ranks the highest on average.

The rest of the thesis is structured as follows: Chapter 2 provides the necessary background on relational causal models and causal discovery in general. Chapter 3 introduces σ -AGG, the proposed abstract representation of cyclic relational causal models, and a theoretically sound operator called relational σ -separation to enable answering relational queries on cyclic relational models. Chapter 4 examines the necessary and sufficient conditions for causal discovery of cyclic relational causal models and establishes the effectiveness of the RCD algorithm (Maier et al., 2013a) for relational causal discovery with cycles. Chapter 5 formalizes the proposed relational dependence and a corresponding non-parametric CI test. Chapter 6 describes the comprehensive study of sampling strategies for one-shot active learning on relational data. Chapter 7 provides a discussion on the overall contribution of the thesis and possible future directions in the study of relational causal discovery with cycles.

CHAPTER 2

BACKGROUND

Discovering cyclic relational causal models largely depends on two categories of existing studies: 1) cyclic causal discovery for propositional models, and 2) relational causal discovery with acyclicity assumption. Therefore it is important to provide necessary background on both these lines of works. I denote random variables with uppercase letters, realizations of random variables with lower case, and bold to denote sets.

2.1 Graphical Causal Models

A structural causal model (SCM) consists of sets of random variables and a set of structural equations that describes how values are assigned to the random variables (Pearl, 2009). A directed graph $G = (\mathbf{V}, \mathbf{E})$ is used to represent the SCM where the set of nodes \mathbf{V} correspond to the random variables and edges \mathbf{E} corresponds to the structural equations. If there is an edge $X \rightarrow Y$, we say that X is a parent of Y and Y is a child of X . The set of parents for some node X is denoted by $parents(X)$. A walk between two nodes $u, v \in \mathbf{V}$ is a tuple $\langle v_0, e_1, v_1, e_2, v_2, \dots, e_n, v_n \rangle$ of alternating nodes and edges in $G(n \geq 0)$, such that $v_0, \dots, v_n \in \mathbf{V}$, and $e_1, \dots, e_n \in \mathbf{E}$, starting with node $v_0 = u$ and ending with node $v_n = v$ where the edge e_k connects the two nodes v_{k-1} and $v_k \in \mathbf{V}$ for all $k = 1, \dots, n$. If the walk contains each node at most once, it is called a *path*. A *directed walk (path)* from $v_i \in \mathbf{V}$ to $v_j \in \mathbf{V}$ is a walk (path) between v_i and v_j such that every edge e_k on the walk (path) is of the form

$v_{k-1} \rightarrow v_k$, i.e., every edge is directed and points away from v_i . We get the *ancestors* of node v_j by repeatedly following the path(s) through the parents: $AN_G(v_j) := \{v_i \in \mathbf{V} : v_i = v_0 \rightarrow v_1 \rightarrow \dots \rightarrow v_n = v_j \in G\}$. Similarly, we define the *descendants* of v_i : $DE_G(v_i) := \{v_j \in \mathbf{V} : v_i = v_0 \rightarrow v_1 \rightarrow \dots \rightarrow v_n = v_j \in G\}$. Each node is an ancestor and descendant of itself. A directed cycle is a directed path from v_i to v_j such that in addition, $v_j \rightarrow v_i \in \mathbf{E}$. All nodes on directed cycles passing through $v_i \in \mathbf{V}$ together form the strongly connected component $SC_G(v_i) := AN_G(v_i) \cap DE_G(v_i)$ of v_i .

The most common graphical representation for causal models is *directed acyclic graphs* (DAGs) which doesn't allow any cycles or feedback loops. A fundamental notion in DAGs is the *d-separation* criteria Pearl (1988):

Definition 1 (*d-separation*). A walk $\langle v_0 \dots v_n \rangle$ in DCG $G = \langle \mathbf{V}, \mathbf{E} \rangle$ is *d-blocked* by $\mathbf{C} \subseteq \mathbf{V}$ if:

1. its first node $v_0 \in \mathbf{C}$ or its last node $v_n \in \mathbf{C}$, or
2. it contains a collider $v_k \notin AN_G(\mathbf{C})$, or
3. it contains a non-collider $v_k \in \mathbf{C}$.

If all paths in \mathcal{G} between any node in set $\mathbf{A} \subseteq \mathbf{V}$ and any node in set $\mathbf{B} \subseteq \mathbf{V}$ are *d-blocked* by a set $\mathbf{C} \subseteq \mathbf{V}$, we say that \mathbf{A} is *d-separated* from \mathbf{B} by \mathbf{C} , and we write $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}}^d \mathbf{B} | \mathbf{C}$.

d-separation exhibits the Markov property in DAGs which states that:

Definition 2 (Causal Markov property). Given a graphical model $G = (\mathbf{V}, \mathbf{E})$, if two variables $X \in \mathbf{V}$ and $Y \in \mathbf{V}$ are *d-separated* given another variable $Z \in \mathbf{V}$ in a DAG representation

then X and Y are conditionally independent given Z in the corresponding distribution of the variables.

2.1.1 Cyclic Graphical Causal Models

DAGs provide ways for natural causal interpretation but cannot reason about cyclic causal models. A more general class of graphs are *directed cyclic graphs (DCGs)* which drop the assumption of acyclicity (and allow feedback loops). These graphs are appropriate for (possibly cyclic) structural causal models (SCMs) where the corresponding Markov properties and causal interpretation are more subtle (Bongers et al., 2021). Cyclic SCMs are useful to represent causal semantics of equilibrium states in dynamical systems (Bongers et al., 2021).

Directed cyclic graphs offer certain properties that help model cyclic causal models. Given a directed cyclic graph $G = (\mathbf{V}, \mathbf{E})$, all nodes on directed cycles passing through node $i \in \mathbf{V}$ together form the strongly connected component $SC_G(i) = AN_G(i) \cap DE_G(i)$ of i where $AN_G(i)$ and $DE_G(i)$ refers to the ancestors and descendants of node $i \in \mathbf{V}$. The set of conditional independence entailed in DCG, G is referred to as independence model $IM(G)$.

Unlike DAGs, DCGs are not guaranteed to satisfy the Markov property in a general case under d -separation. Instead, a different notion of separation, called σ -separation satisfies the Markov property of DCGs (Forré and Mooij, 2017). The σ -separation criterion is very similar to the d -separation criterion where the main difference is σ -separation has as an additional condition for a non-collider to block a path that it has to point to a node in a different strongly connected component (Mooij and Claassen, 2020).

Definition 3 (σ -separation). (Forré and Mooij, 2017)

A walk $\langle v_0 \dots v_n \rangle$ in DCG $G = \langle \mathbf{V}, \mathbf{E} \rangle$ is σ -blocked by $\mathbf{C} \subseteq \mathbf{V}$ if:

1. its first node $v_0 \in \mathbf{C}$ or its last node $v_n \in \mathbf{C}$, or
2. it contains a collider $v_k \notin \text{AN}_{\mathcal{G}}(\mathbf{C})$, or
3. it contains a non-collider $v_k \in \mathbf{C}$ that points to a node on the walk in another strongly connected component (i.e., $v_{k-1} \rightarrow v_k \rightarrow v_{k+1}$ with $v_{k+1} \notin \text{SC}_{\mathcal{G}}(v_k)$, $v_{k-1} \leftarrow v_k \leftarrow v_{k+1}$ with $v_{k-1} \notin \text{SC}_{\mathcal{G}}(v_k)$ or $v_{k-1} \leftarrow v_k \rightarrow v_{k+1}$ with $v_{k-1} \notin \text{SC}_{\mathcal{G}}(v_k)$ or $v_{k+1} \notin \text{SC}_{\mathcal{G}}(v_k)$).

If all paths in \mathcal{G} between any node in set $\mathbf{A} \subseteq \mathbf{V}$ and any node in set $\mathbf{B} \subseteq \mathbf{V}$ are σ -blocked by a set $\mathbf{C} \subseteq \mathbf{V}$, we say that \mathbf{A} is σ -separated from \mathbf{B} by \mathbf{C} , and we write $\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}}^{\sigma} \mathbf{B} | \mathbf{C}$.

σ -faithfulness refers to the property which states that all statistical dependencies found in the distribution generated by a given causal model are entailed by the σ -separation relationships.

Richardson (1996) show that a class of graphs called Partial Ancestral Graphs (PAG) is a sufficient representation for the equivalence class of cyclic causal models represented by DCGs. PAGs have also been shown to be a sufficient representation for causal discovery with cycles and unobserved confounders (Mooij and Claassen, 2020). Since we are assuming no selection bias for simplicity, we will only discuss directed PAGs (DPAG) in this study.

2.1.2 Directed Graphs with Hyperedges (HEDGes)

Forré and Mooij (2017) introduced the concept of directed graphs with hyperedges (HEDGes). A HEDG is a tuple $G = (\mathcal{V}, \mathcal{E}, \mathcal{H})$, where $(\mathcal{V}, \mathcal{E})$ is a directed graph (with or without cycles) and \mathcal{H} a simplicial complex over the set of vertices \mathcal{V} of \mathcal{G} . A simplicial complex \mathcal{H} over \mathcal{V} is a set of subsets of \mathcal{V} such that: 1) all single element sets $\{v\}$ are in \mathcal{H} for $v \in \mathcal{V}$, and 2) if $F \in \mathcal{H}$ then also all subsets $F' \subseteq F$ are elements of \mathcal{H} .

The general directed global Markov property (gdGMP) for the HEDGes is stated as follows:

Definition 4 (gdGMP Forré and Mooij (2017)). *For all subsets $X, Y, Z \subseteq V$ we have the implication:*

$$X \underset{G}{\underset{\sigma}{\parallel}} Y | Z \implies X \underset{P_v}{\parallel} Y | Z$$

Forré and Mooij (2017) introduced an operation called *acyclification* for directed cyclic graphs that generates DAGs with equivalent independence models as the given DCG. It allows a single DPAG to represent the ancestral relationship of a DCG G and all its acyclifications G' .

Definition 5 (Acyclification (Forré and Mooij, 2017)). *Given a DCG $G = (\mathcal{V}, \mathcal{E})$, an acyclification of G is a DAG $G' = (\mathcal{V}, \mathcal{E}')$ with*

- i the same nodes \mathcal{V} ;*
- ii for any pair of nodes i, j such that $i \notin SC_G(j)$: $i \rightarrow j \in \mathcal{E}'$ iff there exists a node k such that $k \in SC_G(j)$ and $i \rightarrow k \in \mathcal{E}$;*
- iii for any pair of distinct nodes i, j such that $i \in SC_G(j)$: $i \rightarrow j \in \mathcal{E}'$ or $i \leftarrow j \in \mathcal{E}'$;*

Proposition 1 ((Mooij and Claassen, 2020)). *For any DCG G and any acyclification G' of G , $IM_\sigma(G) = IM_\sigma(G') = IM_d(G')$ where $IM_\sigma(G)$ and $IM_d(G)$ refers to the independence model of the given DCG G under σ -separation and d -separation respectively.*

2.2 Relational Causal Model (RCM)

We adopt the definition of relational causal model used by previous work on relational causal discovery (Maier et al., 2013a; Lee and Honavar, 2020). We use a simplified Entity-Relationship model to describe relational data following previous work (Heckerman et al., 2007). A relational schema $\mathcal{S} = \langle \mathcal{E}, \mathcal{R}, \mathcal{A}, \text{card} \rangle$ represents a relational domain where \mathcal{E} , \mathcal{R} and \mathcal{A} refer to the set of entity, relationship and attribute classes respectively. A cardinality constraint on a relationship limits how many times an entity instance is involved in a relationship. Figure 1 shows an example relational model that describes a simplified user-media engagement system. The model consists of three entity classes (User, Post, and Media), and two relationship classes (Reacts and Creates). Each entity class has a single attribute. The cardinality constraints are shown with crow’s feet notation— a user can react to multiple posts, multiple users can react to a post, a post can be created by only a single media entity.

2.2.1 Relational Variable

We follow prior work to define a relational variable as a set of random variables (Lee and Honavar, 2017; Maier et al., 2013a).

Definition 6 (Relational Variable). *Given a relational schema $\mathcal{S} = \langle \mathcal{E}, \mathcal{R}, \mathcal{A} \rangle$, its instantiation G and a path predicate ρ , a relational variable $\sigma(v_i, \mathbf{X}, G, \rho)$ is the set of attributes $v_j.X$ selected by ρ of nodes $v_j \in \mathbf{V}$ reachable from $v_i \in \mathbf{V}$ such that $\mathbf{X} \subset \mathcal{A}$, where the path predicate ρ is a function given by:*

$$\rho(v_i, \mathbf{X}, G) : \mathbf{V} \mapsto 2^{\{v_j.X | v_j \in \mathbf{V}\}}.$$

An example path predicate could be $\rho(v_i, X, G) = \{v_j.X | v_j \in \hat{\mathcal{N}}(v_i)\}$ where $\hat{\mathcal{N}}(v_i)$ refers to the direct neighbors of v_i in G . Considering this path predicate we can simplify the notation for the relational variable corresponding to an attribute X by just $\sigma_X(v_i)$. Note that $\sigma_X(v_i)$ can represent a *propositional* variable as a special case. For example, $\sigma_X(v_i) = \{v_i.X\}$ refers to the X attribute of node v_i itself.

We will assume the following for the rest of the thesis:

A 1. *Each node $v \in \mathbf{V}$ has degree of at least 1.*

A 2. *The network structure is fixed and doesn't change during the generation of the observed random variables.*

A *relational skeleton* s is an instantiation of a relational schema \mathcal{S} , represented by an undirected graph of entities and relationships. Figure 2a shows an example skeleton of the relational model from Figure 1. It shows that Alice and Bob both react to post P1. Alice also reacts to post P2. P1 and P2 both are created by media M1. There could be infinitely many possible skeletons for a given RCM. We denote the set of all skeletons for schema \mathcal{S} as $\sum_{\mathcal{S}}$.

Given a relational schema, we can specify relational paths, which intuitively correspond to ways of traversing the schema. For the schema shown in Figure 1, possible paths include $[User, Reacts, Post]$ (the posts a user reacts to), as well as $[User, Reacts, Post, Reacts, User]$ (other users who react to the same post). *Relational variables* consist of a relational path and an attribute. For example, the relational variable $[User, Reacts, Post].Engagement$ corresponds to the overall engagements of the post that a user reacts to. The first item (i.e. $User$) in the relational path corresponds to the *perspective* of the relational variable. A terminal set,

$P|_{i_k}$ is the terminal item on the relational path $P = [I_j, \dots, I_k]$ consisting of instances of class $I_k \in \mathcal{E} \cup \mathcal{R}$.

A relational causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{D} \rangle$, is a collection of relational dependencies defined over schema \mathcal{S} . *Relational dependencies* consist of two relational variables, cause and effect. As an example, consider the following relational dependency $[Post, Reacts, User].Sentiment \rightarrow [Post].Engagement$ which states that the engagement of a post is affected by the sentiment of users who react on that post. In Figure 1, the arrows represents relational dependencies. Note that, all causal dependencies are defined with respect to a specific perspective. The last example was from the perspective of Posts.

2.2.2 Ground Graph and Abstract Ground Graph

A realization of a relational model \mathcal{M} with a relational skeleton is referred to as the *ground graph* $GG_{\mathcal{M}}$. It is a directed graph consisting attributes of entities in the skeleton as nodes and relational dependencies among them as edges. A single relational model is actually a template for a set of possible ground graphs based on the given schema. A ground graph has the same semantic as a graphical model. Given a relational model \mathcal{M} and a relational skeleton s , we can construct a ground graph $GG_{\mathcal{M}_s}$ by applying the relational dependencies as specified in the model to the specific instances of the relational skeleton. 2b shows the ground graph for the relational model from Figure 1. The relational dependencies present in the given RCM may tempt one to conclude a conditional independence statement: $[User].Sentiment \perp\!\!\!\perp [Media].Preference | [Post].Engagement$. However, when the model is unrolled in a ground graph we see the corresponding statement is not true (i.e.

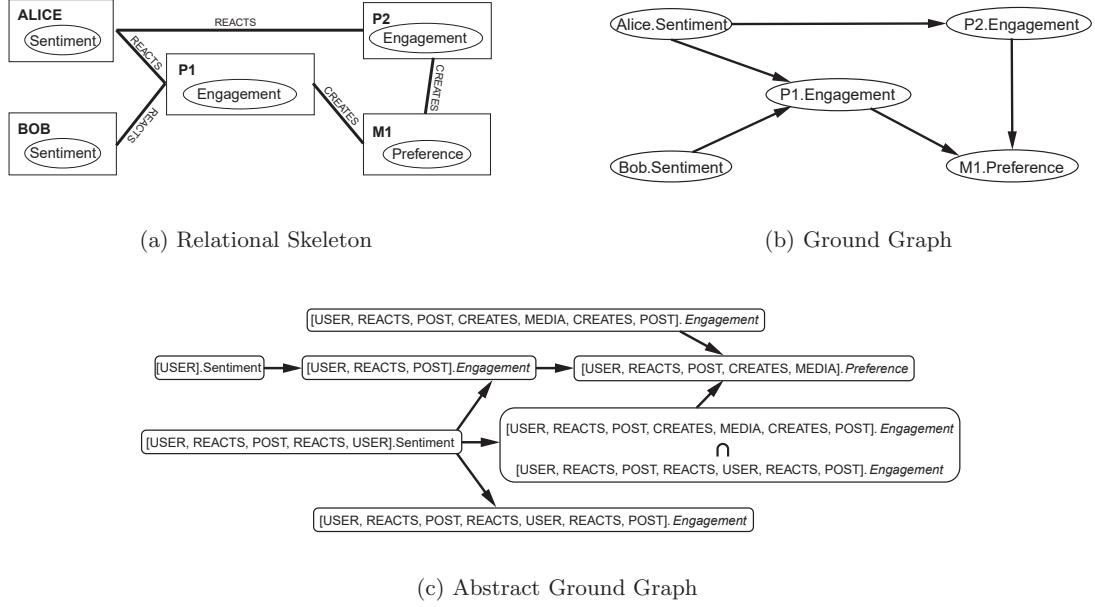


Figure 2: Fragments of a relational skeleton, ground graph, and abstract ground graph corresponding to the relational causal model from Figure 1. The arrows represent relational dependencies.

$[Bob].Sentiment \not\models [M1].Preference \mid [P1].Engagement$ since there is an alternative path through $[Alice].Sentiment$ and $[P2].Engagement$ which is activated when conditioned on $[P1].Engagement$. This shows why generalization over all possible ground graphs is hard.

An *abstract ground graph* (AGG) is an abstract representation that solves the problem of generalization by capturing the consistent dependencies in all possible ground graphs and representing them as a directed graph. AGGs are defined for a specific perspective and *hop*

threshold, h . Hop threshold refers to the maximum length of the relational paths allowed in a specific AGG. There are two types of nodes in AGG, relational variables and intersection variables. Intersection variables are constructed from pairs of relational variables with non-empty intersections (Maier et al., 2013b). For example, $[User, Reacts, Post]$ refers to the set of posts a user reacts to whereas $[User, Reacts, Post, Reacts, User, Reacts, Post]$ refers to the set of other posts reacted by other users who also reacted to the same post as the given user. These two sets of posts can overlap which is reflected by the corresponding intersection variable. Edges between a pair of nodes of AGG exist if the instantiations of those constituting relational variables contain a dependent pair in all ground graphs. We define \overline{W} as the set of nodes augmented with their corresponding intersection nodes for the set of relational variables \overline{W} : $\overline{W} = W \cup \bigcup_{W' \in W} \{W \cap W' | W \cap W' \text{ is an intersection node in } AGG_{\mathcal{M}_s}\}$. Figure 2c presents the AGG from the perspective of $User$ and with $h = 6$ corresponding to the model from Figure 1. The AGG shows that the sentiment of a user is no longer independent of media preference given just engagements of the corresponding posts the user reacts to. We also need to condition on the sentiment of other users who reacted to the same post.

2.2.3 Relational d-separation

Relational model describes a template for many possible instantiations of a relational schema. In order to reason about conditional independence facts entailed in all instances of a given relational template, Maier et al. (2013b) develop a relational counterpart for d -separation criteria. Two sets of relational variables \mathbf{X} and \mathbf{Y} from a given perspective are said to be d -separated by another set \mathbf{Z} if and only if the terminal sets of \mathbf{X} and \mathbf{Y} are d -separated by

the terminal set of \mathbf{Z} from the given perspective in all possible ground graphs of the given model. Maier et al. (2013b) introduce AGG as a means to reason about relational d -separation queries from a given perspective. The soundness and completeness of relational d -separation for AGG relies on the following assumptions:

A 3. *The relational model is acyclic.*

A 4. *There are no unobserved confounders in the relational model.*

Here, soundness refers to the fact that any d -separation relationship found in AGG implies corresponding d -separation relationship in all ground graphs it represents whereas completeness claims that the d -separation facts that hold across all ground graphs are also entailed by d -separation on the AGG. The soundness of relational d -separation under AGG is already proved by Maier et al. (2013b). However, the conditions under which completeness holds have been an open question since (Lee and Honavar, 2015) show that the initial formulation of Maier et al. (2013b) is not complete. We resolve this question in Section 3.4 and show that AGG is also complete under certain realistic assumptions.

2.2.4 Counterexample of Completeness of AGG

Maier et al. (2013a) show that AGG is sound and complete for relational d -separation for general relational causal models. However, Lee and Honavar (2015) point out the following counterexample for which AGG is not complete for relational d -separation.

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

$$\mathcal{D} = \{D_1.X \rightarrow [E_2].Y, D_2.Z \rightarrow [E_2].Y\}$$

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 $P.X, Q.Y, S.Z, S'.Z$ be four relational variables of the same perspective $B = E_1$ where their relational paths are distinct and

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

Figure 3 shows a possible realization of the given example relational causal model \mathcal{M} . The square and rhombus shapes correspond to the entities and relationships, respectively, and the dashed arrows correspond to the relational dependencies. Given the above example, Lee and Honavar (2015) make two claims. The first one says,

Claim 1. $(\overline{P.X} \not\perp \overline{S'.Z} | \overline{Q.Y})_{AGG_{\mathcal{M}}}$.

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

Claim 2. There is no $s \in \sum_{\mathcal{S}}$ and $b \in s(B)$ such that $(P.X|_b \not\perp S'.Z|_b | Q.Y|_b)_{GG_{\mathcal{M}_s}}$.

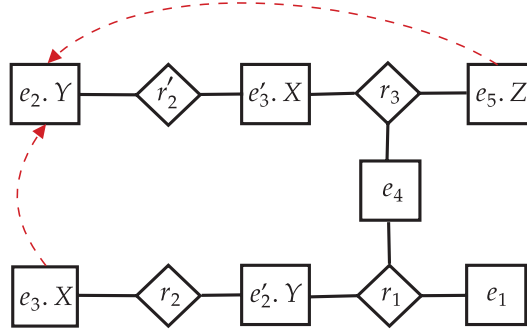


Figure 3: A fragment of the ground graph corresponding to the example model \mathcal{M} visually depicted by Lee and Honavar (2015). Only this structure satisfies the cardinality constraints of the model.

The idea of the proof is simple. Since every terminal set for P , Q , and S' given the base must not be empty and since every cardinality is ONE, terminal sets must be singletons. Let $\{e_3.x\} = P.X|_b$, $\{e_2.y\} = Q.Y|_b$, and $\{e_5.z\} = S'.Y|_b$. However, due to cardinality constraints (i.e., one), there exists only one possible structure (see Figure 3) where $e_3.x$ and $e_5.z$ are the cause of $e_2.y$ while satisfying all previously mentioned conditions except $\{e_5.z\} = S'.Y|_b$. The constraint $\{e_5.z\} = S'.Y|_b$ violates with the set of the rest of conditions. Hence, there exists no such skeleton and base (Lee and Honavar, 2015). In section 3.4, I formulate a general pattern for developing such counterexamples and show that under a certain assumption on the relational skeleton, no such counterexamples exist and thus the completeness of AGG holds.

2.3 Causal Discovery from Observational Data

Present CSL algorithms can be roughly categorized into three groups: 1) constraint-based (Spirtes et al., 2000; Colombo et al., 2012) 2) score-based (Chickering, 2002; Nandy et al., 2018; Hauser and Bühlmann, 2012) 3) hybrid (Tsamardinos et al., 2006; Ogarrio et al., 2016). Heinze-Deml et al. (Heinze-Deml et al., 2018) provides a comprehensive overview of the state-of-the-art CSL algorithms (Heinze-Deml et al., 2018). Constraint-based CSL algorithms conduct a series of CI tests to estimate a skeleton graph and orient the edges of the skeleton graph to infer the causal structure (Spirtes et al., 2000). Score-based methods typically score candidate structures using a penalized likelihood score. Since the search space to find the structure with optimal score is too large, score-based methods utilize a greedy search technique (Chickering, 2002). Hybrid methods combine the idea of both constraint-based and score-based methods. Typically a hybrid method generates the skeleton using a constraint-based approach and then performs score-based search for optimal edge orientation (Tsamardinos et al., 2006).

2.3.1 Cyclic Causal Discovery

Mooij and Claassen (2020) provide the necessary conditions under which constraint-based causal discovery algorithms for acyclic causal models, such as PC (Pearl et al., 2000) and FCI (Spirtes et al., 2000), are sound and complete in the presence of cycles under σ -separation. Their result depends on the following assumptions:

A 5. *The underlying causal model is σ -faithful.*

A 6. *There exists one or more valid acyclifications of the given causal model which contains the same set of ancestral relationships as the given model.*

Corollary 1 ((Mooij and Claassen, 2020)). *The PC algorithm with Meek’s orientation rules is sound, arrowhead-complete, tail-complete, and Markov complete (in the σ -separation setting without selection bias) for directed cyclic graphs. (Mooij and Claassen, 2020)*

2.3.2 Local and Global Causal Structure Learning

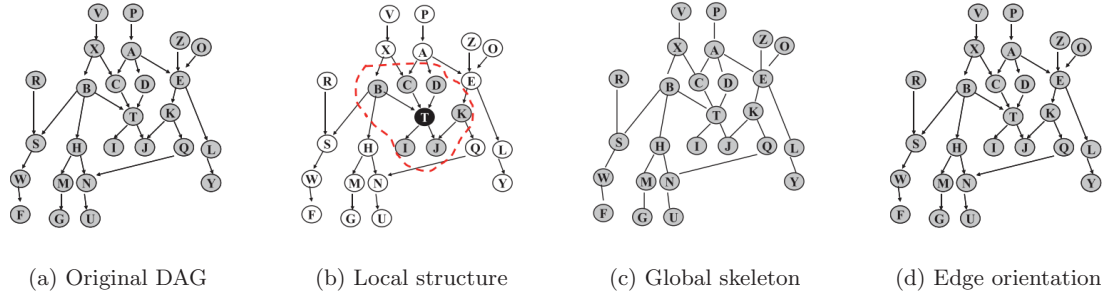


Figure 4: Steps of the general pipeline (LGL) of causal structure learning. The figures are borrowed from (Aliferis et al., 2010a).

