Using Noisy Extractions to Discover Causal Knowledge

Dhanya Sridhar
University of California Santa Cruz
dsridhar@soe.ucsc.edu

Jay Pujara
Information Sciences Institute
jay@cs.umd.edu

Lise Getoor
University of California Santa Cruz
getoor@soe.ucsc.edu

1 Introduction

Knowledge bases (KB) constructed through information extraction from text play an important role in query answering and reasoning. Since automatic extraction methods yield results of varying quality, typically, only high-confidence extractions are retained, which ensures precision at the expense of recall. Consequently, the noisy extractions are prone to propagating false negatives when used for further reasoning. However, in many problems, empirical observations of entities, or observational data, are readily available, potentially recovering information when fused with noisy extractions. For reasoning tasks where both empirical observations and extractions can be obtained, an open and critical problem is designing methods that exploit both modes of identification.

In this work, we study a particular reasoning task, the problem of discovering causal relationships between entities, known as causal discovery. There are two contrasting types of approaches to discovering causal knowledge. One approach attempts to identify causal relationships from text using automatic extraction techniques, while the other approach infers causation from observational data. For example, prior extraction-based approaches have mined causal links such as regulatory relationships among genes directly from scientific text [9, 12]. However, the extracted links often miss complex and longer-range patterns that require observational data. On the other hand, given observations alone, extensive work has studied the problem of inferring a network of cause-and-effect relationships among variables [13, 2]. Observational data such as gene expression measurements are used to infer causal relationships such as gene regulation [7]. Prior approaches use constraints to find valid causal orientations from observational data [5, 6, 7, 13, 4]. Although the constraints offer attractive soundness guarantees, the need for observed measurements of variables remains costly and prohibitive when experimental data is unpublished. Extractions such as interactions between genes mined directly from text provide a coarse approximation of unseen observational data. Combining extractions mined from KBs with observed measurements where available to for causal discovery can alleviate the cost of obtaining experiment-based data.

We propose an approach for fusing noisy extractions with observational data to discover causal knowledge. We introduce CAUSFUSE, a probabilistic model over causal relationships that combines commonly used constraints over observational data with extractions obtained from a KB. CAUSFUSE uses the probabilistic soft logic (PSL) modeling framework to express causal constraints in a natural logical syntax that flexibly incorporates both observational and KB modes of evidence. As our main contributions:

1. We introduce the novel problem of combining noisy extractions from a KB with observational data.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2. We propose a principled approach that uses well-studied causal discovery constraints to recover long-range patterns and consistent predictions, while cheaply acquired extractions provide a proxy for unseen observations.
3. We apply our method gene regulatory networks and show the promise of exploiting KB signals in causal discovery, suggesting a critical, new area of research.

We compare CAUSFUSE with a conventional logic-based approach that uses only observational data to perform causal discovery. We evaluate both methods on transcriptional regulatory networks of yeast. Our results validate two strengths of our approach: 1) CAUSFUSE achieves comparable performance with the well-studied conventional method, suggesting that noisy extractions are useful approximations for unseen empirical evidence; and 2) global logical constraints over observational data enforce consistency across predictions and bolster CAUSFUSE to perform on par with the competing method. The results suggest promising new directions for integrating knowledge bases in causal reasoning, potentially mitigating the need for expensive observational data.

2 Background on Logical Causal Discovery

The inputs to traditional causal discovery methods are m independent observations of n variables $\mathbf{V}$. The problem of causal discovery is to infer a directed acyclic graph (DAG) $\mathcal{G}^* = (\mathbf{V}, \mathbf{E})$ such that each edge $E_{ij} \in \mathbf{E}$ corresponds to V_i being a *direct cause* of V_j, where changing the value of V_i always changes the value of V_j.

Since graphical model $\mathcal{G}^*$ encodes conditional independences among $\mathbf{V}$, causal discovery algorithms exploit the mapping between observed independences in the data and paths in $\mathcal{G}^*$ to specify constraints on the output. The PC algorithm [13] is a canonical such method that performs independence tests on the observations to rule out invalid causal edges. Constraints over causal graph structure can also be encoded with logic [5, 6, 7]. In a logical causal discovery system, independence relations are represented as logical atoms. Logical atoms consist of a predicate symbol $p_i(\cdot)$ with i variable or constant arguments and take boolean or continuous truth values. To avoid confusion with logical variables, for the remainder of this paper, we refer to $V \in \mathbf{V}$ as vertices. As inputs to logical causal discovery, we require the following predicates to represent the outcomes of independence tests among $\mathbf{V}$:

- DEP(A, B), INDEP(A, B) refers to statistical (in)dependence between vertices V_A and V_B as measured by the independence test $V_A \perp\!\!\!\perp V_B$. The conditioning set is the empty set.
- CONDDEP(A, B, S), CONDINDEP(A, B, S) corresponds to statistical (in)dependence between vertices V_A and V_B when conditioned on set $\mathbf{S} \subset \mathbf{V} \setminus \{V_A, V_B\}$. The independence test $V_A \perp\!\!\!\perp V_B | \mathbf{S}$ is performed.

The outputs are of a logical causal discovery system are represented by the following *target* predicates:

- CAUSES(A, B) refers to the absence or presence of a causal edge between V_A and V_B, and is substituted with all pairs of vertices $A, B \in \mathbf{V}$. Finding truth value assignments to these atoms is the goal of causal discovery.
- ANCESTOR(A, B) corresponds to the absence or presence of an *ancestral* edge between all vertices V_A and V_B, where V_A is an ancestor of V_B if there is a directed causal path from V_A to V_B. We may additionally infer the truth values of ancestral atoms jointly with