Aliferis et al. (Aliferis et al., 2010a) developed a general framework (LGL) for causal discovery based on constraint-based CSL methods. Figure 4b, 4c, 4d refers to the three steps of LGL respectively given the original DAG in Figure 4a. In this framework, the core step is to learn the local structure around each variable (GLL). The local structure can be defined as the parent-children set (PC) or the Markov Blanket (MB) (Aliferis et al., 2010a; Guyon and

Elisseeff, 2003; Tsamardinos et al., 2006; Friedman et al., 2013). There are a number of GLL algorithms proposed in recent time (Aliferis et al., 2010a). These algorithms typically find the candidate set for *PC* and *MB* using classification-based feature selection methods which exploit L^1 - or L^2 -norm formulation or its convex combination for the loss function of a classifier (Aliferis et al., 2010a; Guyon et al., 2002; Zhu et al., 2004; Wang et al., 2006). After the candidate set is selected, CI tests are performed on the subsets of the candidate set to find the local structure (Tsamardinos et al., 2006, 2003a,b; Aliferis et al., 2010b). Once the local structure is recovered, the next phase of LGL framework is to piece together the undirected skeleton from the local GLL results and finally apply any favorable edge orientation rules.

2.3.3 Relational Causal Discovery

The first algorithm to employ constraint-based methods to learn causal models from relational data is RPC but it is not complete and may introduce errors (Maier et al., 2010). Maier et al. (Maier et al., 2013a) proposed a sound and complete CSL algorithm (RCD) for relational data based on relational bi-variate orientation. However, it relies on the existence of a relational dependence oracle for CI test. An operational CI test for relational data is key to relational CSL algorithms. Moreover, a relational CI test is also able to detect causal direction and the presence of confounders without additional parametric assumptions for sparse network structures (Arbour et al., 2016a).

Maier et al. (2013a) applied RCD to the MovieLens+ database, a combination of the UMN MovieLens database (www.grouplens.org); box office, director, and actor information

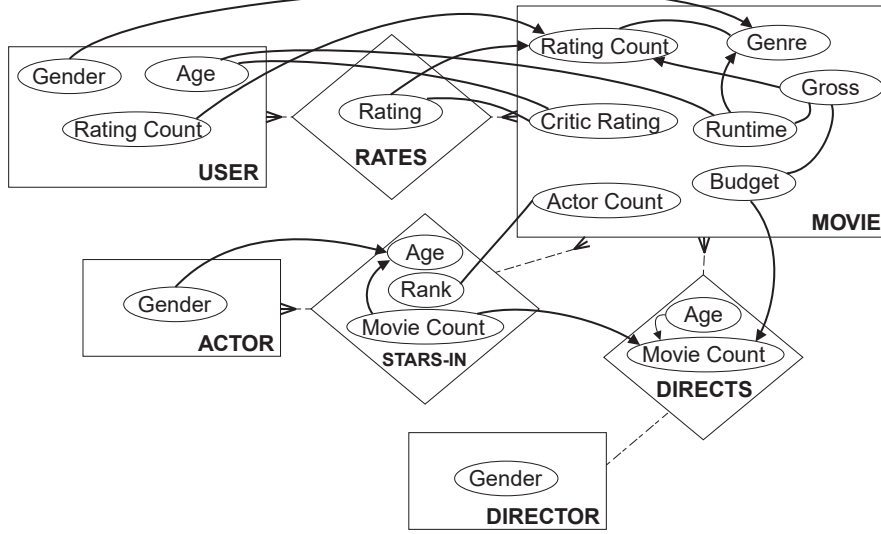


Figure 5: RCD-learned model of MovieLens+ (Maier et al., 2013a).

collected from IMDb (www.imdb.com); and average critic ratings from Rotten Tomatoes (www.rottentomatoes.com). The RCD-generated output is given in Figure 5.

2.3.4 Conditional Independence Test for Relational Data

The Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005) is a statistical test that has been extended to marginal independence testing for structured data (Zhang et al., 2009) and random processes (Chwialkowski et al., 2014). Flaxman et al. (Flaxman et al., 2015) utilize HSIC to develop marginal and conditional independence tests for *propositional* variables in the presence of a latent relational confounder in a single-entity network with an additive noise generating function. A relational dependence exhibits itself through an edge

in the network which is a noisy surrogate of latent homophily. Lee et al. (Lee and Honavar, 2017) extend these tests to explicitly include relational variables. Recently, Lee et al. (Lee and Honavar, 2017) introduced A CSL algorithm based on RCD while using their CI test developed for relational data.

2.4 Active Learning for Relational Data

The most common scenario for active learning is the pool-based scenario where a pool of labeled and unlabeled samples are present and the learner can choose from the unlabeled pool to query for labels (Settles, 2009). An established pool-based active learning algorithm for relational data is ALFNET (Bilgic et al., 2010) which is based on disagreement-based active learning (Seung et al., 1992) using *Iterative Classification Algorithm (ICA)* as relational classifier. Even though it performs well, it is computationally expensive specifically for larger graphs due to the cost of iterative training. Recently proposed neural network-based approaches (Gao et al., 2018; Cai et al., 2017) follow similar expensive procedure of re-training the model over iterations without showing significant improvement over ALFNET. One-shot active learning circumvents this cost of re-training by allowing a single chance to decide which nodes to label. Note that, one-shot active learning is different from one-shot learning (Fei-Fei et al., 2006) or active one-shot learning (Woodward and Finn, 2017). One-shot learning refers to learning from one or few samples and active one-shot learning refers to active learning where one or few samples can be labeled in each iteration. In contrast, one-shot active learning refers to active learning with one iteration to label nodes. To the best of our knowledge there is no prior work investigating one-shot active learning on relational data.

2.4.1 Relational Classification

In contrast to standard classification tasks where data samples are i.i.d, relational classification deals with interconnected samples. The fundamental idea behind relational classification is to effectively exploit the attribute and label correlations between linked nodes to achieve better accuracy in predicting the labels of individual samples. A trivial relational classifier is wVRN (Macskassy and Provost, 2007) which simply infers class association probability based on strong homophily assumption. It requires no learning, rather it classifies the entities of a relational network based only on the relational structure (Macskassy and Provost, 2007). Modern relational classifiers can be categorized into two families: 1) Collective classification and 2) Graph neural networks.

Collective classification refers to the combined classification of a set of connected objects (Sen et al., 2008). The fundamental assumption is that the label of a node not only depends on its own node attributes but also depends on the labels and attributes of its neighboring nodes in the network. Collective classifiers use a vector-based classifier such as logistic regression which is trained iteratively. It learns the conditional probability for estimating node labels. The two most common algorithms for collective classification are Iterative classification algorithm (ICA) (Neville and Jensen, 2000; Lu and Getoor, 2003) and Gibbs Sampling (GS) (Geman and Geman, 1984; Sen et al., 2008).

Graph neural networks (GNN) emerged with the popularity of deep learning architectures. GNN is inspired by the success of convolutional neural networks (CNN) in computer vision. Most of the GNN models are primarily based on redefined notions of convolution for graph data

(Wu et al., 2019a). These convolutional GNNs can be divided into two main categories: spectral and spatial. The most important difference between these fundamental approaches lies in their treatment of the graph laplacian matrix. Spectral methods utilizes eigen-decomposition of the graph laplacian to extract useful information about the graph structure. Spatial methods treat it as spatial connectivity of nodes (Chen et al., 2020). The two most representative algorithms from these two categories are GCN (Kipf and Welling, 2017) and GRAPHsAGE (Hamilton et al., 2017).

2.4.2 Sample Selection from Relational Data

Graph sampling algorithms have been studied for a long time. Kolaczyk et al. (Kolaczyk, 2009) investigated sample properties from a social science perspective. Other works analyzed the statistical properties of sampled subgraphs and how sampling changes topological network properties (Lee et al., 2006; Yoon et al., 2007; Stumpf et al., 2005). Several studies analyzed representativeness (Leskovec and Faloutsos, 2006), correlations of graph properties (Ahmed et al., 2010), biases of topological approaches (Maiya and Berger-Wolf, 2011) and impact on A/B testing (Backstrom and Kleinberg, 2011).

A few recent works studied the effectiveness of sampling methods for relational classification (Ahmed et al., 2013, 2012; Berton et al., 2016; Macskassy, 2009). Ahmed et al. (Ahmed et al., 2012) provided a comprehensive analysis of a variety of graph sampling methods and their effectiveness on relational classification. They sampled subgraphs from a given source graph based on each of the baseline sampling methods. They evaluated the sampling methods based on accuracy of supervised classification models trained with corresponding subgraphs. Their

experiments on four real-world networks show that induced edge sampling (Ahmed et al., 2011) produces better accuracy than any other graph sampling methods (Ahmed et al., 2012). In a more recent work, Berton et al. (Berton et al., 2016) experimentally evaluated effectiveness of centrality-based sampling methods for relational classification. They showed that sampling based on clustering coefficient provides greater accuracy in general. Note that, both these studies considered a supervised classification task and trained the classification model only on the sampled graph. In contrast, our evaluation is based on semi-supervised classification where the full source graph is used in creating the features for training. Moreover, they only considered network-based sampling methods. Moreover, they use a simple classifier, wvRN (Macskassy and Provost, 2007), which relies on label aggregates and has no learning component. Earlier work by Macskassy et al. (Macskassy, 2009) is closely related to ours which is motivated to speed up active learning on graph by sampling a small candidate set of nodes using structural properties from which an Empirical Risk Minimization (ERM) (Roy and McCallum, 2001) method chooses the top candidate to be labeled. However, they also followed the standard active learning procedure of multiple shots for active querying which is costly.

CHAPTER 3

RELATIONAL CAUSAL MODELS WITH CYCLES: REPRESENTATION AND REASONING

Parts of this chapter were previously published as Ahsan, R., Arbour, D., and Zheleva, E.: Relational Causal Models with Cycles: Representation and Reasoning. In Proceedings of the 1st Conference on Causal Learning and Reasoning (CLEaR 2022) Ahsan et al. (2022a)

Many real-world phenomena involve feedback loops or cycles that violate the acyclicity assumption. For example, supply and demand affect price and vice versa, hormone levels in the body affect each other and friends can impact each other’s choices. The existing works on cyclic causal models primarily focus on independently and identically distributed (i.i.d) data instances (Richardson, 1996, 1997; Strobl, 2019b; Rantanen et al., 2020). However, in many real-world systems units are often interconnected in a complex network. Causal reasoning over such relational systems is central to understanding real-world social phenomena, such as social influence and information diffusion.

The development of *relational causal models*, which generalize over structural causal models, is an important step towards capturing interactions between non-i.i.d instances (Maier et al., 2013b,a; Lee and Honavar, 2015; Bhattacharya et al., 2020). Relational models involve multiple types of interacting entities with probabilistic dependencies among their attributes. Maier et al. (2013b) develop a lifted causal representation named *abstract ground graph (AGG)* that abstracts over all instantiations of a relational model. AGG enables reasoning about causal

queries in relational causal models and relational causal discovery. However, existing studies in relational causal models assume acyclicity and do not allow for reasoning about identification in the presence of feedback loops.

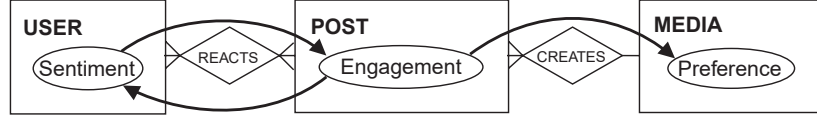
In this work, we specifically study cyclic RCMs and show that they offer the necessary representation to reason about a lot of real-world causal problems where popular assumptions do not hold. To the best of our knowledge, this is the first work that addresses the representation of and reasoning about cyclic relational causal models. We define a new abstract representation, *σ -abstract ground graph* (σ -AGG) which generalizes over cyclic relational models. In order to reason about relational queries in σ -AGG, we introduce *relational σ -separation* and provide proof for its soundness and completeness for all instantiations of a relational model.

3.1 Cyclic Relational Causal Model

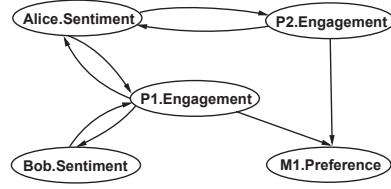
We define cyclic relational causal model which allows the formation of cycles among the relational dependencies of the model.

Definition 7 (Cyclic Relational Causal Model). *A relational model $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ is said to be cyclic if the set of relational dependencies \mathcal{D} constructs one or more directed cycles of arbitrary length.*

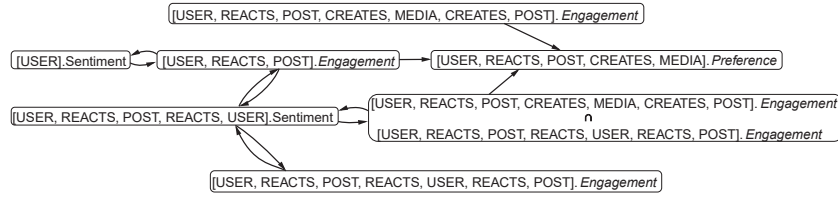
Cycles in RCM represent equilibrium states among a set of relational variables. It implies that the abstract ground graphs are no longer guaranteed to be DAGs. In order to facilitate this, we propose a revised definition of relational dependency provided by Maier et al. (2013b) by relaxing the restriction of having different attribute classes for cause and effect.



(a) Cyclic RCM



(b) Ground Graph

(c) σ -Abstract Ground GraphFigure 6: A cyclic relational causal model, corresponding ground graph, and σ -AGG.

Definition 8 (Relational Dependency). *A relational dependency $[I_j, \dots, I_k].X' \rightarrow [I_j].X$ is a directed probabilistic dependence from any attribute class X' to X through the relational path $[I_j, \dots, I_k]$ such that $I_j, \dots, I_k \in \mathcal{E} \cup \mathcal{R}$, $X, X' \in \mathcal{A}$. Note that it is possible to have $X = X'$.*

3.1.1 Example

Figure 6 shows an example relational model with cyclic dependencies (i.e. dashed arrows). We see a pair of dependencies $[Post, Reacts, User].Sentiment \rightarrow [Post].Engagement$ and $[Post].Engagement \rightarrow [Post, Reacts, User].Sentiment$ which are inverse to each other and form a feedback loop. However, this mere feedback loop prohibits the use of AGG to answer relational causal query that asks whether a user’s Sentiment about a post they reacted to is independent of the preference of media given the posts. Unfortunately, the work by Maier et al. (2013b) is not sufficient to reason about conditional independence relationships in the ground graphs of such relational models since they contain cycles. This motivates us to introduce a new criterion that enables the abstraction of relational queries with cyclic dependencies over all ground graphs.

3.2 Relational σ -separation

Conditional independence facts are only useful when they hold across all ground graphs that are consistent with the model. Maier et al. (2013b) show that relational d -separation is sufficient to achieve that for acyclic models. However, such abstraction is not possible for cyclic models since the correctness of d -separation is not guaranteed for cyclic graphical models (Spirtes, 1995; Neal, 2000). In this work, we propose the following definition of relational σ -separation specifically for cyclic relational models:

Definition 9 (Relational σ -separation). *Let \mathbf{X} , \mathbf{Y} , and \mathbf{Z} be three distinct sets of relational variables with the same perspective $B \in \mathcal{E} \cup \mathcal{R}$ defined over relational schema \mathcal{S} . Then, for relational model structure \mathcal{M} , \mathbf{X} and \mathbf{Y} are σ -separated by \mathbf{Z} if and only if, for all skeletons*

$s \in \sum_{\mathcal{S}}$, $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ are σ -separated by $\mathbf{Z}|_b$ in ground graph $GG_{\mathcal{M}_s}$ for all instances $b \in s(B)$ where $s(B)$ refers to the instances of B in skeleton s .

The definition directly follows from the definition of relational d -separation. If there exists even one skeleton and faithful distribution represented by the relational model for which $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, then $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ are not σ -separated by $\mathbf{Z}|_b$ for $b \in s(B)$.

3.3 σ -Abstract Ground Graph

We refer to the lifted representation for cyclic RCMs as σ -abstract ground graph or σ -AGG. A σ -AGG is constructed using the same *extend* method used to construct AGG (Maier et al., 2013b).

Definition 10 (σ -Abstract Ground Graph). *An abstract ground graph σ -AGG $_{\mathcal{M}} = (V, E)$ for relational model structure $\mathcal{M} = (\mathcal{S}, \mathcal{D})$, perspective $B \in \mathcal{E} \cup \mathcal{R}$, and hop threshold $h \in \mathbb{N}^0$ is a directed graph that abstracts the dependencies \mathcal{D} for all ground graphs $GG_{\mathcal{M}_s}$, where $s \in \sum_{\mathcal{S}}$. The σ -AGG $_{\mathcal{M}_s}$ is a directed cyclic graph with the following nodes and edges:*

1. $V = RV \cup IV$, where
 - (a) RV is the set of relational variables with a path of length at most $h + 1$.
 - (b) IV are intersection variables between pairs of relational variables that could intersect
2. $E = RVE \cup IVE$, where
 - (a) $RVE \subset RV \times RV$ are the relational variable edges

(b) $IVE \subset (IV \times RV) \cup (RV \times IV)$ are the intersection variable edges. This is the set of edges that intersection variables “inherit” from the relational variables that they were created from

Since the construction of an AGG and a σ -AGG $_{\mathcal{M}_s}$ is identical, they share mostly identical properties as defined by Maier et al. (2013b) for AGG. The main difference is the existence of cycles. Consequently, the goal of σ -AGG $_{\mathcal{M}_s}$ is to reason about relational σ -separation queries instead of relational d -separation. Figure 6c shows the σ -AGG $_{\mathcal{M}_s}$ corresponding to the cyclic RCM in 6a with a pairwise feedback loop. It is similar to the AGG in Figure 2c but allows cycles without violating the conditional independence statements under σ -separation which are otherwise undefined with d -separation.

3.4 Theoretical Guarantees of Relational σ -separation

In order to discuss theoretical guarantees of relational σ -separation we first address the open problem of necessary conditions for the completeness of relational d -separation.

3.4.1 Completeness of Relational d -separation Under AGG

Previous work has shown that the original claim of completeness of relational d -separation by Maier et al. (2013a) cannot be guaranteed for any relational model (Lee and Honavar, 2015). A counterexample has been developed as well. In this work, we show that relational d -separation is complete under the following assumption:

A 7. *The degree of any entity in the relational skeleton is greater than 1.*

Note that, this assumption is about the topology of the ground graphs, i.e., the network which defines how entities are connected to each other. It only allows entities that are connected to at least two other entities. For example, if the entities are users in a social network, the framework would only consider users who have degree at least two, i.e., are connected to at least two other users. While this restricts the space of graph topologies allowable under the results in this work, many networks observed in real-world domains, such as social networks, have a minimum degree greater than one. We introduce the following lemma that establishes sufficient conditions for AGGs to be realizable in ground graphs. This result may be of independent interest since it provides sufficient conditions for soundness in the original presentation of relational d -separation under additional assumptions (Maier et al., 2013b; Lee and Honavar, 2015).

Lemma 1. *Under assumption 7, every abstract ground graph can be realized as a ground graph. That is, for every acyclic relational model \mathcal{M} and skeleton $s \in \sum_{\mathcal{S}}$ any relational variable in $AGG_{\mathcal{M}_s}$ has non-empty terminal sets in some ground graph $GG_{\mathcal{M}_s}$.*

Proof. We first consider the conditions under which empty terminal sets can occur, resulting in an AGG that is unrealizable in the ground graphs. There are two necessary and sufficient conditions for empty terminal sets to appear in all ground graphs corresponding to an AGG. First, there must be at least one intersection variable present in the AGG. If no intersection variable exists in the AGG, then the completeness proof of relational d -separation by Maier et al. (2013b) holds. The second condition is that the intersection must be on a path consisting of only one-to-one relationships. In order to understand this condition, let's look at an example

with the following relational paths from a hypothetical relational model which is a generalization of the counterexample given by Lee and Honavar (2015) ¹:

- $P = [E_b, \dots, R_j, \dots, E_x]$
- $Q = [E_b, \dots, R_j, \dots, R_m, E_x, \dots, E_y]$
- $S = [E_b, \dots, R_j, \dots, R_m, \dots, E_z]$
- $S' = P + [R_m, \dots, E_z]$

where E_b, E_x, E_y, E_z are some entity classes, R_j, R_m are relationship classes, “...” are arbitrary valid sequences of entities and relationships, and $+$ represents the concatenation of relational paths. Let’s assume two relational dependencies exist in the given model, $P.X \rightarrow Q.Y$ and $S.Z \rightarrow Q.Y$ where X, Y, Z are attributes of corresponding entity classes. By definition, the corresponding edges $P.X \rightarrow Q.Y$, $S.Z \rightarrow Q.Y$ appear in the AGG. Since S and S' are intersectable an additional edge $S.Z \cap S'.Z \rightarrow Q.Y$ also appears in the AGG. Such a model can be realized in many possible ground graphs. Figure 7 shows the general pattern of such ground graphs. Now, if we restrict the relationships to be strictly one-to-one, then there is only one skeleton structure possible to satisfy the relational dependencies at the cost of $S'.Z$ having empty terminal sets since an instance of R_m can connect to only one instance of E_x . If we allow many-to-many relationships then we can always construct a skeleton where an instance of R_m connects to two instances of E_x (through the dotted line in Figure 7) to produce non-empty terminal sets for both Q and S' .

¹The complete counterexample and figure explaining it are given in Section 2.2.4

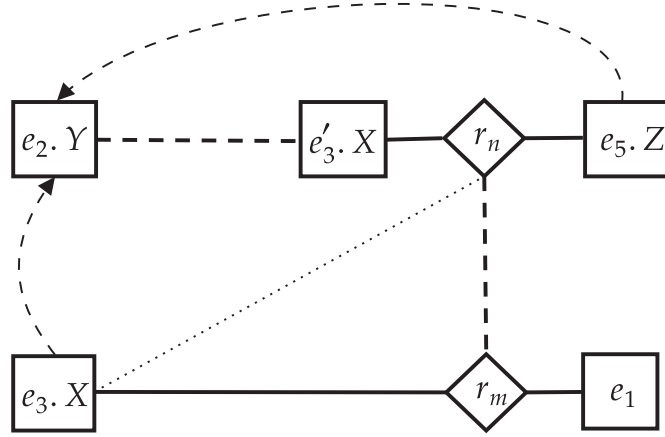


Figure 7: General pattern of the counterexample by (Lee and Honavar, 2015). The notations inside the square/rhombus refer to instances of the corresponding entity/relationship classes. The dashed arrows represent realizations of the relational dependencies. The dashed lines can be replaced with arbitrary length valid relational paths. The dotted line represents a hypothetical connection that can nullify the counterexample under assumption 7.

Since assumption 7 prohibits the second condition, it essentially implies that any relational variable in $\text{AGG}_{\mathcal{M}_s}$ results in non-empty terminal sets in corresponding ground graphs for every acyclic relational model \mathcal{M} and skeleton $s \in \sum_{\mathcal{S}}$ which completes the proof for Lemma 1. \square

The following proposition establishes the completeness of AGG for relational d -separation under the assumption of a minimum degree greater than 1.

Proposition 2. *AGG is sound and complete for relational d -separation under assumption 7.*

Proof. Following lemma 1, the original proof of soundness and completeness of relational d -separation by Maier et al. (2013b) directly applies which proves proposition 2. \square

The correctness of our approach to relational σ -separation relies on several facts which are similar to the case for AGG: (1) σ -separation is valid for directed cyclic graphs; (2) ground graphs are directed cyclic graphs; and (3) σ -AGGs are directed cyclic graphs that represent exactly the edges that could appear in all possible ground graphs. Note that we no longer need assumption 3, but assumptions 4 and 7 are adopted from relational d -separation. Using the previous definitions and lemmas, the following additional assumptions and sequence of results prove the correctness of our approach to identifying independence in cyclic relational models.

A 8. *The given cyclic relational model structure is σ -faithful.*

3.4.2 Soundness and Completeness of σ -separation

Theorem 1. *The rules of σ -separation are sound and complete for cyclic directed graphs.*

Proof. Forré and Mooij (2017) show that for quite general structural equation models HEDGes¹ always follow a directed global Markov property based on σ -separation which completes the proof for soundness since directed cyclic graphs are subsets of HEDGes. The completeness claim is already covered by Assumption 8. \square

¹Definition given in Chapter 2

Theorem 2. *For every cyclic relational causal model $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ and skeleton $s \in \sum_{\mathcal{S}}$ such that relational variables involved in \mathcal{D} are non-empty, the ground graph $GG_{\mathcal{M}_s}$ is a cyclic directed graph.*

Proof. Let's assume for contradiction that there exists an acyclic ground graph g which is a realization of a given cyclic RCM $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ and skeleton $s \in \sum_{\mathcal{S}}$. According to the definition of ground graphs, the edges of ground graphs are directly constructed based on the relational dependencies of the model. Definition 7 states that a cyclic RCM consists of one or more cycles formed by relational dependencies. Assume a cycle in cyclic RCM is formed by the pair of relational dependencies as follows: a) $P.j \rightarrow Q.k$, and b) $Q.k \rightarrow P.j$ where P and Q are relational paths from some perspective b and i, j refers to two attribute classes. By construction of g there must be two nodes a, b in g corresponding to $P.j$ and $Q.k$ respectively. Moreover, the definition of g requires two edges $a \rightarrow b$ and $b \rightarrow a$ to be present in the ground graph. But such edges construct a cycle that is contradictory to the initial claim. Thus, the ground graph g must be cyclic. \square

3.4.3 Soundness and Completeness of $\sigma\text{-AGG}_{\mathcal{M}_s}$

Theorem 3. *For every cyclic relational model structure \mathcal{M} and perspective $B \in \mathcal{E} \cup \mathcal{R}$, the $\sigma\text{-AGG}_{\mathcal{M}_s}$ is sound and complete for all ground graphs $GG_{\mathcal{M}_s}$ with skeleton $s \in \sum_{\mathcal{S}}$.*

The proof follows from the proof of soundness and completeness of AGG (Maier et al., 2013b).

Proof. Let $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ be an arbitrary cyclic relational model structure and $B \in \mathcal{E} \cup \mathcal{R}$ an arbitrary perspective.

Soundness: To prove that $\sigma\text{-AGG}_{\mathcal{M}_s}$ is sound, we must show that for every edge $P_k.X \rightarrow P_j.Y$ in $\sigma\text{-AGG}_{\mathcal{M}_s}$, there exists a corresponding edge $i_k.X \rightarrow i_j.Y$ in the ground graph $GG_{\mathcal{M}_s}$ for some skeleton $s \in \Sigma_{\mathcal{S}}$, where $i_k \in P_k|_b$ and $i_j \in P_j|_b$ for some $b \in s(B)$. There are three subcases, one for each type of edge in an abstract ground graph:

(a) Let $[B, \dots, I_k].X \rightarrow [B, \dots, I_j].Y \in RVE$ be an arbitrary edge in $\sigma\text{-AGG}_{\mathcal{M}_s}$ between a pair of relational variables. Assume for contradiction that there exists no edge $i_k.X \rightarrow i_j.Y$ in any ground graph:

$$\begin{aligned} \forall s \in \Sigma_{\mathcal{S}}, \forall i_k \in [B, \dots, I_k]|_b, \forall i_j \in [B, \dots, I_j]|_b \\ (i_k.X \rightarrow i_j.Y \notin GG_{\mathcal{M}_s}) \end{aligned}$$

By Definition 10 for $\sigma\text{-AGG}_{\mathcal{M}_s}$, if $[B, \dots, I_k].X \rightarrow [B, \dots, I_j].Y \in RVE$, then the model must have dependency $[I_j, \dots, I_k].X \rightarrow [I_j].Y \in \mathcal{D}$ such that $[B, \dots, I_k] \in \text{extend}([B, \dots, I_j], [I_j, \dots, I_k])$. So, by the definition of ground graphs, there is an edge from every $i_k.X$ to every $i_j.Y$, where i_k is in the terminal set for i_j along $[I_j, \dots, I_k]$. Therefore, there exists a ground graph $GG_{\mathcal{M}_s}$ such that $i_k.X \rightarrow i_j.Y \in GG_{\mathcal{M}_s}$, which contradicts the assumption.

(b) Let $P_1.X \cap P_2.X \rightarrow [B, \dots, I_j].Y \in IVE$ be an arbitrary edge in $\sigma\text{-AGG}_{\mathcal{M}_s}$ between an intersection variable and a relational variable, where $P_1 = [B, \dots, I_m, \dots, I_k]$ and

$P_2 = [B, \dots, I_n, \dots, I_k]$ with $I_m \neq I_n$. By Definition 10, if the σ -abstract ground graph has edge $P_1.X \cap P_2.X \rightarrow [B, \dots, I_j].Y \in IVE$, then either $P_1.X \rightarrow [B, \dots, I_j].Y \in RVE$ or $P_2.X \rightarrow [B, \dots, I_j].Y \in RVE$. Then, as shown in case (a), there exists an $i_j \in [B, \dots, I_j]|_b$ such that $i_k.X \rightarrow i_j.Y \in GG_{\mathcal{M}_s}$, which contradicts the assumption.

(c) Let $[B, \dots, I_k].Y \rightarrow P_1.X \cap P_2.X \in IVE$ be an arbitrary edge in $\sigma\text{-AGG}_{\mathcal{M}_s}$ between an intersection variable and a relational variable, where $P_1 = [B, \dots, I_m, \dots, I_j]$ and $P_2 = [B, \dots, I_n, \dots, I_j]$ with $I_m \neq I_n$. The proof follows case (b) to show that there exists a skeleton $s \in \sum_{\mathcal{S}}$ and $b \in s(B)$ such that for all $i_k \in [B, \dots, I_k]|_b$ there exists an $i_j \in P_1 \cap P_2|_b$ such that $i_k.X \rightarrow i_j.Y \in GG_{\mathcal{M}_s}$.

Completeness: To prove that the σ -abstract ground graph $\sigma\text{-AGG}_{\mathcal{M}_s}$ is complete, we show that for every edge $i_k.X \rightarrow i_j.Y$ in every ground graph $GG_{\mathcal{M}_s}$ where $s \in \sum_{\mathcal{S}}$, there is a set of corresponding edges in $\sigma\text{-AGG}_{\mathcal{M}_s}$. Specifically, the edge $i_k.X \rightarrow i_j.Y$ yields two sets of relational variables for some $b \in s(B)$, namely $\mathbf{P}_k.X = \{P_k.X | i_k \in P_k|_b\}$ and $\mathbf{P}_j.Y = \{P_j.Y | i_j \in P_j|_b\}$. Note that all relational variables in both $\mathbf{P}_k.X$ and $\mathbf{P}_j.Y$ are nodes in $\sigma\text{-AGG}_{\mathcal{M}_s}$, as are all pairwise intersection variables. We show that for all $P_k.X \in \mathbf{P}_k.X$ and for all $P_j.Y \in \mathbf{P}_j.Y$ either (a) $P_k.X \rightarrow P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ (b) $P_k.X \cap P'_k.X \rightarrow P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ where $P'_k.X \in \mathbf{P}_k.X$, or (c) $P_k.X \rightarrow P_j.Y \cap P'_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ where $P'_j.Y \in \mathbf{P}_j.Y$.

Let $s \in \sum_{\mathcal{S}}$ be an arbitrary skeleton, let $i_k.X \rightarrow i_j.Y \in GG_{\mathcal{M}_s}$ be an arbitrary edge drawn from $[I_j, \dots, I_k].X \rightarrow [I_j].Y \in \mathcal{D}$, and let $P_k.X \in P_k.X, P_j.Y \in P_j.Y$ be an arbitrary pair of relational variables.

(a) If $P_k \in \text{extend}(P_j, [I_j, \dots, I_k])$, then $P_k.X \rightarrow P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ by Definition 10.

- (b) If $P_k \notin \text{extend}(P_j, [I_j, \dots, I_k])$, but $\exists P'_k \in \text{extend}(P_j, [I_j, \dots, I_k])$ such that $P'_k.X \in P_k.X$, then $P'_k.X \rightarrow P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$, and $P_k.X \cap P'_k.X \rightarrow P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ by Definition 10.
- (c) If $\forall P \in \text{extend}(P_j, [I_j, \dots, I_k])(P.X \notin P_k.X)$, then $\exists P'_j$ such that $i_j \in P'_j|_b$ and $P_k \in \text{extend}(P'_j, [I_j, \dots, I_k])$. Therefore, $P'_j.Y \in P_j.Y$, $P_k.X \rightarrow P'_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$, and $P_k.X \rightarrow P'_j.Y \cap P_j.Y \in \sigma\text{-AGG}_{\mathcal{M}_s}$ by Definition 10.

□

Theorem 4. *The abstract ground graph $\sigma\text{-AGG}_{\mathcal{M}_s}$ is a cyclic directed graph if and only if the underlying relational model structure is cyclic.*

Proof. Let \mathcal{M} be an arbitrary (possibly) cyclic relational model structure, and let $B \in \mathcal{E} \cup \mathcal{R}$ be an arbitrary perspective. It is clear by Definition 10 that every edge in the abstract ground graph $\sigma\text{-AGG}_{\mathcal{M}_s}$ is directed by construction. Assume for contradiction that no cycles exist in $\sigma\text{-AGG}_{\mathcal{M}_s}$ even if the relational dependencies form one or more cycles. Now assume the following two dependencies are part of the given relational model \mathcal{M} : 1. $[I_j, \dots, I_k].X \rightarrow [I_j].Y \in D$, 2. $[I_j].Y \rightarrow [I_j, \dots, I_k].X \in D$ where $I_j, \dots, I_k \in \mathcal{E} \cup \mathcal{R}$. By Definition 10, all edges inserted in $\sigma\text{-AGG}_{\mathcal{M}_s}$ are drawn from some dependency in \mathcal{M} , and edges in $\sigma\text{-AGG}_{\mathcal{M}_s}$ are constructed for all the dependencies in D . As a result, there must be corresponding edges in the $\sigma\text{-AGG}_{\mathcal{M}_s}$ for both dependencies that form a cycle, which contradicts the assumption.

Now, assume that a $\sigma\text{-AGG}_{\mathcal{M}_s}$ is acyclic even if the underlying RCM is cyclic. Using the same argument as above we can say that the edges in the $\sigma\text{-AGG}_{\mathcal{M}_s}$ constructed based on the

dependencies in \mathcal{D} . If a cycle exists in the $\sigma\text{-AGG}_{\mathcal{M}_s}$ it directly implies the existence of a cycle in the RCM which leads to a contradiction. Thus the proof completes from both directions. \square

3.4.4 Soundness and Completeness of relational σ -separation

Theorem 5. *Relational σ -separation is sound and complete for $\sigma\text{-AGG}$. Let \mathcal{M} be a (possibly) cyclic relational model structure, and let \mathbf{X} , \mathbf{Y} , and \mathbf{Z} be three distinct sets of relational variables defined over relational schema \mathcal{S} . Then, $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are σ -separated by $\bar{\mathbf{Z}}$ on the abstract ground graph $\sigma\text{-AGG}_{\mathcal{M}_s}$ if and only if for all skeletons $s \in \sum_{\mathcal{S}}$ and for all perspectives $b \in s(B)$, $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ are σ -separated by $\mathbf{Z}|_b$ in ground graph $GG_{\mathcal{M}_s}$.*

Proof. We must show that σ -separation on an abstract ground graph implies σ -separation on all ground graphs it represents (soundness) and that σ -separation facts that hold across all ground graphs are also entailed by σ -separation on the abstract ground graph (completeness). The proof follows from the proof of soundness and completeness of AGG (Maier et al., 2013b).

Soundness:

Assume that $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are σ -separated by $\bar{\mathbf{Z}}$ on $\sigma\text{-AGG}_{\mathcal{M}_s}$. Assume for contradiction that there exists an item instance b such that $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ are not σ -separated by $\mathbf{Z}|_b$ in the ground graph $GG_{\mathcal{M}_s}$ for some arbitrary skeleton s . Then, there must exist a σ -connecting path p from some $x \in \bar{\mathbf{X}}|_b$ to some $y \in \bar{\mathbf{Y}}|_b$ given all $z \in \bar{\mathbf{Z}}|_b$. By Theorem 3, $\sigma\text{-AGG}_{\mathcal{M}_s}$ is complete, so all edges in $GG_{\mathcal{M}_s}$ are captured by edges in $\sigma\text{-AGG}_{\mathcal{M}_s}$. So, path p must be represented from some node in $\{N_x | x \in N_x|_b\}$ to some node in $\{N_y | y \in N_y|_b\}$, where N_x, N_y are nodes in $\sigma\text{-AGG}_{\mathcal{M}_s}$. If p is σ -connecting in $GG_{\mathcal{M}_s}$, then it is σ -connecting in $\sigma\text{-AGG}_{\mathcal{M}_s}$,

implying that $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are not σ -separated by $\bar{\mathbf{Z}}$. So, $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ must be σ -separated by $\mathbf{Z}|_b$.

Completeness:

Assume that $\mathbf{X}|_b$ and $\mathbf{Y}|_b$ are σ -separated by $\mathbf{Z}|_b$ in the ground graph $GG_{\mathcal{M}_s}$ for all skeletons s for all $b \in s(B)$. Assume for contradiction that $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are not σ -separated by $\bar{\mathbf{Z}}$ on $\sigma\text{-AGG}_{\mathcal{M}_s}$. Then, there must exist a σ -connecting path p for some relational variable $X \in \bar{\mathbf{X}}$ to some $Y \in \bar{\mathbf{Y}}$ given all $Z \in \bar{\mathbf{Z}}$. By Theorem 3, $\sigma\text{-AGG}_{\mathcal{M}_s}$ is sound, so every edge in $\sigma\text{-AGG}_{\mathcal{M}_s}$ must correspond to some pair of variables in some ground graph. So, if p is σ -connecting in $\sigma\text{-AGG}_{\mathcal{M}_s}$, then there must exist some skeleton s such that p is σ -connecting in $GG_{\mathcal{M}_s}$ for some $b \in s(B)$, implying that σ -separation does not hold for that ground graph. So, $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ must be σ -separated by $\bar{\mathbf{Z}}$ on $\sigma\text{-AGG}_{\mathcal{M}_s}$. \square

Maier et al. (2013b) show that relational d -separation is equivalent to the Markov condition on acyclic relational models. However, it doesn't hold for the cyclic relational model. Here, we show how relational σ -separation is equivalent to the Markov condition on cyclic relational models.

3.4.5 Relational σ -separation Markov Condition

Definition 11 (Relational σ -separation Markov Condition). *Let X, Y, Z be relational variables for perspective $B \in \mathcal{E} \cup \mathcal{R}$ defined over relational schema \mathcal{S} . For any solution (\mathcal{X}, ϵ) of a relational model \mathcal{M} which follows a simple SCM,*

$$X \underset{\mathcal{M}}{\overset{\sigma}{\perp\!\!\!\perp}} Y|Z \implies \boldsymbol{\mathcal{X}}_X \underset{\mathbb{P}_{\mathcal{M}}(\boldsymbol{\mathcal{X}})}{\perp\!\!\!\perp} \boldsymbol{\mathcal{X}}_Y|\boldsymbol{\mathcal{X}}_Z, \text{ if and only if}$$

$$x \underset{GG_{\mathcal{M}}}{\overset{\sigma}{\perp\!\!\!\perp}} y|z \implies \boldsymbol{\mathcal{X}}'_x \underset{\mathbb{P}_{GG_{\mathcal{M}}}(\boldsymbol{\mathcal{X}}')}{\perp\!\!\!\perp} \boldsymbol{\mathcal{X}}'_y|\boldsymbol{\mathcal{X}}'_z, \text{ for } \forall x \in X|_b, \forall y \in Y|_b, \forall z \in Z|_b$$

in ground graph $GG_{\mathcal{M}_s}$ for all skeletons $s \in \sum_{\mathcal{S}}$ and for all $b \in s(B)$ where $(\boldsymbol{\mathcal{X}}', \boldsymbol{\mathcal{E}}')$ refers to the solution of the SCM corresponding to the ground graphs.

In other words, σ -separation of two relational variables \boldsymbol{X} and \boldsymbol{Y} given a third relational variable \boldsymbol{Z} would imply \boldsymbol{X} and \boldsymbol{Y} are conditionally independent given \boldsymbol{Z} if and only if, for all instances of $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ in all possible ground graphs, the same condition holds. Since ground graphs of cyclic RCM are directed cyclic graphs and σ -separation on $\sigma\text{-AGG}_{\mathcal{M}_s}$ is sound and complete (by Theorem 5), we can conclude that relational σ -separation is equivalent to the relational Markov property.

3.5 Discussion

Cycles or feedback loops are common elements of many real-world systems. Unfortunately, it is hardly studied in the field of causal inference primarily because of the nice properties of directed acyclic graphs. As a result, cycles and feedback loops are mostly avoided in the domain of the relational causal model. In this study, we take a step forward to bridge this gap by developing an abstract representation and a criterion to reason about statistical relationships in relational models with or without cycles under a general framework. We show that the new criterion called σ -separation can consistently capture the statistical independence relationships

of all possible instantiations of a relational causal model. We believe that this work will open the door for further development including but not limited to causal structure learning of relational models with cycles.

CHAPTER 4

LEARNING RELATIONAL CAUSAL MODELS WITH CYCLES

Influence in complex dynamic systems is often mutual and represented by a feedback loop or cycle in the relational model. Identifying mutual influence in relational models is of great interest in the research community. For example, social scientists and marketing experts are interested to study the social dynamics between people and products in social networks (Bakshy et al., 2011, 2015; Ogburn et al., 2020). However, there is a lack of available methods for discovering mutual influence or cycles in complex relational systems.

Sound and complete algorithms have been proposed for learning relational causal models from observational data (Maier et al., 2013a; Lee and Honavar, 2016a,b). However, they assume acyclicity and thus cannot reason about mutual influence or cycles. In this work, we examine the problem of learning cyclic relational causal models from observational samples under a suitable set of assumptions. We introduce *relational acyclification*, an operation that helps to reason over the scope of cyclic relational models which are identifiable with constraint-based causal discovery algorithms. Following this criterion, we establish sufficient conditions for which, RCD (Maier et al., 2013a), a pioneering relational causal discovery algorithm for acyclic relational models, is sound and complete for cyclic relational models under σ -separation and causal sufficiency assumption. We provide experimental results on synthetic relational models in support of our claims. We also demonstrate the effectiveness of the algorithm on a real-world dataset.

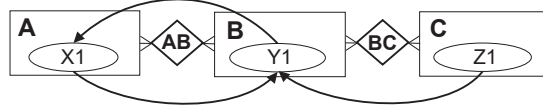
4.1 Relational Causal Discovery with Cycles

Cyclic relational causal models (CRCM) are relational causal models where dependencies form one or more directed cycles (Ahsan et al., 2022a). The cycles or feedback loops can represent equilibrium states in dynamic systems. Consider the example from Figure 1 where sentiments of users and engagements in a media post may reach an equilibrium. Identifying such cycles or feedback loops from observational samples requires proper representation and a learning algorithm. In Chapter 3, I introduce an abstract representation, σ -AGG that entails all the conditional independence relations consistent across all ground graphs of the model and shows that it is sound and complete under σ -separation Ahsan et al. (2022a). Given σ -AGG representation, discovering CRCM transforms into the problem of learning the σ -AGG from observational samples of a relational model. Since σ -AGG is a DCG, we can consider DPAGs to represent the equivalence class of σ -AGG following the previous work of Richardson (1996).

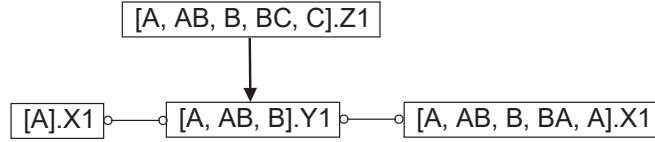
Theorem 1 (Cyclic Relational Causal Discovery). *Given observational samples from a σ -faithful cyclic relational causal model $\mathcal{M} = \langle \mathcal{S}, \mathcal{D} \rangle$ with hop threshold h , learn the maximally oriented DPAG that contains the corresponding σ -AGGs of \mathcal{M} .*

4.1.1 RCD for Cyclic Relational Causal Models

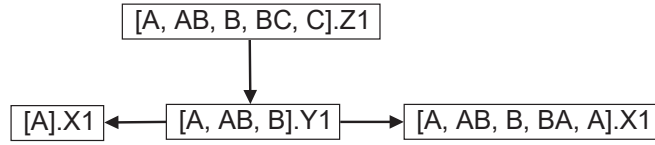
The RCD algorithm developed by Maier et al. (2013a) is the first sound and complete constraint-based algorithm that can learn relational dependencies of a relational causal model (RCM) under the assumption of d -faithfulness, sufficiency, acyclicity, and a maximum hop threshold h . It is designed based on the PC algorithm with additional steps introduced specifically to handle relational aspects of the representation.



(a) Cyclic RCM



(b) True AGG for perspective A



(c) RCD output for perspective A

Figure 8: Counterexample showing RCD produces incorrect output for cyclic RCM under σ -separation.

Following the recent development by Mooij and Claassen (2020) (Corollary 1), and considering that RCD is developed based on the PC algorithm, a natural question arises: *Is RCD sound and complete for cyclic relational causal models?* To the best of our knowledge, no prior work addresses this question. More generally, the effectiveness and theoretical guarantees of existing relational causal structure learning algorithms for cyclic RCMs under σ -separation are not studied in the current literature.

4.1.2 A Counterexample

We present a counterexample that shows that RCD is not sound and complete for discovering cyclic relational causal models in general. Figure 8a shows a CRCM with three entity types A,B,C, and two relationship types AB, BC and maximum hop threshold $h = 2$. The attribute types X1, Y1, and Z1 refer to the attributes of entity types A, B, and C respectively. There are three relational dependencies: 1) $[A, AB, B].Y1 \rightarrow [A].X1$, 2) $[B, AB, A].X1 \rightarrow [B].Y1$, and 3) $[B, BC, C].Z1 \rightarrow [B].Y1$. The first two dependencies form a feedback loop. Figure 8b shows the true σ -AGG built from perspective A with maximum hop threshold $h = 4$ ¹. Figure 8c shows the output of RCD with a σ -separation oracle. We see that RCD orients arrows $[A, AB, B].Y1 \rightarrow [A].X1$ and $[A, AB, B].Y1 \rightarrow [A, AB, B, AB, A].X1$ which refers to the relational dependency $[A, AB, B].Y1 \rightarrow [A].X1$. However, the true model contains a feedback loop between $[A, AB, B].Y1$ and $[A].X1$. This example shows that RCD, even with σ -separation oracle produces incorrect edge orientations.

4.2 Learning Cyclic Relational Causal Models

In this section, we present relational acyclification which enables the discovery of relational causal models with cycles. We also discuss how to read off features of the true model from the output of the discovery algorithm.

¹Maier et al. (2013b) showed that the AGG needs to include nodes with higher hop thresholds than the model in order for it to be a sound and complete representation. However, the hop threshold of relational dependencies should be bounded by the hop threshold of the model. The same arguments hold for σ -AGGs as well. We refer readers to Theorem E.2 of (Maier et al., 2013b)

4.2.1 Relational Acyclification

The counterexample in the previous subsection shows that the RCD algorithm is not sound and complete for general cyclic RCMs under σ -separation. For the given counterexample, RCD orients edges that contradict the given relational model. In order to understand what causes this error and to find a solution, we focus on the acyclification operation introduced by Forré and Mooij (2017) which is a key tool for the generalization results by Mooij and Claassen (2020).

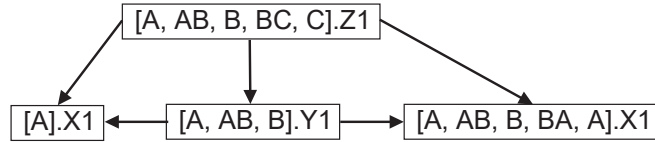


Figure 9: Invalid acyclification of σ -AGG from Figure 8b

Figure 9 shows an acyclification of the σ -AGG presented in Figure 8b following definition 5. Here we see the edges $[A, AB, B, BC, C].Z1 \rightarrow [A].X1$ and $[A, AB, B, BC, C].Z1 \rightarrow [A, AB, B, AB, A].X1$ which does not follow the relational model since the hop threshold of such dependencies ($h = 4$) exceed the hop threshold of the given model ($h = 2$). The definition of acyclification, as given by (Forré and Mooij, 2017) essentially considers all the nodes or entities to be of the same entity type. As a result, applying it directly to relational models creates erroneous results. We propose a new definition of acyclification for relational models which

specifically mentions that the maximum hop threshold of an acyclification can be different than the hop threshold of the original model.

Definition 12 (Relational Acyclification). *Given a relational schema $\mathcal{S} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \text{card})$, σ -AGG $G = (V, E)$, and a hop threshold h , a relational acyclification of G is a σ -AGG $G' = (V, E')$ with hop threshold $h' \geq h$ containing*

- i the same nodes V ;*
- ii for any pair of nodes $P.X, Q.Y$ such that $P.X \notin SC_G(Q.Y)$: $P.X \rightarrow Q.Y \in E'$ iff there exists a node $R.Z$ such that $R.Z \in SC_G(Q.Y)$ and $P.X \rightarrow R.Z \in E$ and $P.X \rightarrow Q.Y$ is a valid relational dependency with maximum hop threshold h' ;*
- iii for any pair of distinct nodes $P.X, Q.Y$ such that $P.X \in SC_G(Q.Y)$: $P.X \rightarrow Q.Y \in E'$ or $P.X \leftarrow Q.Y \in E'$;*

The definition of relational acyclification follows from Definition 5 where the main distinction is that it allows a new bound on the maximum hop threshold which is different than the bound of the original model. The implication of this is that the potential dependencies RCD considers in building the skeleton, may not be sufficient for soundness and completeness.

4.2.2 Maximum Hop Threshold for Relational Acyclification

Definition 12 suggests that the maximum hop threshold used in a relational acyclification of a σ -AGG may be higher than the hop threshold of the given model. It is important to characterize the maximum bound of relational acyclifications for allowing practical implementation of the

RCD algorithm for cyclic models. The following proposition provides the maximum bound on the hop threshold of relational acyclifications.

Proposition 3. *Given a relational model $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ with hop threshold h and corresponding σ -AGG $G = (V, E)$ with a given perspective, the hop threshold h' of any relational acyclification G' of G can be at most $\lfloor \frac{2+l^c}{2} \rfloor h$ where l^c refers to the length of the longest cycle of dependencies in the relational model \mathcal{M} .*

The need for higher hop thresholds arises for the additional edges drawn for any incoming edges to a strongly connected component (Definition 5). Any such incoming edge has a maximum hop threshold h of the given model. In order to reach the farthest node in the cycle where each dependency can be of at most h hop threshold, we need at most $\lfloor \frac{l^c}{2} \rfloor h$ hop threshold where l^c refers to the length of the cycle. So, in total it can be at most $h + \lfloor \frac{l^c}{2} \rfloor h = \lfloor \frac{2+l^c}{2} \rfloor h$. Note that in order to calculate an upper bound on the hop threshold of relational acyclification we need to assume the maximum length of any cycle, l^c in the given relational model.

4.2.3 Soundness and Completeness of RCD for Cyclic Relational Causal Models

We consider RCD as a mapping \mathcal{P}_{RCD} from independence models (on variables V) to DPAGs (with vertex set V), which maps the independence model of a σ -AGG G to the DPAG $\mathcal{P}_{RCD}(IM_\sigma(G))$. We assume the following:

A 9. *There exists one or more valid relational acyclifications with hop threshold not exceeding the hop threshold of the given relational causal model ($h' = h$).*

A 10. *The degree of any entity in the relational skeleton is greater than one.*

Assumption 9 follows from Assumption 6 and also limits the set of relational causal models for which RCD can be shown to be sound and complete. Assumption 10 satisfies the soundness and completeness of σ -AGG (Ahsan et al., 2022a).

Theorem 6. *Considering Assumption 8, 9, 10 and causal sufficiency holds, RCD is*

- (i) *sound: for all σ -AGGs G , $\mathcal{P}_{RCD}(IM_\sigma(G))$ contains G ;*
- (ii) *arrowhead complete: for all σ -AGGs G : if $i \notin AN_{\tilde{G}}(j)$ for any DCG \tilde{G} that is σ -Markov equivalent to G , then there is an arrowhead $j \circ \rightarrow i$ in $\mathcal{P}_{RCD}(IM_\sigma(G))$*
- (iii) *tail complete: for all σ -AGGs G , if $i \in AN_{\tilde{G}}(j)$ in any DCG \tilde{G} that is σ -Markov equivalent to G , then there is a tail $i \rightarrow j$ in $\mathcal{P}_{RCD}(IM_\sigma(G))$;*
- (iv) *Markov complete: for all σ -AGGs G_1 and G_2 , G_1 is σ -Markov equivalent to G_2 iff $\mathcal{P}_{RCD}(IM_\sigma(G_1)) = \mathcal{P}_{RCD}(IM_\sigma(G_2))$*

in the σ -separation setting given sufficient hop threshold.

Proof. The main idea of the proof is very similar to the proof of Theorem 1 from Mooij and Claassen (2020) where they prove the soundness and completeness of FCI for cyclic models under σ -separation.

To prove soundness, let G be a σ -AGG and $\mathcal{P} = \mathcal{P}_{RCD}(IM_\sigma(G))$. The acyclic soundness of RCD means that for all AGGs G' , $\mathcal{P}_{RCD}(IM_\sigma(G'))$ contains G' . Hence, by Definition 12 and Assumption 9, \mathcal{P} contains G' for all acyclifications G' . But then \mathcal{P} must contain G which can be easily shown using Proposition 3 of Mooij and Claassen (2020).

To prove arrowhead completeness, let G be a σ -AGG and suppose that $i \notin AN_{\tilde{G}}(j)$ in any DCG \tilde{G} that is σ -Markov equivalent to G . Since G' is σ -Markov equivalent to G , this implies in particular that for all AGGs \tilde{G} that are d -Markov equivalent to G' , $i \notin AN_{\tilde{G}}(j)$. Because of the acyclic arrowhead completeness of RCD, there must be an arrowhead $j \ast \rightarrow i$ in $\mathcal{P}_{RCD}(IM_{\sigma}(G')) = \mathcal{P}_{RCD}(IM_{\sigma}(G))$. Tail completeness is proved similarly.

To prove Markov completeness: Definition 12 and Proposition 1 imply both $IM_{\sigma}(G_1) = IM_d(G'_1)$ and $IM_{\sigma}(G_2) = IM_d(G'_2)$. From the acyclic Markov completeness of RCD¹, it then follows that G'_1 must be d -Markov equivalent to G'_2 , and hence G_1 must be σ -Markov equivalent to G_2 . \square

The statement of this theorem can be seen as a special case of the generalization claim (Theorem 2) by Mooij and Claassen (2020). There is an important point to discuss Assumption 9. Even though Assumption 9 limits the scope of possible relational causal models, it is possible to modify RCD in a way so that it can work for models with relational acyclification having a hop threshold higher than the hop threshold of the given model ($h' > h$). The intuition here is that the skeleton building process should consider this new hop threshold h' (which is upper bounded by $\lfloor \frac{2+l^c}{2} \rfloor h$) rather than the true hop threshold h . However, it requires further proof of soundness and completeness with this modified skeleton. We leave this for future work.

¹Since relational d-separation is equivalent to the Markov condition and it is sound and complete on abstract ground graph (Maier et al., 2013b)

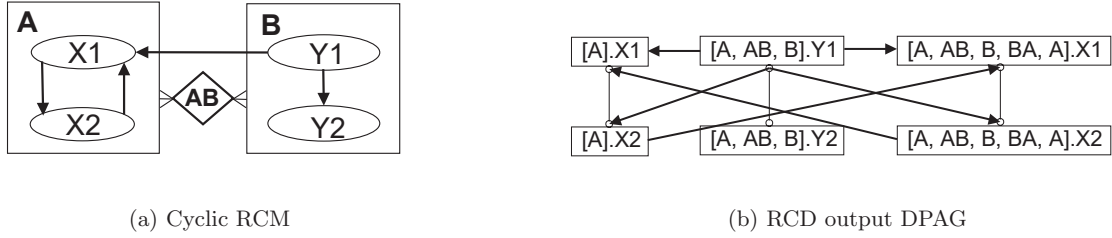


Figure 10: An example cyclic relational model and its corresponding DPAG output by RCD under σ -separation.

4.2.4 Identification of Relational (Non-)cycles

Mooij and Claassen (2020) show that the patterns in strongly connected components in DCGs can be used as a sufficient condition for identifying the absence of certain cyclic causal relations in a complete DPAG. Given Definition 12, the same condition holds for relational models and σ -AGGs as well. We present the necessary and sufficient conditions for identifying non-cycles in the output of RCD following Proposition 10 by (Mooij and Claassen, 2020):

Proposition 4. *Let G be a σ -AGG and denote by $\mathcal{P} = \mathcal{P}_{RCD}(IM_{\sigma}(G))$ the corresponding complete DPAG output by RCD. Let $i \neq j$ be two nodes in \mathcal{P} . If there is an edge $i \circ \rightarrow j$ in \mathcal{P} , and all nodes k for which $k \ast \rightarrow i$ is in \mathcal{P} also have an edge of the same type $k \ast \rightarrow j$ (i.e., the two edge marks at k are the same) in \mathcal{P} , then there exists a DCG \tilde{G} with $j \in SC_{\tilde{G}}(i)$ that is σ -Markov equivalent to G , but also a DCG H with $j \notin SC_H(i)$ that is σ -Markov equivalent to G .*

In other words, under the conditions of this proposition, it is not identifiable from \mathcal{P} alone whether j and i are part of a causal cycle, but they are candidates of being part of a cycle. Figure 10 shows an example of this identifiability criteria. Figure 10b shows the output DPAG of an example cyclic RCM from Figure 10a. The edges between nodes $[A].X1$, $[A, AB, B].Y1$ and $[A, AB, B].Y1$, $[A, AB, B, AB, A].X1$ satisfies the conditions given in Proposition 4. It means they could be part of a cycle but it is not possible to confirm that based on the output alone.

4.3 Experiments

In this section, we examine the effectiveness of RCD for cyclic RCMs using both synthetically generated cyclic RCMs satisfying relational acyclification criteria and a demonstration with a real-world dataset. Since there is no other algorithm designed to discover cyclic RCMs, we compare against the vanilla RCD with d -separation oracle.

4.3.1 Experimental Setup

We follow the procedure introduced by Maier et al. (2013a) for synthetic experiment except we allow feedback loops in the model. We generate 100 random cyclic causal models over randomly generated schema for each of the following combinations: entities (1–3); relationships (one less than the number of entities) with cardinalities selected uniformly at random; attributes per item drawn from $\text{Pois}(\lambda = 1) + 1$; and the number of relational dependencies (4, 6, 8, 10, 12) limited by a hop threshold of 2 and at most 3 parents per variable. We enforce a feedback loop among the dependencies. Note that a single feedback loop can introduce arbitrary length cycles based on the structure of the model. This procedure yields a total of 15,000 synthetic models. Note that this generates simple Bayesian networks when there is a single entity type. We refer to the version of RCD with d -separation and σ -separation oracles as d -RCD and σ -RCD respectively.¹

¹Code available at <https://github.com/edgeslab/sRCD>

4.3.2 Evaluation Criteria

The goal of the evaluation is to compare the learned causal models with the true causal models. However, the output object for cyclic RCMs is PAGs instead of CPDAGs. Moreover, it is expected that the skeleton of the output PAG might be different from the true causal model. For this reason, we evaluate the algorithms based on ancestral relationships. We identify the ancestral relationships entailed by the output and the σ -AGG of the true model and report what percentage (recall) of the actual ancestral relationships are contained in the output. For a sound and complete algorithm, we expect to see 100% recall. We omit precision since we are only comparing to the true model, not all the models in the equivalence class. Moreover, we consider the identification criterion given in Proposition 4 and evaluate the algorithms based on their ability to correctly identify edges as possible cycle candidates. We report recall for this evaluation as well.

4.3.3 Results

Figure 11 shows the comparison of d -RCD and σ -RCD based on *isPossibleAncestor* (top row) and *isPossibleCycle* (bottom row) queries on synthetically generated relational models. The columns represent the increased number of entity types (left to right). The x-axis shows the number of dependencies and y-axis shows recall values. In the leftmost column, we see the results for single entity models. The top left and bottom left figures are equivalent to running the PC algorithm with d - and σ -separation oracles respectively. The rest of the figures represent proper relational models. As expected, we see 100% recall for σ -RCD in all these plots. However, the result for d -RCD shows some intuitive patterns. For a single entity, d -RCD

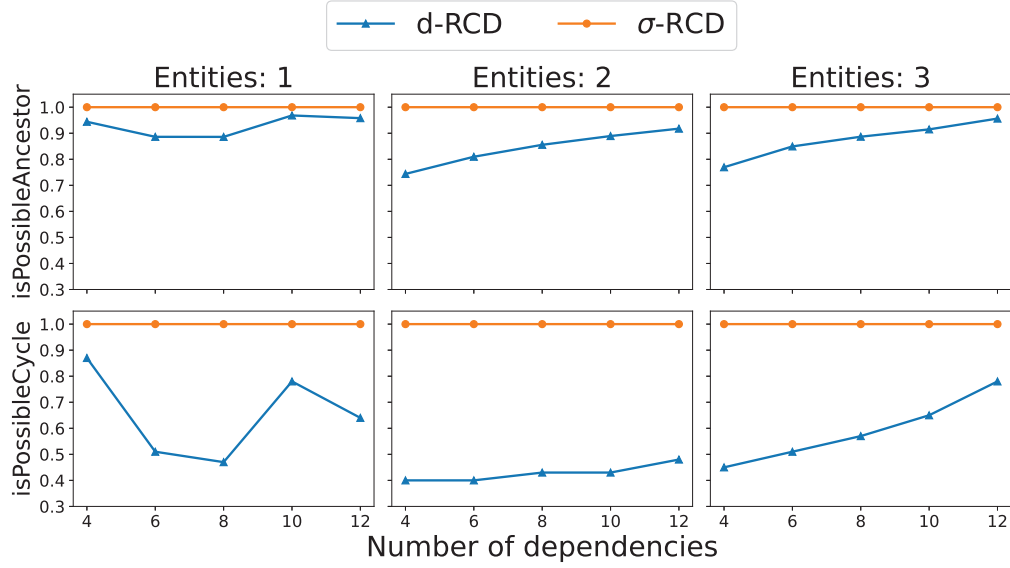


Figure 11: Comparison of d -RCD and σ -RCD based on the recall of *isPossibleAncestor* (top row) and *isPossibleCycle* (bottom row) queries. The number of entity types increased from left to right.

suffers most for 6 and 8 dependencies and get relatively better recalls on lower and higher extremes in x-axis. On the other hand, for multiple entity relational models, we see a general upward trend from left to right which is intuitive since higher number of dependencies makes the models increasingly denser. The difference in the trend between non-relational and relational cases for the low number of dependencies is due to the nature of relational data. Because of multiple entities and overlapping relational paths, there are usually more nodes in a σ -AGG than a DCG with the same number of dependencies.

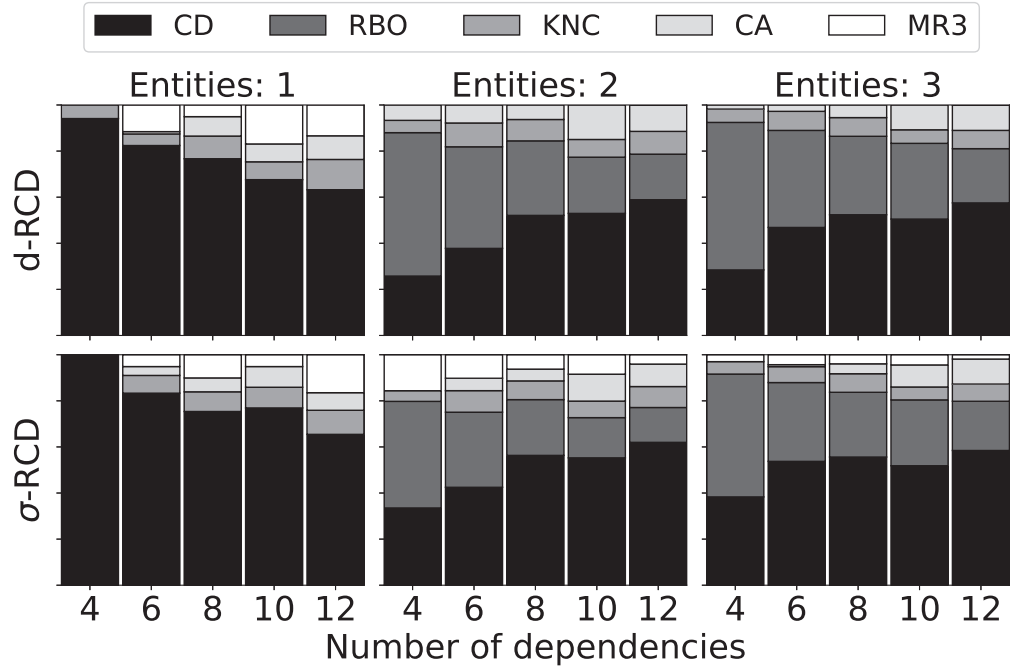


Figure 12: Frequency of edge orientation rules for d -RCD (top) and σ -RCD for different numbers of entity types and dependencies.

Figure 12 shows the percentage of orientation rules used for d -RCD (top row) and σ -RCD (bottom row). The leftmost column refers to the single entity case where no RBO is in effect. We can see some subtle differences in the distribution of rules for d -RCD and σ -RCD. For the small number of dependencies (i.e. 4) only CD (collider detection) rule activates with σ -RCD where d -RCD utilizes both CD and KNC (known non-collider). The increased number of dependencies shows the difference in the overall distribution. For the middle and right column,

a significant difference is seen in the percentage of times rule MR3 (Meek rule 3) is executed for σ -RCD compared to d -RCD. These differences indicate that the algorithms learn fundamentally different structures.

4.3.4 Demonstration on Real-world Data

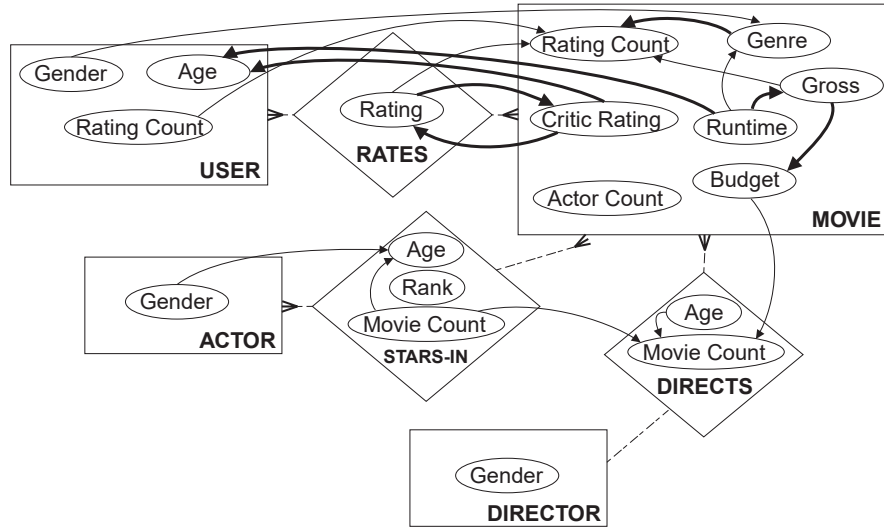


Figure 13: A possible cyclic relational model of MovieLens+ based on the output of RCD (Maier et al., 2013a).

Maier et al. (2013a) show the output of RCD on a sample of MovieLens dataset (www.grouplens.org) based on an approximate conditional independence test using the sig-

nificance of coefficients in linear regressions ¹. Their output contains undirected edges which are potential candidates for cycle edges. Figure 13 shows a possible cyclic relational model which corresponds to the original output. Following Proposition 4, we can infer that the edge between *[Movie].Rating Count* and *[Movie].Genre* cannot be part of any cycles or feedback loops. Some undirected edges can be oriented based on domain knowledge (i.e. Budget can cause gross income but not the other way around). There exist many possible orientations of dependencies that agree with the RCD output. We show one plausible case with a feedback loop between *user ratings* and *critic ratings* of a movie. It is possible that rating information is public and users and critics influence each other with their ratings.

4.4 Discussion

Despite several methods developed for cyclic causal discovery from i.i.d samples, no such algorithm exists for cyclic relational causal models even though cycles are ubiquitous in real-world relational systems. In this work, we investigate the necessary conditions for discovering cyclic relational causal models from observational samples. We introduce relational acyclification operation which facilitates the theoretical guarantees for identifiability of such models. We prove that an existing state-of-the-art relational discovery algorithm, RCD is sound and complete for cyclic relational models for which valid relational acyclification exists. To the best of our knowledge, this discovery is the first of its kind. We hope that this work will play an important role in the study of mutual influence and interference in complex relational systems.

¹The original output is given in Figure 5

CHAPTER 5

RELATIONAL DEPENDENCE TEST

Parts of this chapter were previously published as Ahsan, R., Fatemi, Z., Arbour, D., and Zheleva, E.: Non-parametric Inference of Relational Dependence. In Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022) Ahsan et al. (2022b)

Measuring statistical dependence is a fundamental task in statistics. However, most existing independence tests assume that the observed data is independent and identically distributed (i.i.d.). This makes them unsuitable for capturing the dependencies in real-world relational systems, from social networks to protein-protein interactions, in which data instances depend on each other. In this chapter, we characterize the notion of statistical dependence in relational data. Furthermore, we provide a non-parametric consistent method to operationalize the test for relational dependence. We empirically evaluate our proposed method on a variety of synthetic and semi-synthetic networks and demonstrate its effectiveness compared to the state-of-the-art kernel-based relational independence test.

5.1 Relational Dependence

Relational dependence refers to a statistical dependence, either marginal or conditional, between two variables where at least one of the variables is relational. The goal of a relational dependence test is to determine whether to reject the null hypothesis of independence between these variables or not. The representation of relational data for such a test is non-trivial because

data instances are not i.i.d. A common practice to deal with relational data is *propositionalization* (Kramer et al., 2001), which refers to the process of projecting a set of connected data samples down to a single, propositional table. In the context of relational dependence testing, flattening has three main deficiencies. First, the entities in the flattened data are not i.i.d. Second, choosing the appropriate aggregation function is non-trivial as discussed in Chapter 1. Failing to appropriately define the aggregate, in this case, could lead to increased type I errors in marginal tests and both type I and II errors for conditional tests. Third, flattening raises statistical concerns for relational causal discovery, one of the application areas of relational conditional independence tests, by violating the causal Markov condition (Maier et al., 2013c). Lee and Honavar (2017) address the first deficiency by proposing a solution framework based on graph kernels using an existing i.i.d. kernel-based CI test method. However, their approach does not directly address the other two concerns.

5.1.1 Example of Relational Dependence

Let's look at the problem with a few concrete examples. We consider an entity Person which exhibits attributes like smoking (S) and drinking (D) that represent a person's smoking and drinking behavior respectively and G represents the network of social ties. We might be interested to detect the impact of social influence on a person's smoking behavior. It can be formalized as both marginal and conditional independence tests based on the choice of a relational variable. For example, detecting whether one's smoking behavior is marginally independent of one's friends' smoking behavior could be carried out by a test of $v_i.S \perp\!\!\!\perp \sigma_S(v_i)$. Similarly, a conditional test of $v_i.S \perp\!\!\!\perp \sigma_S(v_i) | v_i.D$ should be able to detect whether one's smoking behavior

is conditionally independent of neighbors' smoking behavior given one's drinking behavior. We pick three such diverse cases for evaluation and formally introduce those in the experimental evaluation section.

5.1.2 General Relational Independence

In this work, we propose a general notion of relational dependence which captures complex dependencies between relational variables without relying on flattening or explicit aggregate representations. We extend the definition of conditional independence for non-parametric functions by Daudin (1980) and propose the following definitions of marginal and *relational* conditional independence:

Definition 13 (Relational Marginal Independence). *Two relational variables, $\sigma_X(v_i)$ and $\sigma_Y(v_i)$ are said to be marginally independent of each other if and only if, $\mathbb{E}[g_X(\sigma_X(v_i))g_Y(\sigma_Y(v_i))] = \mathbb{E}[g_X(\sigma_X(v_i))]\mathbb{E}[g_Y(\sigma_Y(v_i))]$ for any smooth square measurable functions $g_X(\cdot), g_Y(\cdot)$.*

Definition 14 (Relational Conditional Independence). *Two relational variables, $\sigma_X(v_i)$ and $\sigma_Y(v_i)$ are said to be independent of each other given a third, $\sigma_Z(v_i)$ if and only if, $\mathbb{E}[g_X(\sigma_X(v_i))g_Y(\sigma_Y(v_i))|g_Z(\sigma_Z(v_i))] = \mathbb{E}[g_X(\sigma_X(v_i))|g_Z(\sigma_Z(v_i))]\mathbb{E}[g_Y(\sigma_Y(v_i))|g_Z(\sigma_Z(v_i))]$ for any smooth square measurable functions $g_X(\cdot), g_Y(\cdot), g_Z(\cdot)$.*

Here, $g_X(\cdot), g_Y(\cdot), g_Z(\cdot)$ are *aggregate* functions that map σ to a real-valued vector. They could be *sum*, *mean* or any other complex non-linear function. The rejection of the null hypothesis of marginal independence would mean that the variables are possibly dependent, either due

to a directed path between them or due to a direct, causal relationship, or the presence of a confounding relationship. For a relational conditional independence (RCI) test, the rejection of the null hypothesis would imply that the two variables are not independent given the conditioning set. Note that because we are considering the dependence between sets of relational variables and their propositional counterparts we circumvent the three problems with flattening described earlier.

5.2 Relational Dependence Test

In this section, we discuss the components which operationalize the definition of relational dependence into an empirical test. We first describe a non-parametric relational aggregate formed by local kernel means. Then we formulate marginal and conditional independence tests using the kernel mean embedding. Then, we discuss the theoretical boundaries for the consistency of the proposed test. Finally, we introduce techniques for large-scale approximation of the proposed relational kernels that can speed up the independence test significantly. We make the following assumptions for the proposed approach:

A 11. *Each node $v \in V$ has a degree of at least 1.*

A 12. *The adjacency matrix of G is symmetric with edge weights bounded by some real constant.*

A 13. *Dependence between two instances i and j implies the existence of a path in the graph between v_i and v_j .*

5.2.1 Non-parametric Aggregate Representations

One of the central problems in estimating dependence in relational settings is defining a sufficient representation for the sets of observations for individual instances of a relational variable. Prior work (Maier et al., 2013a; Arbour et al., 2016b; Lee and Honavar, 2017) considered aggregation functions, usually one, which are specified *a priori* by the practitioner. However, in many scenarios, it is unreasonable to expect practitioners to reason over a very complex joint distribution or to know the exact parametric form of dependence. For example, the possible aggregation in effect of the spread of obesity on social networks (Christakis and Fowler, 2007) can be different from people’s influence on the Twitter platform (Bakshy et al., 2011). A generalized definition and associated operationalization of relational dependence can help the practitioner by directly measuring dependence without prior domain knowledge about aggregations on the given relational system.

Adopting the kernel mean as an aggregation function removes the burden of reasoning over parametric families and predefined aggregates. Specifically, the kernel mean embedding considers the mean of a variable after applying a projection $\phi(\cdot)$ into some RKHS, $\mu = \int \phi(x)p(x)dx$, with the corresponding empirical estimate of $\hat{\mu} = \frac{1}{N} \sum_i^N \phi(x_i)$ where N is the number of observations and x_1, \dots, x_N are observations from a random variable X (Smola et al., 2007).

We present the practical implementation of the kernel mean as a relational aggregate. For a given node v_i , we define the kernel mean aggregate of its neighbors with respect to the attribute

$$X \text{ as } \mu(v_i) = \frac{1}{\deg(v_i)} \sum_{m \in \hat{\mathcal{N}}(v_i)} \phi(m.x)$$

where $\hat{\mathcal{N}}(\cdot)$ refers to a path predicate that is restricted to immediate neighbors for ease of exposition. Because ϕ may map to an infinite dimension, it is impractical to explicitly represent this quantity.

Fortunately, because our statistics of interest are concerned with the covariance, the kernel trick, i.e. considering the inner product rather than the feature representations directly, can be employed. Specifically, the inner product between relational kernel mean is given as

$$\langle \mu(v_i), \mu(v_j) \rangle = \frac{\sum_{m \in \hat{\mathcal{N}}(v_i)} \sum_{p \in \hat{\mathcal{N}}(v_j)} k(m.x, p.x)}{\deg(v_i) \deg(v_j)},$$

which can be written for an entire sample in terms of a matrix product between the network adjacency matrix, A , the inverse degree matrix D^{-1} where $D_{i,i} = \frac{1}{\deg(v_i)}$, and the kernel matrix K_X , by observing $(D^{-1}A\phi(\mathbf{x}))(D^{-1}A\phi(\mathbf{x}))^T = D^{-1}AK_XAD^{-1}$.

In contrast to the propositional kernel mean, the convergence of the relational to its population counterpart is not necessarily guaranteed because of sample dependence. We discuss convergence and consistency guarantees under the assumption of weak dependence after describing the relational independence tests.

5.2.2 Relational Marginal Independence Test

With the relational kernel mean defined we now turn to the central task of this work, non-parametric inference of relational dependence (NIRD). As a test statistic, we use the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005). HSIC measures the maximum distance between an embedding of the observed joint distribution, and the product of the

marginals, i.e., $\|\mathbb{E}[\phi(x) \otimes \phi(y)] - \mathbb{E}[\phi(x)] \otimes \mathbb{E}[\phi(y)]\|^2$. We perform a hypothesis test using HSIC as the test statistic where the null hypothesis refers to independence. The test produces a p-value which is used to decide whether to reject the null or not. Testing relational independence using HSIC is straightforward with the relational kernel mean by using the kernel matrix defined earlier in the empirical HSIC estimator. Defining the centering matrix $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, an empirical estimate of HSIC is given by $\frac{1}{n^2}\text{trace}(K_X H K_Y H)$, where K_X and K_Y are kernel matrices corresponding to the random variables X and Y , respectively. Independence testing with HSIC can be performed by using the corresponding relational kernel in the test statistic.

5.2.3 Relational Conditional Independence Test

A similar construction can be employed to test for relational conditional independence, defined in Definition 14. Following Strobl et al. (2019), we consider the following L^2 spaces,

$$\begin{aligned} F_{XZ} &\triangleq \left\{ \tilde{f} \in L_{XZ}^2 \mid E(\tilde{f} \mid Z) = 0 \right\} \\ F_{YZ} &\triangleq \left\{ \tilde{g} \in L_{YZ}^2 \mid E(\tilde{g} \mid Z) = 0 \right\} \\ F_{Y \cdot Z} &\triangleq \left\{ \tilde{h}' \mid \tilde{h}' = h'(Y) - E(h' \mid Z), h' \in L_Y^2 \right\} \end{aligned}$$

Each of these quantities can easily be constructed by considering regressions, e.g. \tilde{f} can be obtained by taking the residuals after performing a regression. We consider a mean of the feature basis representation as an aggregation function whenever one of the variables is relational. Under the assumption that the direct sum of the reproducing kernel Hilbert spaces, $k_x k_y$ and k_z is dense in L_2 , Strobl et al. (2019) (proposition 5) showed that conditional linear covariance

of zero implies uncorrelatedness, i.e., $\mathbb{E}[\tilde{f}\tilde{g}] = 0 \implies X \perp\!\!\!\perp Y|Z \implies \Sigma_{XY|Z} = 0$. This motivates the use of a multiple output kernel ridge regression as an estimator of the conditional expectation, $\beta = (\phi(z)^T \phi(z) + \lambda I)^{-1} \phi(z)^T \phi(\ddot{x})$ where $\ddot{X} \triangleq (X, Z)$ is the concatenation of x and z . Informally, this can be seen as applying the “kernel trick” of considering linear operations on non-linear transformations of the data allowing for observations to be dependent. The test is then constructed by considering the residuals, $\widetilde{\phi(\ddot{x})} = \phi(\ddot{x}) - \phi(z)\beta_{\ddot{x}z}$, $\widetilde{\phi(y)} = \phi(y) - \phi(z)\beta_{yz}$ and sum of the squared covariances between them. The final form of the test is given by $\frac{1}{n^2} \text{trace} \left(\widetilde{K_{\ddot{X}}} H \widetilde{K_Y} H \right)$ where $\widetilde{K_{\ddot{X}}}$ and $\widetilde{K_Y}$ refers to the kernel matrices for the residuals $\widetilde{\phi(\ddot{x})}$ and $\widetilde{\phi(y)}$ respectively.

There are two considerations in employing this procedure in a relational setting, namely how to handle relational variables in the conditioning set (Z) and the test set (X), respectively. When a member of the conditioning set is relational, the test procedure is identical after replacing $\phi(z)$ with its relational counterpart, $\frac{1}{|\hat{\mathcal{N}}(z)|} \sum_{m \in \hat{\mathcal{N}}(z)} \phi(m)$. When a member of the test set is relational, the problem is reduced to predicting each member of the set independently by considering the regression of the perspective of the relational variable, as described by Maier et al. (2013a). After regressing individual members, the mean of residuals is then considered for the marginal tests, $\widetilde{\sigma(\phi(x))} = \frac{1}{|\hat{\mathcal{N}}(x)|} \sum_{m \in \hat{\mathcal{N}}(x)} \phi(m) - \phi(z)\beta_{mz}$.

5.2.4 Consistency of Relational Independence Test

In order to reason about the behavior of test statistics under non-i.i.d. samples and understand asymptotic behavior, we need to characterize the behavior of dependence amongst instances as a function of some notion of distance between instances. There are a number

of formalisms for reasoning about dependent data (Andrews and Pollard, 1994; Bickel and Bühlmann, 1999; Dedecker et al., 2007). In this work we focus on weak dependence (Dedecker et al., 2007), which we describe next.

5.2.4.1 Weak Dependence

In order to accommodate dependent observations and maintain consistency of the testing procedure we will assume that observations are weakly dependent. Weak dependence provides a flexible notion of dependence that requires only the definition of distance between instances and the presence of a measurable probability space Arbour (2017). Within this work we will make use of the notion of weak dependence, i.e. τ -dependence.

Definition 15. (Dedecker et al., 2007) Let π be a filtration¹ over the set of nodes in a graph, G , defined by performing a breadth first search at an arbitrary node, $v \in G$. Further, define X to be a \mathbb{L}^p -integrable random variable. The **weak-dependence coefficient** is defined as $\tau_{p,r}(X) = \sup_{(i,j)} \| \sup_g \text{Cov}(g(X_{\pi(i)}), g(X_{\pi(j)})) \|_p$, where $i \leq j$ and $j - i \leq r$, and $g(\cdot)$ is a Lipschitz function.

Intuitively, the weak dependence coefficient, $\tau_{p,r}(X)$ measures the covariance between a vector, X_i and another random vector X_j drawn from the same process separated by at least distance of r . We call a process *weakly dependent* if τ tends to zero as the r tends to infinity. Note that this is a strictly weaker condition than alternative assumptions on dependence such

¹A filtration is an ordering of a set such that for any two subsets, $S_{1,\dots,j}, S_{1,\dots,k}$, $j \leq k \rightarrow S_{1,\dots,j} \subseteq S_{1,\dots,k}$.

as strong mixing and m -dependence which require independence at a finite distance, whereas weak-dependence only requires it asymptotically.

5.2.4.2 Weak Dependence in Relational Domains

We provide a natural extension of weak dependence within the relational setting by replacing usual definition of distance to the shortest path distance between two nodes in a graph. The role of τ in this case can be interpreted as measuring the decay of dependence between instances as a function of shortest-path distance. We will assume from here out that as the distance between any two nodes in the network tends to infinity, the dependence between them converges to zero. More formally, we will employ the following assumption similar to Arbour (2017):

A 14. *$(X_t)_{t \in \pi}$ is a strictly stationary τ -dependent process with $\sum_{r=1}^{\infty} r^2 \sqrt{\tau_r(X)} \leq \infty$ for some filtration π , where r is shortest-path graph distance.*

The notion of weak dependence within the network setting is not novel to this work, Xiang and Neville (2011) make use of the τ -coefficient in the context of deriving asymptotic consistency for transductive learning with an assumption of linear dependence amongst instances. In this work, I consider weak dependence with arbitrary dependence for independence testing of relational data. Consistency of the relational independence testing is provided by the following theorem and corollary, after applying two additional assumptions.

A 15. *The maximum degree of any node in the network is bounded by a real constant.*

A 16. *The network structure is fixed and doesn't change during the generation of the observed random variables.*

Assumption 15 ensures that the average shortest path distance from any node to all other nodes in the graph tends to infinite as the number of nodes tends to infinite, which is necessary in order to have convergence of weakly dependent sequences. Assumption 16 ensures that the observed neighborhoods for nodes correspond to the structure which generated the data.

Theorem 7. *Under the aforementioned assumptions the Hilbert-Schmidt independence criterion of two weakly dependent propositional variables converges in L_1 to its population counterpart, i.e., $|HSIC_n - HSIC_{population}| \xrightarrow{d} 0$.*

I present the proofs of consistency for HSIC and relational HSIC under weak dependence. The approach here is to extend the results of Chwialkowski et al. (2014) and (Leucht and Neumann, 2013), who analyze degenerate U and V -statistics (which includes HSIC as a specific instantiation) under weak dependence in spaces that admit euclidean distances to the more general setting of graph structured spaces. Much of the results carry through after modifications to accommodate the fact that the number of reachable instances at a specific distance is irregular. Arbour (2017) present a modification of the relevant proof which shows the convergence of the distribution of degenerate V -statistics. Here, I describe the application of that proof to our setting and the extension to relational variables.

5.2.4.3 Relational Marginal Independence Test

We now turn our attention to the Hilbert-Schmidt independence criterion for relational marginal independence test.

Theorem 8. *Under the aforementioned assumptions the Hilbert-Schmidt independence criterion of two weakly dependent propositional variables converges in L_1 to its population counterpart, i.e., $|\overline{HSIC_n} - HSIC_{population}| \xrightarrow{d} 0$.*

Proof. Recall that the Hilbert-Schmidt independence criterion (HSIC) is a test of dependence, i.e. a hypothesis test of paired samples where the null hypothesis is that the two samples are generated independently, $\mathbb{P}_{x,y} = \mathbb{P}_x \mathbb{P}_y$. Our focus is on the empirical estimator of HSIC, which can be written as degree-four V -statistic with a *core*¹ defined by:

$$h(x_1, x_2, x_3, x_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})k(y_{\pi(1)}, y_{\pi(2)}) + k(y_{\pi(3)}, y_{\pi(4)}) - 2k(y_{\pi(2)}, y_{\pi(3)}) \quad (5.1)$$

where k is the relational kernel and S_n is the set of permutations over a set of n elements. Convergence then follows as a direct application of Theorem 1 by Arbour (2017) and the weak law of large numbers. \square

Corollary 2. *Under the aforementioned assumptions the Hilbert-Schmidt independence criterion between a weakly relational and a weakly dependent propositional variable converges in L_1 to its population counterpart, i.e., $|HSIC_n - HSIC_{population}| \xrightarrow{d} 0$.*

Proof. The central items to be shown in order to apply the results of Theorem 1 by Arbour (2017) are (1) relational kernels define a valid V -statistic, and (2) the relational variable remains weakly-dependent. Item (1) follows directly by denoting one of the variables in equation

¹In order to prevent confusion, we follow Chwialkowski et al. (2014) and do not follow the canonical convention of calling h the kernel.

Equation 5.1 to be a set of instances return by the path predicate and k to be the relational kernel defined in the main text. Item (2) follows as a consequence of assumption 5 which bounds the degree of each node by a finite constant, c . As a result, any path predicate which defines a finite length path will return a set no larger than $c < c' < \infty$. As a result, so long as the initial random variable is weakly dependent, the relational variable constructed from the initial random variable will also be weakly-dependent, albeit with a slower rate of convergence since the coefficient τ_r (the weak dependence coefficient) will necessarily decay more slowly. \square

It is important to note that Theorem 8 and Corollary 2 show convergence in distribution but do not claim any guarantees regarding the rates of convergence with respect to the number of nodes and level of dependence. The rate of convergence will depend on the weak dependence coefficient. In the case that the coefficient is 0, this reduces to results that correspond to prior work on iid data (Zhang et al., 2011). While there is prior work studying this in more restrictive assumptions on the dependence between instances (London et al., 2013), we are not aware of similar results for the case of weak dependence in general structured domains even in the simpler case of regression. This would be an important direction for future work.

5.2.5 Large Scale Approximations

While the proposed model is theoretically appealing, the associated time and space complexity render it infeasible for most modern network settings. To address this, we appeal to an approximation of the kernels known as Random Fourier Features (Rahimi and Recht, 2008). Random Fourier Features exploit Bochner’s theorem, which states that a continuous, time-invariant kernel is positive definite if and only if the kernel is the Fourier transform of some

non-negative measure. For example the Gaussian kernel can be represented with the following Fourier transformation $\hat{k}(\omega) = \frac{1}{2\pi} \int e^{-j\omega^\top \delta} k(\delta) d\delta$. This property implies that a kernel can be approximated via the following procedure:

- Draw D samples, from some distribution (i.e Normal), to approximate the Gaussian kernel where the variance σ corresponds to the bandwidth of the kernel.
- Construct the Fourier basis explicitly as $z(x) = \sqrt{\frac{2}{d}} [\cos(w_1^T x), \sin(w_1^T x), \dots]$.
- Perform linear operations using z .

Following (Zhang et al., 2018; Strobl et al., 2019), we approximate HSIC using random Fourier features by considering $\widehat{\text{HSIC}}(X, Y) = \|\frac{1}{n} Z_X^T H Z_Y\|^2$ where Z is a $n \times d$ dimensional matrix with each row consisting of the random Fourier features for an observation. We can represent the relational kernel mean as $D^{-1}AZ$, and the corresponding test statistic as $\|\frac{1}{n} Z_X^T A D^{-1} H Z_Y\|$ where D and A are the diagonal degree and adjacency matrix as before. In several experiments we show that using approximate statistic leads to significant performance improvements with minimal effect on the efficacy of the test, even with only a few random features.

5.2.6 Extension to Multi-relational Systems

In our problem definition we assumed a single-entity, single relationship relational schema for ease of exposition. Here, we discuss necessary extensions for a multiple entity, multi-relational system. We consider a set of item classes \mathcal{I} to be the union of entities and relationship classes, $\mathcal{I} = \mathcal{E} \cup \mathcal{R}$, following prior work (Lee and Honavar, 2017; Maier et al., 2013a). We refer to the attribute class of an item class $I \in \mathcal{I}$ as $\mathcal{A}(I)$. Moreover, let $G(I)$ denote a set of items of an item class $I \in \mathcal{I}$.

Here, we point out two major differences in a multi-relational system:

1. The relational dependence is specifically defined between two item classes $I \in \mathcal{I}$ and $J \in \mathcal{J}$.
2. The path predicate ρ is likely to be defined with relational queries rather than random walks over a neighborhood.

Now, we revisit definition 1 from the main text with the new notation as follows:

Definition 16 (Relational Variable). *Given a relational schema $\mathcal{S} = \langle \mathcal{E}, \mathcal{R}, \mathcal{A} \rangle$, its instantiation G , two item classes $I, J \in \mathcal{I}$ and a path predicate ρ , a relational variable $\sigma(v_i, \mathbf{X}, G, \rho)$ is the set of attributes $v_j.\mathbf{X}$ selected by ρ of items $v_j \in G(J)$ reachable from items $v_i \in G(I)$ such that $\mathbf{X} \subset \mathcal{A}(J)$, where the path predicate ρ is a function given by:*

$$\rho(v_i, G) : G(I) \mapsto \mathcal{P}(G(J))$$

The necessary assumptions and relational dependence definitions still hold. The major difference arises in the compact representation of the relational kernel. Equation 1 stays valid with an updated notion of path predicate. However, the compact representation in equation 2 is no longer trivial since the adjacency matrix A is no longer directly applicable. There are two potential workarounds. First, since the compact representation is not mandatory for our method to work, we can still work with equation 1 for multi-relational systems. Second, we can essentially consider the bipartite graph between sets of items between item classes $I, J \in \mathcal{I}$ and

use the adjacency matrix A_{IJ} of this bipartite graph instead of A . Similarly a corresponding degree matrix D_{IJ} can be constructed from A_{IJ} .

5.3 Experimental Evaluations

To evaluate the effectiveness of the proposed test, we run experiments with multiple network datasets, relational dependence cases, and synthetic attribute generators.

5.3.0.1 Network Datasets

We consider networks from two synthetic graph generators and three non-PII real-world networks. First, for the Barabási-Albert (BA) model, we vary the parameter that controls the number of nodes a new node can attach to. For the Erdős-Rényi (ER) model, we vary the probability of edge creation between each pair of nodes. For each set of parameters, we generate 100 networks with size 100. The small size of the synthetic networks is driven by the baseline method which does not scale well, as shown in Figure 17a. We also demonstrate the applicability of our approach through a Facebook ego-network with 4,039 nodes and 88,234 edges (Leskovec and Mcauley, 2012) and a Twitter ego-network with 11,176 nodes and 1,44,653 edges where we sampled a subgraph of 10,000 nodes (Leskovec and Mcauley, 2012). We also demonstrate results on the 50 Women dataset (Michell and Amos, 1997). This dataset has the smoking, sport, drug, alcohol consumption habits of 50 female students, along with their friendship information, over the course of three years.

5.3.1 Four Cases of Relational (In)dependence

We choose three representative relational dependence cases and one relational independence case to cover a range of possible tests. We consider attributes $Z, X, Y \in \mathcal{A}$ which measure characteristics in time steps $t-1, t, t+1$ respectively. All the cases are represented with arrows showing the direction of dependence:

1. **Case 1:** $\sigma_X(v_i) \rightarrow v_i.Y$
2. **Case 2:** $\sigma_X(v_i) \leftarrow v_i.Z \rightarrow v_i.Y \leftarrow \sigma_X(v_i)$
3. **Case 3:** $v_i.X \leftarrow \sigma_Z(v_i) \rightarrow v_i.Y \leftarrow v_i.X$
4. **Case 4:** $\sigma_X(v_i) \leftarrow v_i.Z \rightarrow v_i.Y$

where $\sigma_X(v_i)$ and $\sigma_Z(v_i)$ are relational variables on the attributes X and Z of the direct neighbors of v_i . Case 1 refers to marginal independence between a relational and a propositional variable ($\sigma_X(v_i) \perp\!\!\!\perp v_i.Y$). Cases 2 and 3 introduce conditional independence given a confounder. Case 2 refers to a propositional confounder ($\sigma_X(v_i) \perp\!\!\!\perp v_i.Y | v_i.Z$) whereas case 3 refers to a relational confounder ($v_i.X \perp\!\!\!\perp v_i.Y | \sigma_Z(v_i)$). A test should be able to reject the null hypothesis of no dependence in the first three cases. Case 4 represents conditional independence and the test should not reject the null hypothesis and it should produce high errors. Note that direction is ignored in the test.

5.3.1.1 Synthetic Attribute Generation

Here, we describe the synthetic attribute generation procedure for the three cases mentioned in the main text. Note that, only the generation of $v_i.Y$ differs in null and alternate hypothesis while others stay the same. We consider polynomial dependency model for most of our experiments. $v_i.X$ for case 0 and $v_i.Z$ for cases 1,2 is drawn from a uniform distribution $U(0, 1)$ while $v_i.X$ is always *binarized* to resemble the effect of treatment assignment. The outcome $v_i.Y$ is generated according to the following equation for marginal dependence (case 0):

$$v_i.Y \sim \begin{cases} U(0,1) & null \\ \beta_d \cdot (g(\sigma_x(v_i)))^2 + \epsilon & alternate \end{cases} \quad (5.2)$$

Conditional dependence (case 1) is reflected by the following equation:

$$v_i.X \sim \beta_c \cdot (v_i.Z)^2 + \epsilon$$

$$v_i.Y \sim \begin{cases} \beta_c \cdot (v_i.Z)^2 + \epsilon & null \\ \beta_d \cdot (g(\sigma_X(v_i)))^2 + \beta_c \cdot (v_i.Z)^2 + \epsilon & alternate \end{cases} \quad (5.3)$$

Here, β_d and β_c are dependence and confounding coefficients respectively. β_c is considered 1.0 in our experiments. ϵ is noise drawn from standard normal ($N(0,1)$) distribution. g refers to the *mean* aggregate function. We can get the generating function for case 2 by replacing $g(\sigma_X(v_i))$ and $v_i.Z$ with $v_i.X$ and $g(\sigma_Z(v_i))$ respectively in equation Equation 5.3. Next, we consider the following procedure to simulate linear threshold model (LTM) for the diffusion experiment which falls under case 0:

$$T_i \sim U(0,1)$$

$$v_i.x_{t+1} = \mathbb{1}(\text{mean}(\sigma_{x_t}(v_i)) > T_i) \quad (5.4)$$

$$v_i.y_{t+1} = \mathbb{1}(g(\sigma_{x_t}(v_i)) > T_i)$$

where we reassign $v_i.x$ values to simulate each diffusion step based on its value in previous step.

The $v_i.y$ values are assigned based on $v_i.x$ values in the last diffusion step.

5.3.2 Experimental Setup

We empirically evaluate the proposed approach, NIRD, to the state-of-the-art RCI test method, KRCIT (Lee and Honavar, 2017).¹ We report the average Type I and Type II errors with significance level 0.05 over 100 trials for each set of parameters. We use *Radial Basis Function kernel* (RBF) as the base kernel. KRCIT is implemented with HSIC as the kernel-based marginal independence test method and KCIT (Zhang et al., 2011) as the kernel-based conditional independence (CI) test. We use the approximate method of NIRD in all experimental evaluation with 20 and 50 random Fourier features for marginal and conditional test respectively. We estimate the null distribution via permutation on the non-relational variable since the marginal distribution remains unchanged. We compare both RCI methods (NIRD, KRCIT) to a recent i.i.d. CI test method, Sobolov Independence Criterion (SIC) (Mroueh et al., 2019). We study NIRD’s strengths and weaknesses in five experimental setups:

- **Relational dependence sensitivity:** We evaluate the sensitivity of the dependence tests to different relational dependence strengths. We report results on polynomial models while varying the dependence coefficient in range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ for the alternate hypothesis and report both Type I and Type II errors. Edge connectivity is 3 for Barabási-Albert and edge probability is 0.02 for Erdős-Rényi model.
- **Network sensitivity:** We examine performance over a variety of network structures. We vary edge connectivity of Barabási-Albert in range $\{1, 2, 3\}$ and edge probability of

¹Code available at <https://github.com/edgeslab/nird-uai22>

Erdős-Rényi in range $\{0.005, 0.01, 0.015, 0.02, 0.025\}$. For these experiments we use a fixed dependence coefficient value of 0.5.

- **Scalability:** To evaluate the scalability of the proposed method against the baseline, we generate Erdős-Rényi synthetic networks (edge probability 0.02) with varying number of nodes and report the execution time for both NIRD and KRCIT on marginal and conditional independence test. We vary the network size (x-axis) in the range $\{100, 200, 300, 400, 500\}$. This experiment is executed on a 2.4GHz 8-core machine with 50GB memory.
- **Diffusion:** We demonstrate the application of our proposed method on testing for contagion through information diffusion on the semi-synthetic Facebook network. We use a linear threshold model (Granovetter, 1978) to characterize dependency in our attribute generation process and simulate the diffusion process by reassigning the X attributes over several diffusion steps. We generate the Y values based on X generated in the last diffusion step. The initial state of all nodes is 0, then nodes get activated (set to 1) with probability 0.1. We expect the distribution of $\sigma_X(v_i)$ to change with increasing diffusion steps and we want to investigate at what step it is easier to detect relational dependence. We vary the number of steps and sample size to observe the Type II error for our relational dependence test. We also investigate the impact of activation probability on the Type-II error on the Twitter ego-network with 10,000 nodes.
- **Real world demonstration:** We demonstrate the applicability of our test for detecting peer influence in a well-studied real-world social network (*50 Women*) where our test discovers smoking-, drug- and sport-related peer dependencies that concur with previous research.

5.3.3 Results

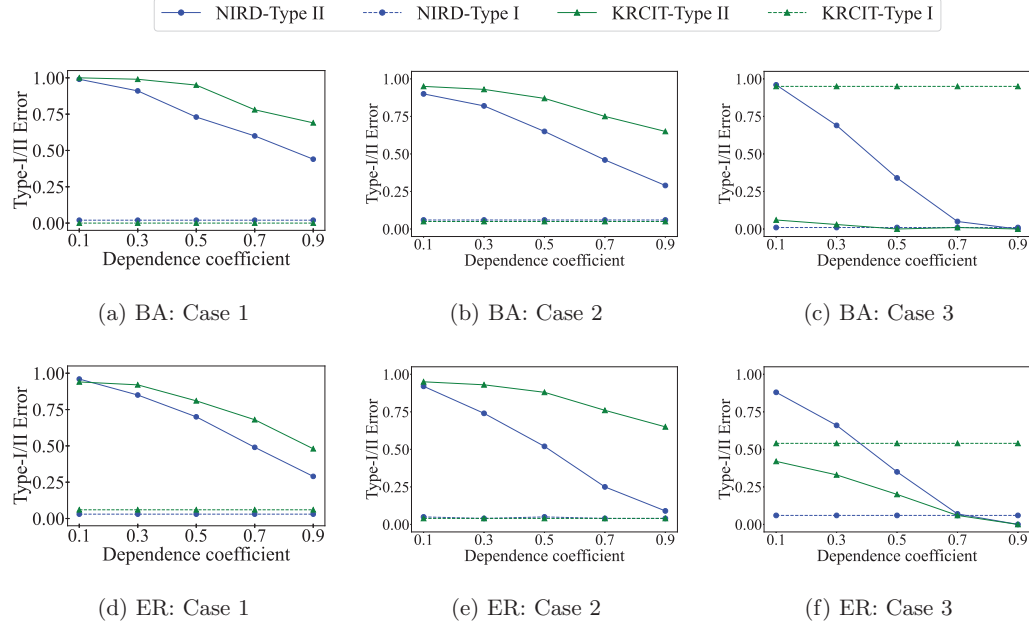


Figure 14: Relational dependence impact on Type I/II errors.

5.3.3.1 Relational Dependence Sensitivity

Figure 14 shows Type I and Type II errors for the polynomial dependency model on synthetic data. The rows correspond to the network models and the columns to relational dependence cases. The solid and dashed lines correspond to Type II and Type I errors respectively. The test is most challenging when the dependence coefficient, β_d (x-axis) is low. The figure shows

that both RCI methods are well calibrated with low Type I error (max 0.06 by KRCIT in Erdős-Rényi) for the first two cases. In these cases, NIRD consistently produces lower Type II errors compared to KRCIT. It is most visible in Erdős-Rényi model (14e) with 86% reduction in Type II error for $\beta_d = 0.9$. The performance gain of NIRD increases slightly from case 1 to case 2 as the difficulty increases. In case 3, KRCIT is poorly calibrated and exhibits an unusually high Type I error. Across cases, NIRD shows desired behavior: it is consistently well-calibrated and its Type II error decreases with the increase of relational dependence. Case 4 provides a sanity check and both methods produce high Type II errors (0.9 to 1.0) with good calibration. The error is nearly constant irrespective of strength of dependence coefficients or network model parameters used. In order to test for sensitivity to noise, we repeat these experiments varying the noise variance over multiple trials instead of drawing from a fixed distribution. The results look very similar. From Figure 15 we can see a slight change of type-II errors compared to Figure 14. However, the trend seems to be very similar.

5.3.3.2 Network Sensitivity

Figure 16 shows Type I and Type II errors for two network models. The x-axis represents the corresponding parameter values for each model. We observe that increased parameter values exhibit higher Type II errors in general for Barabási-Albert model but not for Erdős-Rényi. A possible reason is that Barabási-Albert exhibits a more skewed degree distribution compared to Erdős-Rényi. Note that the increased parameter values indicate higher density of the network. We expect Erdős-Rényi to show a similar trend if the edge probability is further increased. NIRD outperforms KRCIT in terms of Type II error (except in Figures 16c and 16f which is

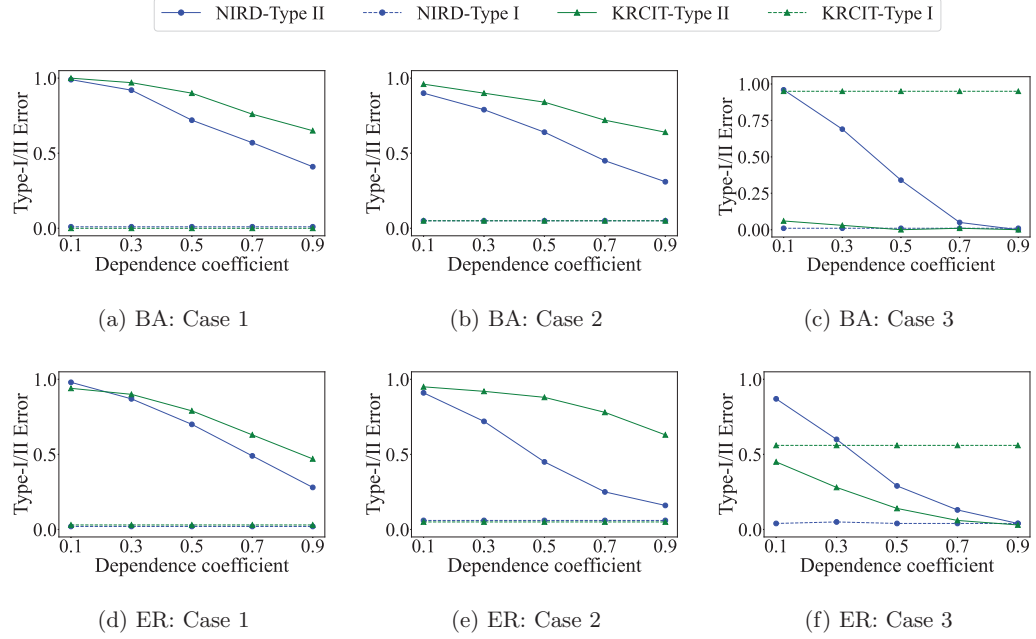


Figure 15: Relational dependence impact on Type I/II errors while variance of noise varied $\sim \mathcal{N}(1, 0.2)$ over multiple trials.

due to poor calibration of the baseline method) irrespective of network density. Type II error is reduced as high as 65% for Erdős-Rényi model with edge probability 0.025 (Figure 16e). Moreover, Type I error for NIRD is consistent whereas KRCIT suffers in case 2 (Figures 16c, 16f).

5.3.3.3 Scalability

Figure 17a shows execution time in minutes (y-axis) for both marginal (case 1) and conditional (case 3) independence test for different network sizes in terms of number of nodes

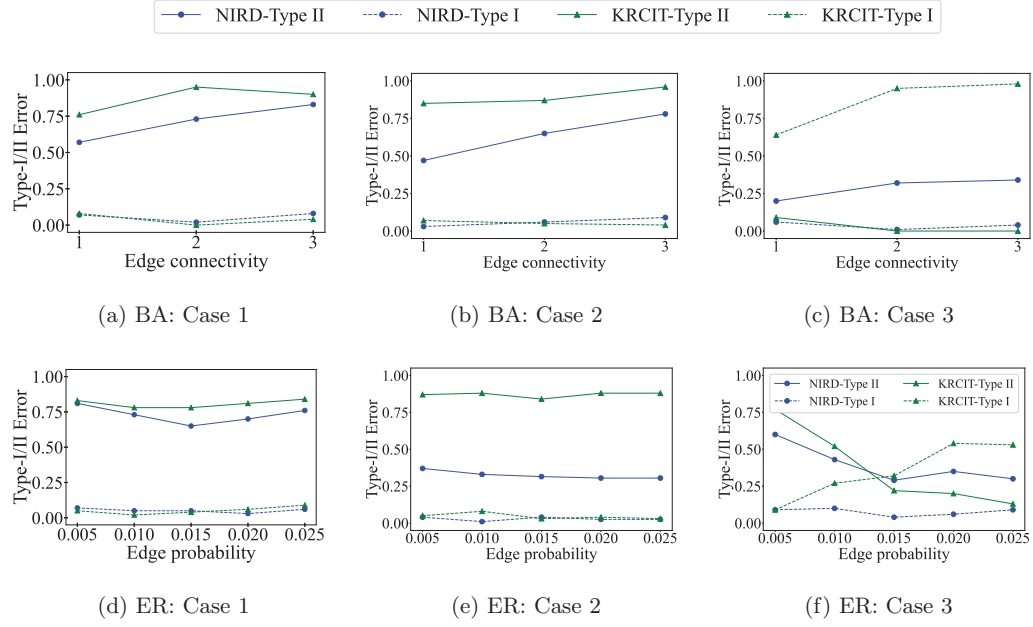
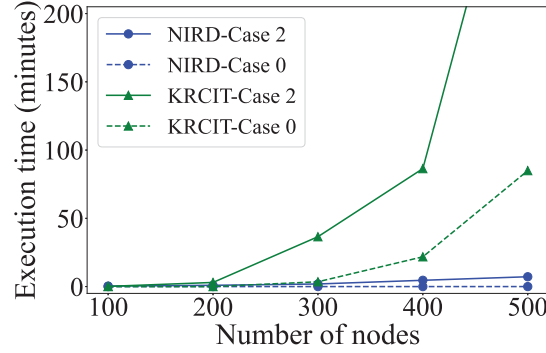


Figure 16: Impact of network parameters on Type I/II errors.

(x-axis). The solid and dashed lines represent the conditional and marginal test result respectively. KRCIT exhibits an exponential complexity whereas NIRD shows much less sensitivity to network size. This is expected given the complexity of the corresponding algorithms.

5.3.3.4 Diffusion

Figure 18a shows the impact of the number of diffusion steps (lines) and sample size (x-axis) on the effectiveness of NIRD. At initial activation (1 diffusion step) there is a high Type II error across sample sizes which decreases with higher number of steps. We see a significant decrease in error with just 5 diffusion steps. Further steps drastically lower the Type II error



(a) Impact of network size on execution time

Figure 17: Test scalability.

and at 20 steps and larger samples it can reject the null hypothesis consistently. This suggests that relational dependence is easier to detect after several diffusion steps rather than at early activation. It also demonstrates the effectiveness and scalability of NIRD in terms of detecting social phenomena in real world networks. Note that it is computationally infeasible to run the baseline method on such size of samples.

In order to understand the impact of activation probabilities on the results of the test, we run a different experiment with varying activation probabilities. We consider a similar semi-synthetic setup with Twitter ego-network which is a larger real world network consisting 11,176 nodes and 1,44,653 edges (Leskovec and McAuley, 2012). We consider a sample of 10,000 nodes and vary the initial activation probabilities. Figure 18b shows the Type-II errors (y-axis) for different diffusion step sizes (x-axis). The lines correspond to the initial activation probabilities

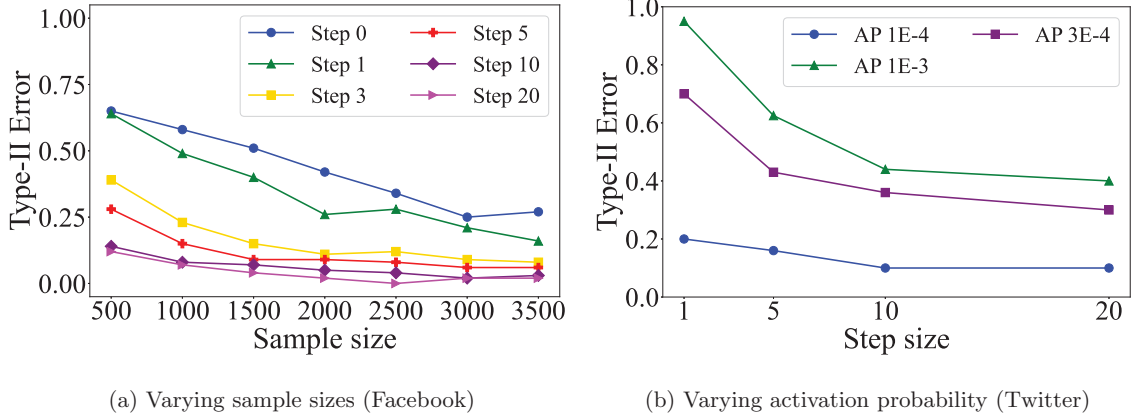


Figure 18: Type II error for the Linear Threshold Model on Facebook and Twitter ego-network.

(AP) for the diffusion process. We see the general trend of decreasing Type-II error with higher step sizes. It seems to be almost saturated with step 10. Moreover, the result indicates that the test is sensitive to activation probabilities and with higher activation probability, it shows higher type II error.

5.3.4 Comparison to Sobolev Independence Criterion (SIC)

To show the effectiveness of relational CI methods vs. CI methods developed for i.i.d. data, we compare both RCI methods (NIRD, KRCIT) to a recent i.i.d. CI test, the Sobolev Independence Criterion (Mroueh et al., 2019). SIC is an interpretable dependency measure between multivariate random variables characterized by integral probability metric between the joint distribution and the product of the marginals. We perform the SIC test on the flattened representation of the relational data, similar to KRCIT. Figure 19 extends the results

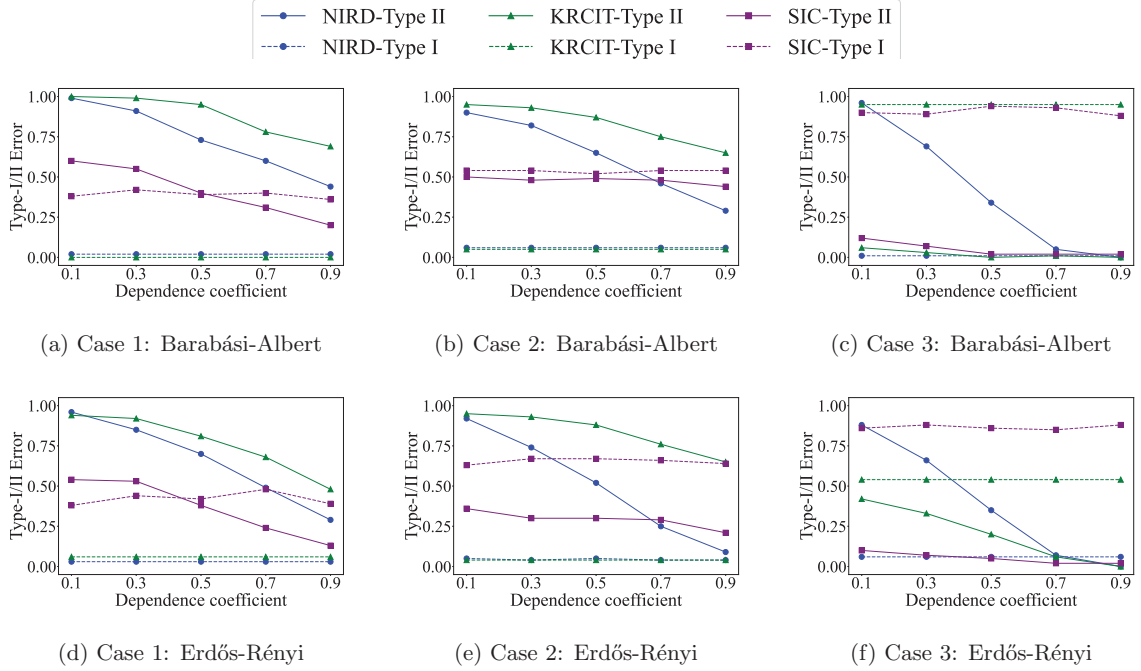


Figure 19: Type I/II errors with polynomial dependency model on synthetic networks for all three cases.

shown in Figure 1 in the main text. In all three cases, the i.i.d. baseline SIC exhibits high Type I error which shows its poor calibration to reasoning over the relational data.

5.4 Real-world Demonstration

One of the main challenges in social studies is to identify the effect of friends on their peers and the strength of such effects in different domains, e.g. health and violence. Studies show that patterns of interactions among adolescents can reveal possible reasons for changes in their

behavior over time. The central question in such studies is how to identify and measure the existence of such effects. The proposed independence test can facilitate reasoning over the existence of dependence between peers’ behaviors in social networks by providing a mechanism for falsifying statistical hypotheses.

As a demonstration, we examine the 50 Women dataset (Michell and Amos, 1997). This dataset has the smoking, sport, drug, alcohol consumption habits of 50 female students, along with their friendship information, over the course of three years. Each of the behavioral variables are coded as categorical variables indicating how regularly women engage in each of the behaviors. Assuming independence between the behavior peers as the null hypothesis, the goal of this analysis is to explore whether the habits of a student’s friends are associated with her habits in subsequent years.

Table I shows p-values estimated by our kernel test method considering four attributes in 50 Women dataset. We use column *Period* to indicate the years we consider for the test, e.g., in *Period* $1 \rightarrow 2, 3$, we explore students’ behavior change from first year to the second and third year. We consider both the original categorical coding and a binarization of the categorical attributes, which is 1 if the student uses a substance at least once during the year and 0 otherwise. The number of students who did not engage in the behavior during the first time point is shown in column t_0 , e.g., in the first row of the table, 4 students did not drink alcohol in the first year. We exclude t_0 for categorical data (indicated by NA) because the frequency of the habit is intrinsic to the hypothesis of interest in these cases. The last two columns (*NIDR_all* and *NIDR_t0*) show p-values measured by NIRD. In *NIDR_all* and

TABLE I: Real-world demonstration: exploration of the dependence between the habits of students and their first-hop neighbors in 50 Women dataset

period	attribute	attribute type	t0	NIRD_all	NIRD_t0
1 \rightarrow 2	alcohol	binary	4	0.425532	0.000000
1 \rightarrow 2	alcohol	categorical	NA	0.000000	NA
1 \rightarrow 2	drug	binary	35	0.000000	0.138298
1 \rightarrow 2	drug	categorical	NA	0.000000	NA
1 \rightarrow 2	smoke	binary	35	0.000000	0.021277
1 \rightarrow 2	smoke	categorical	NA	0.000000	NA
1 \rightarrow 2	sport	binary	12	0.925532	0.978723
1 \rightarrow 2	sport	categorical	NA	0.925532	NA
1 \rightarrow 2,3	alcohol	binary	5	0.114583	0.197917
1 \rightarrow 2,3	alcohol	categorical	NA	0.000000	NA
1 \rightarrow 2,3	drug	binary	35	0.000000	0.583333
1 \rightarrow 2,3	drug	categorical	NA	0.000000	NA
1 \rightarrow 2,3	smoke	binary	36	0.000000	0.000000
1 \rightarrow 2,3	smoke	categorical	NA	0.000000	NA
1 \rightarrow 2,3	sport	binary	12	1.000000	0.166667
1 \rightarrow 2,3	sport	categorical	NA	1.000000	NA
2 \rightarrow 3	alcohol	binary	3	0.125000	0.666667
2 \rightarrow 3	alcohol	categorical	NA	0.281250	NA
2 \rightarrow 3	drug	binary	32	0.000000	0.125000
2 \rightarrow 3	drug	categorical	NA	0.000000	NA
2 \rightarrow 3	smoke	binary	31	0.000000	0.281250
2 \rightarrow 3	smoke	categorical	NA	0.000000	NA
2 \rightarrow 3	sport	binary	20	0.864583	0.479167
2 \rightarrow 3	sport	categorical	NA	0.864583	NA

$NIDR_{t0}$ we consider all women (whether they have the habit in a year or not), and women who do not have the habit in the first time point, respectively. Overall we find:

- Sports activity of peers is not associated with whether a student plays a sport or not. High values of $NIDR_{t0}$ and $NIDR_{all}$ are enough evidences to accept the null hypothesis of independence.
- Peer smoking habits are associated with students' frequency of smoking: $NIDR_{all} = 0$ and $NIDR_{t0} < 0.022$ for all time periods, except period $2 \rightarrow 3$ where $NIDR_{t0} \approx 0.28$.
- Peer drug use is not associated with subsequent drug use in previously non-drug using students ($NIDR_{t0} > 0.05$). However, when we consider the effect of drug users on the overall population, it seems to be associated ($NIDR_{all} = 0$).
- Peer alcohol consumption is associated with the level subsequent alcohol use ($NIDR_{all} = 0$, except in period $2 \rightarrow 3$ where $NIDR_{all} > 0.1$), but not with the decision for a non-drinking student to begin drinking.

Different studies (Michell and Amos, 1997; Michael Pearson, 2000) deploy 50 women data to explore the association between gender, risk-taking or social position and smoking or drug usage in groups of youngsters. In particular our results comport with Pearson et al. (Michael Pearson, 2000) who show that drug usage and smoking are contagious among group of friends who are highly connected and people who are loosely connected to a friendship group.

5.5 Discussion

In this work we examine the problem of defining and measuring dependence in relational data. We proposed NIRD, a consistent, non-parametric test for detecting relational dependence that improves the state-of-the-art in relational dependence testing by capturing a wider range of possible relational dependencies than previous methods and improved computation time. We evaluated the effectiveness of our method across diverse relational settings and found that our proposed test exhibits significantly less sensitivity to network properties and dependence type. Our work paves the way for a promising future research direction on causal structure learning from relational data.

CHAPTER 6

EFFECTIVENESS OF SAMPLING STRATEGIES ON RELATIONAL DATA

Parts of this chapter were previously published as Ahsan, R., and Zheleva, E.: Effectiveness of Sampling Strategies for One-shot Active Learning from Relational Data. In the 16th International Workshop on Mining and Learning with Graphs (MLG 2020) Ahsan and Zheleva (2020)

Real-world networks have millions of nodes, and one of the biggest challenges in dealing with real-world relational data is its large size. A general and standard remedy is taking a smaller but representative sample from the relational data considering both the structure and distribution of the data. There are several sampling methods proposed over the years but there is a lack of understanding regarding their effectiveness. The closest works comparing their effectiveness rely on relational classification tasks for the comparison Berton et al. (2016); Ahmed et al. (2013). These studies are relatively old and don't consider modern deep-learning-based relational classification methods. In this study, I conduct a comprehensive empirical evaluation of existing sampling methods for relational data for the task of relational classification and propose a new sampling technique based on the latent structural properties of the data.

Standard practice for reducing the labeling complexity of classification methods is active learning which allows the classifier itself to select samples to be labeled by an oracle (Settles, 2009). State-of-the-art active learning strategies repeatedly select batch of samples in multiple iterations until a pre-specified budget of labels is reached. ALFNET (Bilgic et al., 2010),

RAL Kuwadekar and Neville (2011), (Kuwadekar and Neville, 2011) and ANRMAB (Gao et al., 2018) are some examples of such active learning approaches for relational data. However, these strategies generally learn a model at each iteration, in order to compute utility scores for all the unlabeled samples over all iterations which incurs a substantial computational cost. In order to address this issue for large networks, we consider a constrained problem setup where the active learner is allowed only a single iteration to select samples to label. We refer to it as *one-shot active learning*.

A popular approach to active learning is carefully selecting a representative sample of the source data. Prior works have empirically evaluated the effectiveness of sampling methods for one-shot active learning in the context of relational classification (Berton et al., 2016; Ahmed et al., 2013). However, the main difference between previous works with ours is twofold. First, they consider only the labeled subgraph for classification whereas we use the full source graph which gives us the opportunity to utilize the structural properties better. Second, they primarily consider a naive relational classifier, wVRN (Macskassy and Provost, 2007) whereas we compare the performance of collective classification and modern neural network-based approaches.

In this study we consider a wide variety of sampling strategies to compare their relative performance for one-shot active learning. We consider both graph sampling algorithms as well as sampling strategies specifically designed for semi-supervised node classification (Wu et al., 2019b). We also propose a sampling approach based on the Weisfeiler-Lehman algorithm (Weisfeiler and Lehman, 1968) which shows promising results in the empirical evaluation. Our proposed Weisfeiler-Lehman Sampling (WLS) relies solely on the structural role of nodes for

label acquisition decisions. One of its main advantages is that it is computationally efficient and yet harnesses structural information effectively. Our empirical evaluation shows that even though there isn't one sampling method that performs the best consistently across datasets and classifiers, Weisfeiler-Lehman ranks the highest on average.

The rest of the chapter is organized as follows: Section 6.1 presents the preliminary concepts used in this work. Section 6.2 defines the one-shot active learning problem for relational classification and provides an overview of existing solutions to consider along with the proposed WLS method. Section 6.3 describes the experimental evaluation and results. Section 6.4 presents a concluding discussion about the work.

6.1 Preliminaries

Here, we first describe the basic notations used for this work. Note that, even though we follow the definition of the relational model from section??, we reintroduce the notations here specifically for a relational classification task. Moreover, we present an overview of the classic Weisfeiler-Lehman (Weisfeiler and Lehman, 1968) algorithm that the proposed method is based on.

6.1.1 Basic Notations for Relational Classification

We consider an undirected graph $G = (\mathbf{V}, \mathbf{E})$ where \mathbf{V} and \mathbf{E} are the set of vertices and edges correspondingly. Each node V_i is associated with a feature vector \vec{X}_i and a corresponding class label Y_i which may be unknown, $V_i = \langle \vec{X}_i, Y_i \rangle$. A set of individual attributes comprises the vector $\vec{X}_i = \langle X_i^1, X_i^2, \dots, X_i^p \rangle$ where $1, 2, \dots, p$ are feature dimensions. The set of node features for all nodes is denoted by $\mathbf{X} = \{\vec{X}_i | V_i \in V\}$ and the set of class labels for all nodes is denoted

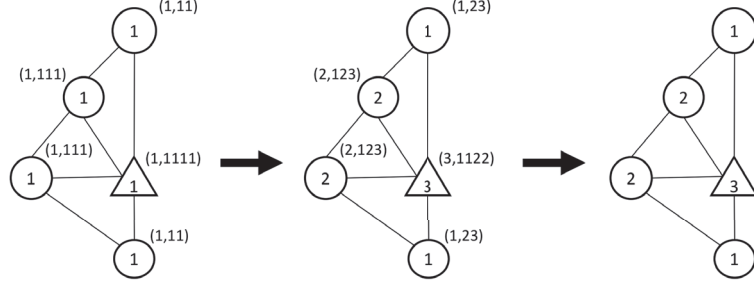


Figure 20: Illustration of classic Weisfeiler-Lehman algorithm: 1) same initial label to all nodes, 2) first relabeling after sorting signature strings, 3) final stable labels.

by $\mathbf{Y} = \{Y_i | V_i \in V\}$. The domain of class labels Y_i is discrete and the set of possible labels is denoted by $\mathcal{Y} = \{y_1, y_2, \dots, y_m\}$. The domain for \vec{X}_i can be either discrete or continuous. An edge $E_{ij} = \langle V_i, V_j \rangle$ represents an explicit link between two nodes V_i and V_j in the network. Let \mathcal{N}_i denote the set of neighboring nodes of V_i , $\mathcal{N}_i = \{V_j | \langle V_i, V_j \rangle \in E\}$.

6.1.2 Weisfeiler-Lehman Algorithm

The Weisfeiler-Lehman algorithm (Weisfeiler and Lehman, 1968) is a graph labeling algorithm that generates canonical ordering of the vertices of a given graph. The classic Weisfeiler-Lehman algorithm is presented in Algorithm 1. One key benefit it offers is the relative representation of the vertices based on their structural roles in the graph. Because of this feature, it has inspired several works in the network domain, especially for graph classification (Shervashidze and Borgwardt, 2009), graph embedding (Shervashidze et al., 2011) and link prediction (Zhang and Chen, 2017). The algorithm starts by assigning the same initial label

to all vertices (line 1). For each node, it forms a multiset of labels from its direct neighbors' color labels (line 4). After sorting the elements in the multiset and concatenating them to the node's label, it generates signature strings (lines 5-6). These signature strings are then sorted and compressed and used to assign new labels to the nodes (lines 8-9). This process continues until the labels have stabilized.

The effectiveness of the Weisfeiler-Lehman algorithm has been demonstrated for graph classification (Shervashidze et al., 2011) and link prediction (Zhang and Chen, 2017). Weisfeiler-Lehman is used to encode the subgraph properties for a given link for link prediction. This subgraph property is then used as input features to a neural network model to predict the existence of links, exploiting the ability of Weisfeiler-Lehman to encode relative structural roles of nodes in subgraph (Zhang and Chen, 2017). Our work is the first to study the application of Weisfeiler-Lehman encoding in the context of active learning sampling.

6.2 One-shot Active Learning for Relational Classification

In this section, we first formulate the problem and then describe the sampling methods we consider as potential solutions.

6.2.1 Problem Definition

One-shot active learning is a constrained version of the active learning problem. The main difference between them is that in the case of one-shot active learning the learner can query only once to acquire labels from the oracle.

Problem 2 (One-shot Active Learning for Relational Classification). *Given an undirected graph $G = (\mathbf{V}, \mathbf{E})$, node features \mathbf{X} , a labeling budget B , a relational classifier C , and a*

Algorithm 1 WEISFEILER-LEHMAN GRAPH LABELING

Input: Graph $G = (V, E)$, initial labels $l^0(v) = 1$ for all $v \in V$

Output: Final labels $l(v)$ for all $v \in V$

- 1: Let $l(v) = l^0(v)$ for all $v \in V$
 - 2: **while** $l(v)$ has not converged **do**
 - 3: **for each** $v \in V$ **do**
 - 4: Build a multiset $\{l(v')|v' \in \Gamma(v)\}$ concatenating
 - 5: its neighbor's labels
 - 6: Sort elements in the multiset in ascending order
 - 7: Concatenate the sorted multiset to $l(v)$ to generate
 - 8: a signature string $s(v) = \langle l(v), \{l(v')|v' \in \Gamma(v)\} \rangle$
 - 9: **end for**
 - 10: Sort all of the strings $s(v)$ for all v in ascending order
 - 11: Map each string $s(v)$ to a new compressed label,
 - 12: using a function f such that $f(s(v)) = f(s(w))$ if and
 - 13: only if $s(v) = s(w)$.
 - 14: **end while**
-

labeling oracle, select a set of nodes of size B to be labeled by the oracle in one shot such that Micro-F1 score of classifier C on unseen data is maximized upon training on the labeled set.

Note that, the active learning budget B is typically much less than the size of the available pool of unlabeled nodes. In such a scenario the classifier can either exploit the full graph structure or restrict itself to the subgraph induced by the labeled nodes. In this work, we focus on the first option with the availability of the full graph structure, making it a semi-supervised classification problem.

6.2.2 Sampling for One-shot Active Learning

Since the initial data has no labels available, smart sampling strategies are key to solving the one-shot active learning problem. The existing sampling methods for relational data can be categorized into three groups: 1) Network sampling methods, 2) Non-network sampling methods and 3) Hybrid sampling methods.

6.2.2.1 Network Sampling Methods

This category of sampling method is the state-of-the-art sampling for relational data. Their effectiveness in preserving the structural properties of networks makes them good candidates for one-shot active learning. These methods can be grouped into four major types:

- **Node Sampling:** This is a standard sampling strategy where the algorithm can sample nodes based on their structural properties (e.g., highest or lowest degree). One of the deficiencies of this sampling strategy is that it doesn't preserve the connectivity of the original graph. We consider two different node sampling methods. NS-DC-H refers to a sampling

method that prioritizes nodes with a high *degree of centrality* (Berton et al., 2016). We also consider sampling proposed by Berton et al. (Berton et al., 2016) which prioritizes nodes with high clustering coefficients (NS-CT-H).

- **Edge Sampling:** This is another standard sampling strategy where one can sample edges instead of nodes. Generally, all nodes incident to the edges is subsequently added to form the induced subgraph. The advantage of this method is that, unlike *Node Sampling*, it can preserve the connectivity of the original graph better. However, due to the independent selection of edges, it fails to preserve clustering properties. We consider random edge sampling (ES-RS) where edges are selected at random.
- **Topological Sampling:** Both *Node Sampling* and *Edge Sampling* methods exhibit shortcomings in preserving the structural properties of the graph. In order to overcome these shortcomings, several topology-based sampling algorithms have been proposed. These algorithms mostly utilize either breadth-first search or random walks over the graph to construct a representative sample (Ahmed et al., 2013). For example, *Snowball Sampling* (SS) selects nodes and edges following a breadth-first search from a randomly selected seed node. It stops when a certain threshold is reached. Another example is *Forest Fire Sampling* (FFS) which also follows a breadth-first search, but only considers a proportion (we consider 70%) of the neighborhood for exploration.
- **Graph Clustering:** Nguyen et al. (Nguyen and Smeulders, 2004) showed that clustering the data can help improve the performance of an active learning strategy. Inspired by this, we consider sampling based on graph clustering. We choose modularity-based clustering

(Newman, 2006) since it is a standard method for community detection in networks and has been used for active learning in the past (Bilgic et al., 2010). We generate modularity-based clusters and then iteratively select random nodes from each cluster until the labeling budget is exhausted. We refer to this sampling method as MS.

6.2.2.2 Non-network Sampling Methods

We consider sampling methods that are not based on networks for a comprehensive comparison. The primary reason for choosing this is to evaluate how much the structural information helps in prediction. The trivial choice for this category is random sampling (RS). Moreover, we consider sampling based on k-means clustering over the node features \mathbf{X} in order to utilize the strength of clustering in active learning (Bilgic et al., 2010). We create the k-means clusters for a given k value. Then in each iteration, we pick a random node from each cluster until the labeling budget is exhausted. We refer to this method as KMS.

6.2.2.3 Hybrid Sampling Methods

Several recent works use an intuitive idea of combining the power of both structural properties and node features. Most of those works follow a standard active learning strategy with multiple iterations. FEATPROP proposed by Wu et al. (Wu et al., 2019b) can be considered a hybrid approach for one-shot active learning. It clusters the samples using K-Means clustering based on a distance function derived by both propagated node features and graph structure. Then it iteratively selects the closest nodes to the cluster centers until the active learning budget is exhausted.

6.2.2.4 Weisfeiler-Lehman Sampling

We propose a new sampling algorithm, *Weisfeiler-Lehman Sampling* (WLS) for one-shot active learning for networks. It is adapted from the Weisfeiler-Lehman node labeling algorithm (Weisfeiler and Lehman, 1968). WLS considers the structural role of a node as the main predictor for label acquisition. Since Weisfeiler-Lehman method has been proven to be useful for encoding relative structural roles of nodes in link prediction problem (Zhang and Chen, 2017), we investigate its effectiveness for one-shot active learning.

The basic idea behind WLS is to explore different local neighborhoods of the graph and pick the nodes based on their relative structural roles. In order to achieve this, WLS utilizes the final color labels produced by Weisfeiler-Lehman algorithm. The color labels encode relative neighborhood properties of nodes which plays an important role in isomorphism testing. For any two isomorphic graphs, the first nodes exhibit similar structural properties in the corresponding orderings. The key benefit of this process is that the algorithm is able to pick the structural roles specific to each network, so they do not have to be defined a priori. Note that, in our case, a neighborhood is formed around a given seed node. An exploration budget B_e is introduced to limit the number of hops of neighborhood the algorithm can explore from the given seed node.

The WLS algorithm is presented in Algorithm 2. It starts with an empty set of labeled nodes (\mathcal{L}) and a set of all nodes \mathcal{U} (line 1). Then the algorithm keeps selecting k informative nodes until the labeling budget B is exhausted (lines 3-10). Note that, the number k here is the batch size for WLS. The true labels for these selected nodes are acquired from the oracle. Then the labeled set \mathcal{L} and unlabeled set \mathcal{U} are updated accordingly (lines 11-12). At the end

Algorithm 2 WEISFEILER-LEHMAN SAMPLING

Input: A network $G = (\mathbf{V}, \mathbf{E})$

Parameter: Batch size k , labeling budget B , exploration budget B_e

Output: Set of labeled nodes \mathcal{L}

```

1:  $\mathcal{L} = \emptyset, \mathcal{U} = \mathbf{V}$ 
2: while  $|\mathcal{L}| < B$  do
3:    $\mathcal{L}^k = \emptyset$ 
4:   Pick  $k$  random seed nodes,  $\mathcal{S}^k$  from  $\mathcal{U}$ 
5:   for each  $V_i \in \mathcal{S}^k$  do
6:      $\mathcal{N}_i \leftarrow$  Up to  $B_e$  hop neighborhood subgraph of  $V_i$ 
7:      $R_i \leftarrow \text{WEISFEILER-LEHMAN}(\mathcal{N}_i)$ 
8:      $v_i \leftarrow$  top ranked  $v \in (\mathcal{N}_i \cap \mathcal{U})$  in  $R_i$ 
9:      $\mathcal{L}^k = \mathcal{L}^k \cup v_i$ 
10:  end for
11:   $\mathcal{L} = \mathcal{L} \cup \mathcal{L}^k$ 
12:   $\mathcal{U} = \mathcal{U} \setminus \mathcal{L}^k$ 
13: end while
14: return  $\mathcal{L}$ 

```

of the iterations, the labeled set \mathcal{L} is ready to be used as training samples for classification. The core functionality of the algorithm lies in lines 3-10. Here, it first chooses k distinct random seed nodes from the unlabeled pool \mathcal{U} (line 4). Then, for each of the seed nodes, it constructs a subgraph with up to B_e hop neighbors of the seed node V_i (line 6). This subgraph is sent to the Weisfeiler-Lehman method (line 7) to produce the labels which we will consider as the canonical ordering of nodes based on structural properties. The algorithm always picks the first node in the produced ordering of the subgraph (line 8). It breaks ties arbitrarily. The selected node is added to the current set \mathcal{L}^k (line 9). The algorithm selects k nodes for label acquisition for corresponding k seed nodes. These nodes may or may not overlap with the seed nodes. In order to avoid duplication, once a node is picked it is no longer considered in the neighborhood of any other nodes. At the end of the iterations, the algorithm returns the set of k selected nodes (line 14). We use exploration budget $B_e = 2, 3$ for our evaluation and WLS-2, WLS-3 represent the corresponding versions of our algorithm.

The computational complexity of WLS directly depends on the complexity of WEISFEILER-LEHMAN algorithm. Let's denote the complexity of WEISFEILER-LEHMAN algorithm as $W(n)$ where n is the number of nodes to label. Also, let $N_i(h)$ refer to the average number of nodes in the h -hop neighborhood of any node V_i . The time complexity of WLS becomes $\mathcal{O}(B * W(N_i(h)))$.

Dataset	$ V $	$ E $	Number of Features	Number of Classes	Class Entropy	Average Degree	Clustering Coeff.	Homophily	Class Label
Citeseer	3312	4660	3703	6	1.71	2.81	0.1711	0.74	Topic
Cora	2708	5278	1433	7	1.83	3.90	0.2376	0.80	Topic
Hateful	3218	9620	1036	2	0.47	5.98	0.0785	0.72	Hatefulness
PubMed	19717	44327	500	3	1.06	4.50	0.0602	0.80	Topic

TABLE II: Properties of the datasets used in experimental evaluation.

6.3 Experimental Evaluation

6.3.1 Data

We conduct experiments on four real-world datasets, three of which are based on citation networks: Cora, Citeseer, and Pubmed ¹. The first two correspond to publications in computer science and the third one is based on publications on Diabetes diseases. The fourth dataset is sampled from the *Hateful Users on Twitter* dataset (Ribeiro et al., 2018). The original network contains around $100k$ users whereas around $5k$ users are annotated as either “hateful” or “normal”. Our sample consists of the annotated nodes and the edges between them.

We pre-process all datasets by removing all nodes that are not connected to the largest connected component. Table II summarizes the properties of the datasets after pre-processing. The column titled *Class Entropy* represents the entropy of the distribution of classes in a

¹All datasets available at <https://lincs.soe.ucsc.edu/data>.

dataset. The higher the entropy means more balanced class distribution and vice versa. Cora and Hateful seem to be the best and worst datasets in terms of class balance.

The next three columns of Table II shows several important network properties. We can see all the datasets exhibit a reasonably good amount of homophily where homophily is measured by the proportion of edges that connect two nodes from the same class. One interesting property to notice here is the distinction of clustering coefficient among the datasets. Based on these properties we can categorize the datasets into two groups. Citeseer and Cora fall into *Group I* with a smaller number of nodes and edges, relatively high clustering coefficient, and low average degree. They also consist of a higher number of classes with reasonable class balance. On the other hand, Hateful and Pubmed form *Group II* with a higher number of nodes and edges, high average degree, and low clustering coefficient. They exhibit strong class imbalance, especially Hateful.

6.3.2 Experimental Setup

6.3.2.1 The Hash Function for WLS

We use a specific hashing function, PALLETTE-WL (Zhang and Chen, 2017), compatible with the standard Weisfeiler-Lehman algorithm (1) for implementing WLS. PALLETTE-WL not only avoids higher computational cost by using a refined normalized hash function, but it also preserves vertex orders across iterations. This technique has been shown to be effective in link prediction for utilizing subgraph features (Zhang and Chen, 2017). In our work, we use the PALLETTE-WL method for the implementation of the Weisfeiler-Lehman algorithm. Note that, the complexity for PALLETTE-WL is $\mathcal{O}(n^2)$ where n is the number of nodes to label (Zhang

and Chen, 2017). So, according to the description in Section 6.2.2.4, the overall complexity of WLS becomes $\mathcal{O}(k * W(N_i(h)^2))$.

6.3.2.2 Relational Classifiers

We consider *Logistic Regression* as the local classifier for ICA and *Count* function as the aggregator. We choose *Simplified GCN* (SGC) (Wu et al., 2019c) and GRAPH-SAGE (Hamilton et al., 2017) classifiers as representatives of GNN. SGC is a faster approximation of the popular relational classifier GCN.

6.3.2.3 Evaluation Methodology

We randomly split 80% of the nodes for training and keep the other 20% for testing. We run all our experiments 5 times and take the average.

Most of the datasets in this study contain multiple classes and there is a considerable class imbalance present in the data as shown in Table II. To reduce the impact of class imbalance in evaluation, we used stratification while splitting the train and test samples. We also considered class-weighted loss functions for the classifiers. We considered Micro-F1 score as the evaluation metric. This is a popular metric used to evaluate multi-class classification.

We varied the active learning budget B from 32 up to 224. We used the same budget for all datasets for consistency. The maximum budget, 224 represents 10% of the training nodes for all datasets except PubMed. We consider batch size $k = 8$ for some sampling methods in our experiments. It represents the number of nodes selected per iteration for WLS whereas for MS and KMS it represents the number of clusters.

6.3.2.4 Packages and Hardware

We use NetworkX 2.3 (Hagberg et al., 2008) for representing and processing graphs. Scikit-Learn library is used for implementation of *Logistic Regression* and K-means clustering. We use StellarGraph (Data61, 2018) package for implementing SGC and GRAPH-SAGE¹. For running all our experiments and recording execution time we use Ubuntu 18.04 OS running on a 96-core Intel(R) Xeon(R) Platinum 8275CL @ 3.00GHz processor with 185GB memory.

6.3.3 Results

Figure 21 shows results for all the candidate sampling methods using four classifiers (wvRN, ICA, SGC, GRAPH-SAGE) on four different datasets. In the figure, the rows represent different datasets and the columns represent different classifiers used. Moreover, the y-axis represents Micro-F1 score and the x-axis shows the number of training nodes considered as active learning budget B . We can observe a great deal of variance in terms of performance of different sampling methods. To better understand the relative performance, we list down all the Micro-F1 scores for the highest budget (224) in table:f1. Each row in this table corresponds to a specific dataset and a specific relational classifier. The bold cell represents the best Micro-F1 score in the corresponding row. For example, in the first row, for Citeseer dataset and wvRN classifier, NS-DC-H performs the best. We rank the algorithms based on this table and present the final ranking in Table IV. We used the following process to generate the final ranking: first, rank all sampling algorithms for each row of Table III separately and then calculate the average rank of

¹The code is available online at <https://github.com/edgeslab/sampling-osal>

TABLE III: Micro-F1 scores of 11 sampling methods across 4 datasets and 4 classifiers for an active learning budget of 224 nodes.

Dataset	Classifier	RS	NS-DC-H	NS-CT-H	ES-RS	FeatProp	WLS-2	WLS-3	SS	FFS	KMS	MS
Citeseer	wvRN	0.378199	0.424171	0.332701	0.419431	0.388132	0.405213	0.400474	0.298803	0.278156	0.354976	0.361611
Citeseer	ICA	0.712322	0.581991	0.716588	0.702844	0.648728	0.733745	0.737602	0.361347	0.425220	0.718483	0.727962
Citeseer	SGC	0.758588	0.700552	0.742451	0.715568	0.670169	0.737108	0.748460	0.457079	0.441471	0.758101	0.749000
Citeseer	GSAGE	0.682464	0.656398	0.688152	0.661611	0.663981	0.676303	0.697630	0.356634	0.403149	0.677725	0.669668
Cora	wvRN	0.418511	0.627767	0.361617	0.499396	0.558551	0.495372	0.449497	0.308099	0.276676	0.449253	0.443461
Cora	ICA	0.765392	0.721932	0.753320	0.741247	0.786318	0.742455	0.760161	0.536877	0.493188	0.754930	0.771429
Cora	SGC	0.762181	0.812475	0.752918	0.756856	0.800000	0.775742	0.785630	0.412851	0.397802	0.756111	0.765896
Cora	GSAGE	0.783501	0.803219	0.764185	0.776660	0.793159	0.773843	0.785111	0.562559	0.467780	0.774245	0.788330
Hateful	wvRN	0.813704	0.333333	0.817037	0.848519	0.837407	0.812593	0.810741	0.695926	0.456296	0.813704	0.814444
Hateful	ICA	0.888519	0.896296	0.888519	0.905556	0.900741	0.890000	0.889259	0.872593	0.888889	0.884444	0.879259
Hateful	SGC	0.899083	0.545549	0.899083	0.545549	0.899083	0.899083	0.899083	0.781238	0.427704	0.899083	0.899083
Hateful	GSAGE	0.867598	0.340110	0.884077	0.856113	0.888896	0.899083	0.857594	0.720681	0.514434	0.899083	0.899083
Pubmed	wvRN	0.409381	0.479817	0.401217	0.435446	0.409381	0.427840	0.430680	0.424074	0.420188	0.409432	0.408773
Pubmed	ICA	0.715112	0.503296	0.734888	0.695538	0.550963	0.690974	0.709533	0.537221	0.472110	0.730781	0.727890
Pubmed	SGC	0.523560	0.669704	0.525552	0.592995	0.496082	0.626141	0.610570	0.608615	0.549057	0.518602	0.594904
Pubmed	GSAGE	0.775913	0.767546	0.776318	0.773986	0.744371	0.775355	0.779513	0.672577	0.605269	0.756542	0.769320

TABLE IV: Average ranks of sampling methods for different categories of relational classifiers over all datasets.

Sampling	Avg. Rank		
	All	GNN	wvRN, ICA
WLS-3	4.19 ± 1.81	3.38 ± 1.80	5.00 ± 1.41
WLS-2	4.56 ± 1.87	4.38 ± 1.87	4.75 ± 1.85
MS	4.88 ± 2.60	4.25 ± 1.39	5.50 ± 3.28
FeatProp	5.28 ± 3.20	6.00 ± 3.24	4.56 ± 2.99
RS	5.31 ± 2.21	4.62 ± 2.29	6.00 ± 1.89
ES-RS	5.44 ± 2.71	7.12 ± 1.45	3.75 ± 2.63
KMS	5.66 ± 2.54	5.62 ± 2.83	5.69 ± 2.22
NS-DC-H	5.75 ± 4.07	5.88 ± 3.92	5.62 ± 4.21
NS-CT-H	5.88 ± 2.98	5.50 ± 2.69	6.25 ± 3.19
SS	9.19 ± 1.94	9.00 ± 2.06	9.38 ± 1.80
FFS	9.88 ± 1.76	10.25 ± 1.30	9.50 ± 2.06

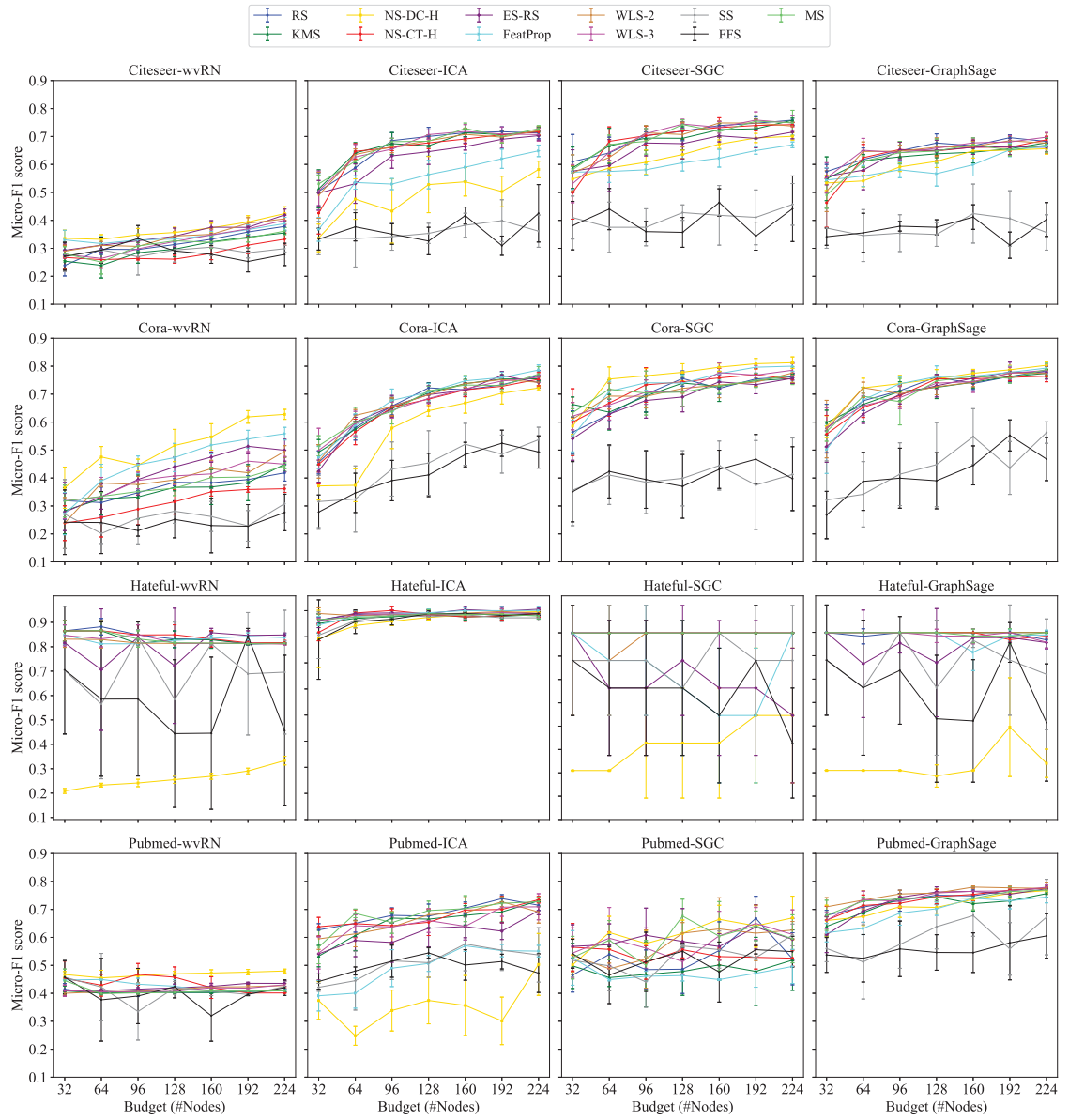


Figure 21: Macro-F1 scores by different sampling methods using different relational classifiers.

each sampling algorithm over all 16 rows. The first column of Table IV represents the sampling methods sorted by their average ranks over all 16 combinations. The next three columns show the average rank along with the standard deviation of the corresponding sampling methods based on different categories of relational classifiers. The column *All* shows the average ranks for all classifiers whereas the column *GNN* shows average ranks for only GNN-based classifiers (over 8 rows). The last column shows average ranks for wvRN and ICA classifiers (over 8 rows). Both versions of our proposed method (WLS-3, WLS-2) top the overall ranking and show relatively low standard deviation. This establishes its robustness across multiple datasets and classifiers. Next, we present the main takeaways by analyzing the performance of the sampling methods from several different perspectives:

Group I vs Group II datasets We can observe from the results that our proposed method WLS performs relatively better in *Group II* where the network is larger and exhibits a higher average degree. On the other hand, random sampling (RS) and Degree Centrality (NS-DC-H) work better in *Group I*. This indicates that in smaller networks with high clustering coefficients, simple node sampling or even random sampling is good enough for one-shot active learning. On the other hand, larger graphs with low clustering coefficients require more sophisticated methods like WLS.

Network Sampling vs others. Next, we observe how different categories of sampling methods perform across all setups. Figure 21 shows that graph sampling methods like *Snowball Sampling* (SS) or *Forest Fire Sampling* (FFS) exhibit poor performance for relational classification. This is intuitively provided that these methods are heavily biased in spatial exploration.

They fail to explore diverse local regions of the graph. Another expected observation is that the non-network sampling approach KMS suffers in almost all cases since it can not exploit any of the relational information. Note that even though it shows a high Micro-F1 score for the Hateful dataset, that could be due to the high class imbalance in that dataset. Node sampling approaches seem to do best in Cora and Pubmed where higher homophily is observed. In general graph-based clustering method (MS) shows a consistently good performance across all setups. Surprisingly, the hybrid approach FEATPROP only performs great in Cora but produces relatively poor results for other cases.

GNN vs others. Most of the sampling methods produce higher Micro-F1 score when used with GNN approaches compared to wvRN and ICA. GNN approaches show consistently better performance across all datasets except for Hateful. Surprisingly, FEATPROP works best with ICA even though it was primarily designed for SGC. It is interesting to note that certain sampling methods show significant variation in performance based on the classifier. For example, degree centrality (NS-DC-H) shows good result using SGC (3rd row) but quite poor using ICA (2nd row) on Citeseer dataset. In contrast, WLS-2 and WLS-3 show less variance across different classifiers and datasets. The last two columns of Table IV show the difference in performance for GNN-based classifiers versus previous relational classifiers. The cells marked in bold represent the top three average ranks in each category and the top three overall sampling methods also perform the best for GNN-based classifiers. However, ES-RS takes the top spot in the last column which supports the findings by Ahmed et al. (Ahmed et al., 2012)

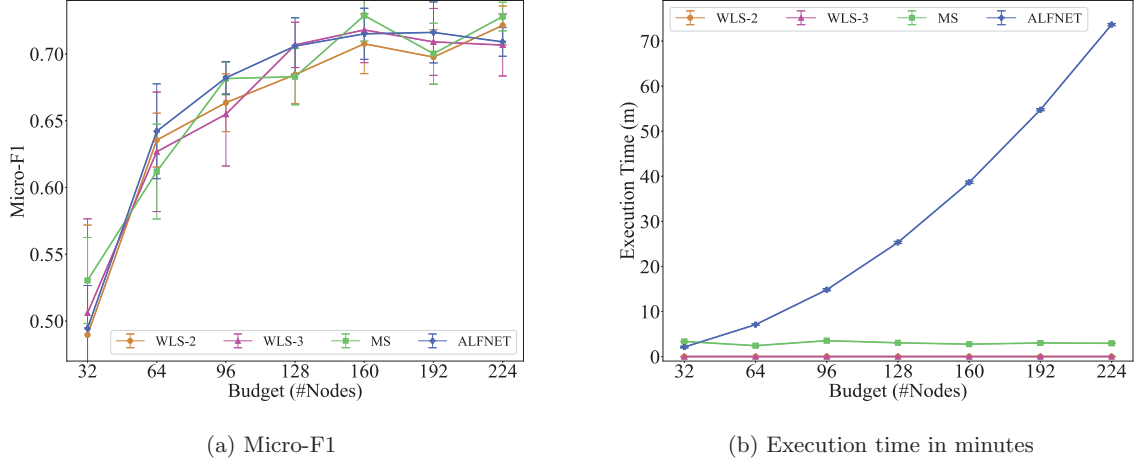


Figure 22: Top three sampling methods vs ALFNET using ICA classifier on Citeseer dataset.

One-shot vs Multi-shot. In order to show the effectiveness of one-shot active learning, we compare the sampling methods with ALFNET, a state-of-the-art active learning algorithm for relational data which requires iterative training over the acquired samples. Figure 22 shows both Micro-F1 score (22a) and execution times (22b) for the sampling methods and ALFNET on the Citeseer dataset. We choose only the top 3 (second row of Table III) sampling methods (WLS-3, WLS-2, MS) for convenience of comparison. In 22a, we can see the sampling methods show competitive results compared to ALFNET. However, in 22b the difference in execution time is significant. Note that, the execution times presented here is in minutes, and lines for WLS-2, WLS-3 overlap with each other. This big difference in execution time and competitive Micro-F1 score justifies the motivation behind one-shot active learning.

6.4 Discussion

We address a constrained classification problem, *one-shot active learning* for relational data. The objective is to reduce both the labeling and computation costs of relational classification in large real-world network datasets. We explore a wide variety of sampling methods as solutions and proposed a node sampling method based on Weisfeiler-Lehman algorithm. We experimentally evaluate all these sampling methods on four real-world network datasets and four popular relational classifiers. The main takeaways are as follows:

- WLS performs best with GNN-based classifiers whereas ES-RS shows the best results for wvRN and ICA classifier. WLS also shows overall best performance across all setups.
- Network-based node sampling methods work well for smaller networks with high clustering coefficients.
- One-shot active learning methods produce competitive results compared to state-of-the-art multi-shot active learning methods with much smaller computational costs.

CHAPTER 7

CONCLUSION

Learning relational causal models from observational data is of central importance for understanding real-world complex relational systems. It can help identify and understand the mutual influence and interference in social networks, organizational networks, medical diagnostics, and many other real-world phenomena. However, few relational causal discovery methods have been developed in recent years and they make strong assumptions (i.e. acyclicity) which make them unsuitable to reason about real-world dynamic systems with feedback loops and cycles. Moreover, the tools and techniques (i.e., conditional independence test) used in the existing algorithms lack generalizability and scalability. This thesis aims to address these deficiencies in the current literature.

In this thesis, I first develop an abstract representation for *cyclic relational causal models* that can capture the conditional independence statements consistent across all possible instantiations of the model. I introduce *relational σ -separation* criteria which can answer relational queries on cyclic relational causal models. I provide theoretical guarantees for representation and reasoning with the proposed methods under some suitable assumptions. Based on the newly developed representation and understanding of cyclic relational models, I focus on learning cyclic relational causal models from observational samples. I develop and characterize *relational acyclification* that helps to reason about the identifiability of cycles in relational systems. I establish necessary conditions and assumptions for learning these models using existing

relational causal discovery algorithms with theoretical guarantees. I provide adequate experimental evaluations in support of my claims. In order to facilitate the practical application of relational causal discovery methods for real-world data, I address the issues of generalizability and scalability of existing conditional independence tests for relational data. I develop a general definition of relational dependence for relational variables and I propose a nonparametric test for measuring relational dependence based on the given definition. I provide asymptotic guarantees for the convergence of the test. The proposed relational dependence test can generalize over a family of aggregate functions and can scale much better than the state-of-the-art. I show the effectiveness of the test in real-world applications. To benefit from the advancement of machine learning and relational classification for relational causal discovery, I study the effectiveness of varieties of sampling strategies for relational classification. I conduct experimental evaluations on several real-world datasets and state-of-the-art relational classifiers to compare the performance of different sampling strategies. Moreover, I propose a computationally efficient and robust sampling strategy for relational data which produces competitive performance in my experimental study.

7.1 Limitations and Future Directions

During the development of this thesis, I came across several limitations and challenges of both prior works and my own proposals which present a great opportunity for the future direction of research in this field of study. In the following subsection, I point out some of the limitations of the works presented in this thesis and possible future research directions.

1. **Latent confounders:** Presence of latent confounders is a big challenge for causal inference and discovery methods in general. All of the existing studies in relational causal discovery, including this thesis, assume that there are no latent confounders present in the relational system. However, this is counter-intuitive in many real-world applications and it is an important limitation of the current literature. Understanding and identifying the presence of latent confounders in relational models can help discovery methods be useful in practical scenarios. There are few algorithms proposed for causal discovery in the presence of latent confounders (i.e. FCI (Spirtes et al., 2000), BackShift (Rothenhäusler et al., 2015)). However, these algorithms were developed for propositional, i.i.d data. The abstract representation proposed in this thesis poses an opportunity to adopt the existing propositional causal discovery methods for relational causal discovery. A promising future direction is to understand the implication of latent confounders on relational models and how they are different than their propositional counterparts.
2. **Cardinality constraint:** The completeness of relational σ -separation (and relational d -separation) depends on the assumption that any node in the relational skeleton has degree of more than one. Even though it can still be reasonable for different kinds of application areas, it prohibits application to general real-world cases. A possible future direction could be relaxing this assumption to allow the broader area of application while ensuring some theoretical guarantees. It has been shown that a weaker sense of completeness can be used to learn an acyclic relational causal model from data (Lee and Honavar, 2015). A similar approach can be taken for cyclic relational causal models.

3. **Relational acyclification:** In this thesis we introduced relational acyclification criteria and established the soundness and completeness of RCD for cyclic relational causal models based on a strong constraint on the hop threshold of the relational acyclifications. The constraint states that the RCD is sound and complete only for the cyclic models for which the hop thresholds for its acyclifications do not exceed the hop threshold of the model itself. It essentially limits the scope of the algorithm to a smaller set of models. An intuition to alleviate this problem is to allow a higher hop threshold during the skeleton-building phase of the RCD algorithm. However, this requires further study of the equivalence of different cyclic causal models with a different cap on hop thresholds which can be an important direction for future research.
4. **Convergence rate of relational dependence test:** The relational dependence test proposed in this thesis is shown to converge in an asymptotic sense. However, it doesn't help estimate sample complexity for convergence in real-world cases. The rate of convergence will depend on the weak dependence coefficient. In the case that the coefficient is 0, this reduces to results that correspond to prior work on i.i.d data (e.g., Zhang et al. (2011)). While there is prior work studying this in more restrictive assumptions on the dependence between instances (e.g. London et al. (2013)), I am not aware of similar results for the case of weak dependence in general structured domains even in the simpler case of regression. Thus, any development in the characterization of the convergence rate for the proposed test would help practitioners utilize the test for practical scenarios.

5. **Learning from intervention:** Most of the work in relational causal discovery, including this thesis, primarily focuses on causal discovery entirely from observational data. While it is reasonable in many cases, it misses the opportunity to exploit interventional data to guide the learning process further. There are several causal discovery algorithms proposed for propositional data which can utilize intervention for causal learning (Rothenhäusler et al., 2015; Hauser and Bühlmann, 2012; Hyttinen et al., 2012). In recent work, Besserve and Schölkopf (2022) took advantage of the automatic differentiation techniques to optimize for a special kind of intervention, called *soft interventions* in cyclic causal models. Earlier, Schmidt and Murphy (2009) developed a representation that can be used to model the effects of interventions through a directed cyclic graph. However, no such approach was developed for relational models. There are several challenges to reasoning with interventions in relational causal models. To start with, we first need a better understanding of interventions in relational variables. Intervention can happen on both the structure (i.e cutting friendship ties) and on values (i.e forcing attribute values). Moreover, intervention on a relational variable (i.e friends’ preference) may be tied with intervention on a propositional variable (i.e individual preference). To the best of my knowledge, the characterization of intervention on relational variables is not studied in the existing literature. This can be a great direction for future research.
6. **Score-based and hybrid approaches:** A general observation is that all the studies in relational causal discovery primarily focus on constraint-based approaches. However, there are several score-based methods and hybrid methods (constraint-based + score-

based) available for propositional data (Chickering, 2002; Hauser and Bühlmann, 2012; Nandy et al., 2018; Tsamardinos et al., 2006). Score-based methods can open a new avenue for relational causal discovery methods in the future.

7. **Practical considerations:** The contributions in my thesis lay a theoretical foundation for reasoning with relational causal models in the presence of cycles or feedback loops and learning such models from data. I hope future developments would be able to benefit from my work, especially with its application to practitioners. For this reason, I'd like to briefly discuss the key challenges of adopting my work for real-world applications. Even though the proposed relational dependence test, NIRD, is relatively more scalable than the baseline method, it still needs improvements to run on real-world data. Either theoretical or empirical estimates of the convergence rate of the test would also help identify the scope of the test. Finally, latent confounders are unavoidable in real-world scenarios. Adopting FCI-based approaches for relational causal discovery would be beneficial for the practical applications of this work.

CITED LITERATURE

Bibliography

- Ragib Ahsan, David Arbour, and Elena Zheleva. Relational causal models with cycles: Relational causal models with cycles: Representation and reasoning. In *1st Conference on Causal Learning and Reasoning (CLear 2022)*, 2022a.
- Ragib Ahsan, Zahra Fatemi, David Arbour, and Elena Zheleva. Non-parametric inference of relational dependence. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, 2022b.
- Ragib Ahsan and Elena Zheleva. Effectiveness of sampling strategies for one-shot active learning from relational data. 2020.
- Sanghack Lee and Vasant G Honavar. Lifted representation of relational causal models revisited: Implications for reasoning and structure learning. In *ACI@UAI*, 2015.
- Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenophon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010a.
- Marc Maier, Katerina Marazopoulou, David Arbour, and David Jensen. A sound and complete algorithm for learning causal models from relational data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI’13, pages 371–380, Arlington, Virginia, United States, 2013a. AUAI Press. URL <http://dl.acm.org/citation.cfm?id=3023638.3023676>.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Robert R Tucci. Introduction to judea pearl’s do-calculus. *arXiv preprint arXiv:1305.5506*, 2013.
- Lise Getoor, Nir Friedman, Daphne Koller, Avi Pfeffer, and Benjamin Taskar. Probabilistic relational models. *Introduction to statistical relational learning*, 8, 2007.
- Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.
- Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.
- Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.
- Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4): 1659–1676, 2020.
- Marc Maier, Katerina Marazopoulou, and David Jensen. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*, 2013b.

- Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020.
- T Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, 1996.
- Joris M Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pages 1159–1168. PMLR, 2020.
- Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19, 2000.
- Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.
- Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 28, 2015.
- Eric V Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8(1):33–56, 2019a.
- Sanghack Lee and Vasant Honavar. A characterization of markov equivalence classes of relational causal models under path semantics. In *UAI*, 2016a.
- Sanghack Lee and Vasant Honavar. On learning causal models from relational data. In *AAAI*, pages 3263–3270, 2016b.
- Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.
- Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI’11, pages 804–813, Arlington, Virginia, United States, 2011. AUAI Press. ISBN 978-0-9749039-7-2. URL <http://dl.acm.org/citation.cfm?id=3020548.3020641>.
- Seth R. Flaxman, Daniel B. Neill, and Alexander J. Smola. Gaussian processes for independence tests with non-iid data in causal inference. *ACM Trans. Intell. Syst. Technol.*, 7(2):22:1–22:23, November 2015. ISSN 2157-6904. doi: 10.1145/2806892. URL <http://doi.acm.org/10.1145/2806892>.

- Sanghack Lee and Vasant Honavar. A kernel conditional independence test for relational data. 1 2017. 33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017 ; Conference date: 11-08-2017 Through 15-08-2017.
- M. Maier, K. Marazopoulou, D. Arbour, and D. Jensen. Flattening network data for causal discovery: What could go wrong? In *WIN*, 2013c.
- Lilian Berton, Didier Augusto Vega-Oliveros, Jorge Carlos Valverde-Rebaza, Andre Tavares da Silva, and Alneu de Andrade Lopes. The impact of network sampling on relational classification. In *SIMBig*, pages 62–72, 2016.
- Nesreen K Ahmed, Jennifer Neville, and Ramana Kompella. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2): 1–56, 2013.
- Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5), oct 2021. doi: 10.1214/21-aos2064. URL <https://doi.org/10.1214%2F21-aos2064>.
- Patrick Forré and Joris M Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv preprint arXiv:1710.08775*, 2017.
- Sanghack Lee and Vasant Honavar. Towards robust relational causal discovery. In *Uncertainty in Artificial Intelligence*, pages 345–355. PMLR, 2020.
- David Heckerman, Chris Meek, and Daphne Koller. Probabilistic entity-relationship models, prms, and plate models. *Introduction to statistical relational learning*, pages 201–238, 2007.
- Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.
- David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- Preetam Nandy, Alain Hauser, Marloes H Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.
- Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.
- Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.
- Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

- Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.
- Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.
- Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. *arXiv preprint arXiv:1301.6695*, 2013.
- Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.
- Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, pages 49–56, 2004.
- Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, pages 589–615, 2006.
- Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380, 2003a.
- Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003b.
- Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. *Journal of Machine Learning Research*, 11(1), 2010b.
- M. Maier, B. Taylor, H. Oktay, and D. Jensen. Learning causal models of relational domains. In *AAAI*, 2010.
- David Arbour, Katerina Marazopoulou, and David Jensen. Inferring causal direction from relational data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 12–21, 2016a.
- Xinhua Zhang, Le Song, Arthur Gretton, and Alex J Smola. Kernel measures of independence for non-iid data. In *Advances in neural information processing systems*, pages 1937–1944, 2009.
- Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616, 2014.
- Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- Mustafa Bilgic, Lilyana Mihalkova, and Lise Getoor. Active learning for networked data. In *ICML*, 2010.

- H. Sebastian Seung, Manfred Oppor, and Haim Sompolsky. Query by committee. In *COLT*, 1992.
- Li Gao, Hong Yang, Chuan Zhou, Jia Wu, Shirui Pan, and Yue Hu. Active discriminative network representation learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.
- Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. Active learning for graph embedding. *arXiv preprint arXiv:1705.05085*, 2017.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.
- Mark Woodward and Chelsea Finn. Active one-shot learning. *arXiv preprint arXiv:1702.06559*, 2017.
- Sofus A Macskassy and Foster Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of machine learning research*, 8(May):935–983, 2007.
- Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29:93–106, 2008.
- Jennifer Neville and David Jensen. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pages 13–20, 2000.
- Qing Lu and Lise Getoor. Link-based classification. In *ICML*, 2003.
- Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019a.
- Zhiqian Chen, Fanglan Chen, Lei Zhang, Taoran Ji, Kaiqun Fu, Liang Zhao, Feng Chen, and Chang-Tien Lu. Bridging the gap between spatial and spectral domains: A survey on graph neural networks. *arXiv preprint arXiv:2002.11867*, 2020.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.
- Eric D Kolaczyk. Sampling and estimation in network graphs. In *Statistical Analysis of Network Data*, pages 1–30. Springer, 2009.
- Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical review E*, 73(1):016102, 2006.
- Sooyeon Yoon, Sungmin Lee, Soon-Hyung Yook, and Yup Kim. Statistical properties of sampled networks by random walks. *Physical Review E*, 75(4):046114, 2007.

- Michael PH Stumpf, Carsten Wiuf, and Robert M May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, 102(12):4221–4224, 2005.
- Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006.
- Nesreen K Ahmed, Jennifer Neville, and Ramana Kompella. Reconsidering the foundations of network sampling. In *Proceedings of the 2nd Workshop on Information in Networks*, 2010.
- Arun S Maiya and Tanya Y Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–113, 2011.
- Lars Backstrom and Jon Kleinberg. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*, pages 615–624, 2011.
- Nesreen K Ahmed, Jennifer Neville, and Ramana Kompella. Network sampling designs for relational classification. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- Sofus A. Macskassy. Using graph-based metrics with empirical risk minimization to speed up active learning on networked data. In *KDD*, 2009.
- Nesreen Ahmed, Jennifer Neville, and Ramana Rao Kompella. Network sampling via edge-based node selection with graph induction. Computer Sciences Technical Report 1648, Purdue University, 2011.
- Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, 2001.
- Thomas Richardson. A characterization of markov equivalence for directed cyclic graphs. *International Journal of Approximate Reasoning*, 17(2-3):107–162, 1997.
- Eric V. Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, pages 1–24, 2019b.
- Kari Rantanen, Antti Hyttinen, and Matti Järvisalo. Discovering causal graphs with cycles and latent confounders: An exact branch-and-bound approach. *Int. J. Approx. Reason.*, 117: 29–49, 2020.
- Peter L Spirtes. Directed cyclic graphical representations of feedback models. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.
- Radford M Neal. On deducing conditional independence from d-separation in causal graphs with feedback (research note). *Journal of Artificial Intelligence Research*, 12:87–91, 2000.
- S. Kramer, N. Lavrač, and P. Flach. Propositionalization approaches to relational data mining. In *Relational data mining*, pages 262–291. Springer, 2001.

- JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.
- D. Arbour, D. Garant, and D. Jensen. Inferring network effects in observational data. In *KDD*, 2016b.
- Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.
- Donald WK Andrews and David Pollard. An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review/Revue Internationale de Statistique*, pages 119–132, 1994.
- Peter J Bickel and Peter Bühlmann. A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli*, pages 413–446, 1999.
- Jerome Dedecker, Paul Doukhan, and Gabriel Lang. *Weak Dependence: With Examples and Applications*. Springer, 2007.
- David Arbour. Method for enabling causal inference in relational domains. 2017.
- Rongjing Xiang and Jennifer Neville. Relational learning with one network: An asymptotic analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 779–788. JMLR Workshop and Conference Proceedings, 2011.
- Anne Leucht and Michael H Neumann. Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.
- Ben London, Bert Huang, Ben Taskar, and Lise Getoor. Collective stability in structured prediction: Generalization from one example. In *International Conference on Machine Learning*, pages 828–836. PMLR, 2013.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.
- Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.
- Jure Leskovec and Julian J. McAuley. Learning to discover social circles in ego networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 539–547. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4532-learning-to-discover-social-circles-in-ego-networks.pdf>.
- Lynn Michell and Amanda Amos. Girls, pecking order and smoking. *Social Science & Medicine*, 44(12):1861 – 1869, 1997. ISSN 0277-9536. doi: [https://doi.org/10.1016/S0277-9536\(96\)00295-X](https://doi.org/10.1016/S0277-9536(96)00295-X). URL <http://www.sciencedirect.com/science/article/pii/S027795369600295X>.
- Y. Mroueh, T. Sercu, M. Rigotti, I. Padhi, and C. Nogueira dos Santos. Sobolev independence criterion. *NeurIPS*, 2019.

- Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.
- Lynn Michell Michael Pearson. Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs: Education, Prevention and Policy*, 7(1):21–37, 2000. doi: 10.1080/dep.7.1.21.37. URL <https://doi.org/10.1080/dep.7.1.21.37>.
- Ankit Kuwadekar and Jennifer Neville. Relational active learning for joint collective classification models. In *ICML*, 2011.
- Yuexin Wu, Yichong Xu, Aarti Singh, Yiming Yang, and Artur Dubrawski. Active learning for graph neural networks via node feature propagation. In *Workshop on Graph Representation Learning, NeurIPS*, 2019b.
- Boris Weisfeiler and AA Lehman. A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsia*, 2(9):12–16, 1968.
- Nino Shervashidze and Karsten Borgwardt. Fast subtree kernels on graphs. In *Advances in neural information processing systems*, pages 1660–1668, 2009.
- Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12: 2539–2561, 2011.
- Muhan Zhang and Yixin Chen. Weisfeiler-lehman neural machine for link prediction. In *KDD*, 2017.
- Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In *ICML*, 2004.
- Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.
- Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. *ICWSM*, 2018.
- Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6861–6871. PMLR, 2019c.
- Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
- CSIRO’s Data61. Stellargraph machine learning library. <https://github.com/stellargraph/stellargraph>, 2018.
- Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *UAI*, 2012.
- Michel Besserve and Bernhard Schölkopf. Learning soft interventions in complex equilibrium systems. In *Uncertainty in Artificial Intelligence*, pages 170–180. PMLR, 2022.
- Mark W. Schmidt and Kevin P. Murphy. Modeling discrete interventional data using directed cyclic graphical models. In *UAI*, 2009.

APPENDICES

Appendix A

COPYRIGHT

Publication Agreement

This is a publication agreement¹ (“this agreement”) regarding a written manuscript currently entitled

Non-Parametric Inference of Relational Dependence

(“the article”) to be published in PMLR (“the proceedings”). The parties to this Agreement are:

Ragib Ahsan

(name of corresponding author who signs on behalf of any other authors, collectively “you”) and PMLR, (“the publisher”).

1. By signing this form, you warrant that you are signing on behalf of all authors of the article, and that you have the authority to act as their agent for the purpose of entering into this agreement.
2. You hereby grant a Creative Commons copyright license in the article to the general public, in particular a Creative Commons Attribution 4.0 International License, which is incorporated herein by reference and is further specified at <http://creativecommons.org/licenses/by/4.0/legalcode> (human readable summary at <http://creativecommons.org/licenses/by/4.0>).
3. You agree to require that a citation to the original publication of the article in the proceedings as well as a hyperlink to the PMLR web site linking to the original paper be included in any attribution statement satisfying the attribution requirement of the Creative Commons license of paragraph 2.
4. You retain ownership of all rights under copyright in all versions of the article, and all rights not expressly granted in this agreement.
5. To the extent that any edits made by the publisher to make the article suitable for publication in the proceedings amount to copyrightable works of authorship, the publisher hereby assigns all right, title, and interest in such edits to you. The publisher agrees to verify with you any such edits that are substantive. You agree that the license of paragraph 2 covers such edits.


¹The language of this publication agreement is based on Stuart Shieber’s model open-access journal publication agreement, version 1.2, available at <http://bit.ly/1m9UsNt>.

6. You further warrant that:

1. The article is original, has not been formally published in any other peerreviewed journal or in a book or edited collection, and is not under consideration for any such publication.
2. You are the sole author(s) of the article, and that you have a complete and unencumbered right to make the grants you make.
3. The article does not libel anyone, invade anyone's copyright or otherwise violate any statutory or common law right of anyone, and that you have made all reasonable efforts to ensure the accuracy of any factual information contained in the article. You agree to indemnify the publisher against any claim or action alleging facts which, if true, constitute a breach of any of the foregoing warranties or other provisions of this agreement, as well as against any related damages, losses, liabilities, and expenses incurred by the publisher.

7. This is the entire agreement between you and the publisher, and it may be modified only in writing. It will be governed by the laws of the Commonwealth of Massachusetts. It will bind and benefit our respective assigns and successors in interest, including your heirs. It will terminate if the publisher does not publish, in any medium, the article within one year of the date of your signature.

I HAVE READ AND AGREE FULLY WITH THE TERMS OF THIS AGREEMENT.

- Corresponding Author:
 - Signed: 
 - Date: 06/17/22

Publication Agreement

This is a publication agreement¹ (“this agreement”) regarding a written manuscript currently entitled

Relational Causal Models with Cycles: Representation and Reasoning

(“the article”) to be published in PMLR (“the proceedings”). The parties to this Agreement are:

Ragib Ahsan

(name of corresponding author who signs on behalf of any other authors, collectively “you”) and PMLR, (“the publisher”).

1. By signing this form, you warrant that you are signing on behalf of all authors of the article, and that you have the authority to act as their agent for the purpose of entering into this agreement.
2. You hereby grant a Creative Commons copyright license in the article to the general public, in particular a Creative Commons Attribution 4.0 International License, which is incorporated herein by reference and is further specified at <http://creativecommons.org/licenses/by/4.0/legalcode> (human readable summary at <http://creativecommons.org/licenses/by/4.0>).
3. You agree to require that a citation to the original publication of the article in the proceedings as well as a hyperlink to the PMLR web site linking to the original paper be included in any attribution statement satisfying the attribution requirement of the Creative Commons license of paragraph 2.
4. You retain ownership of all rights under copyright in all versions of the article, and all rights not expressly granted in this agreement.
5. To the extent that any edits made by the publisher to make the article suitable for publication in the proceedings amount to copyrightable works of authorship, the publisher hereby assigns all right, title, and interest in such edits to you. The publisher agrees to verify with you any such edits that are substantive. You agree that the license of paragraph 2 covers such edits.


¹The language of this publication agreement is based on Stuart Shieber’s model open-access journal publication agreement, version 1.2, available at <http://bit.ly/1m9UsNt>.

6. You further warrant that:

1. The article is original, has not been formally published in any other peerreviewed journal or in a book or edited collection, and is not under consideration for any such publication.
2. You are the sole author(s) of the article, and that you have a complete and unencumbered right to make the grants you make.
3. The article does not libel anyone, invade anyone's copyright or otherwise violate any statutory or common law right of anyone, and that you have made all reasonable efforts to ensure the accuracy of any factual information contained in the article. You agree to indemnify the publisher against any claim or action alleging facts which, if true, constitute a breach of any of the foregoing warranties or other provisions of this agreement, as well as against any related damages, losses, liabilities, and expenses incurred by the publisher.

7. This is the entire agreement between you and the publisher, and it may be modified only in writing. It will be governed by the laws of the Commonwealth of Massachusetts. It will bind and benefit our respective assigns and successors in interest, including your heirs. It will terminate if the publisher does not publish, in any medium, the article within one year of the date of your signature.

I HAVE READ AND AGREE FULLY WITH THE TERMS OF THIS AGREEMENT.

- Corresponding Author:
 - Signed: 
 - Date: 04/29/22

Appendix B

VITA

Ragib Ahsan

✉ ragib.ahsan@gmail.com

📁 [ragib06.github.io](https://github.com/ragib06)

Research

Interests *Causal Inference, Relational Classification, Recommender Systems*
Focus *Causal discovery from relational data involving feedback loops*
Application *Social Science, Epidemiology*

Education

2022 **Ph.D.**, *Computer Science*, University of Illinois at Chicago.
Advised by Dr. Elena Zheleva
2021 **M.S.**, *Computer Science*, University of Illinois at Chicago.
In fulfillment of Ph.D. candidacy requirements, GPA – 3.87
2012 **B.S.**, *Computer Science and Engineering*, Bangladesh University of
Engineering and Technology, Bangladesh.
Major in AI, GPA – 3.65

Publications

UAI'22 ***Non-Parametric Inference of Relational Dependence***
R. Ahsan, Z. Fatemi, D. Arbour, E. Zheleva
38th conference on Uncertainty in Artificial Intelligence (UAI), 2022
CLear'22 ***Relational Causal Models with Cycle: Representation and Reasoning***
R. Ahsan, D. Arbour, E. Zheleva
1st conference on Causal Learning and Reasoning (CLear), 2022
MLG'20 ***Effectiveness of Sampling Strategies for One-shot Active Learning from Relational Data***
R. Ahsan, E. Zheleva
KDD Workshop on Mining and Learning with Graphs (MLG), 2020
WWW'20 ***Correcting for Selection Bias in Learning-to-Rank Systems***
Z. Ovaisi, **R. Ahsan**, E. Zhang, K Vasilaky, E. Zheleva
Proceedings of The World Wide Web Conference, 2020

Experience

2018–Present **Research Assistant**, DEPARTMENT OF COMPUTER SCIENCE, UIC.
◦ Primarily working on my dissertation research as part of the EDGES Lab group.
◦ Working on application of Machine Learning and Causal Inference in a privacy-based collaborative project.
Summer 2021 **Research Intern**, PINTEREST LABS, IL-Remote.
◦ Worked on a research project on spam behaviour analysis under the Trust & Safety team.
◦ Major contribution includes feature extraction for spam detection from longitudinal data.

- Summer 2019 **Research Intern**, DATA SCIENCE, Anthem Inc.
- Worked on a research project to build a medical knowledge graph for a specific medical condition from public medical journals.
 - Worked with state-of-the-art graph database (GraphDB) and PubMed API.
- 2015–2018 **Teaching Assistant**, DEPARTMENT OF COMPUTER SCIENCE, UIC.
- My primary job was to conduct labs, office hours and grading assignments and exams.
 - Major courses: Machine Learning, Data Structure.
- 2012–2013 **Software Engineer**, PLAYDOM, Disney Interactive, Bangladesh.
- Worked as a front end developer in both mobile and web platforms to contribute to some world class social games by Disney.
- 2011 **Student Developer**, APERTIUM, *Google Summer of Code 2011*.
- Contributed to an open source machine translation project named "Apertium" to introduce Bengali-English language pair.

Academic Projects

- CS 594 De-bias Learning-To-Rank algorithm with Yahoo Learning To Rank (YLTR) dataset.
- CS 594 Implemented Joint-RNN and LSTM for keyphrase detection from tweets.
- CS 421 Implemented Mini Watson - an NLP based small scale question answering system.
- CS 412 Participated in an archived Kaggle challenge on Cause-Effect Pair detection.

Honors and Awards

- 2020 Conference Sub-reviewer in *KDD, WSDM, WWW, AISTATS*
- 2018 Finalist (Top 50) in ITA Tech Challenge 2015 and 2018.
- 2011 Ranked 10th at ITEE contest organized by MOSICT and JICA.
- 2009 Ranked 4th in 2nd IUT ICT Fest Programming Contest, Bangladesh.
- 2009 Ranked 8th in ACM ICPC Regional: Dhaka Site, Bangladesh.

Technical skills

- Programming C++, PYTHON, OBJECTIVE-C, JAVA, JAVASCRIPT, MATLAB
- Library PyTorch, StellarGraph, Networkx, MPI, PETSc, Scikit-learn, NLTK
- Cloud Parse, Google App Engine, OpenShift, AWS

References

Dr. Elena Zheleva
Assistant Professor
Department of Computer Science
University of Illinois at Chicago
Chicago, IL, USA
✉ ezheleva@uic.edu

Dr. David Arbour
Research Scientist
Adobe Research
California, USA
✉ arbour@adobe.com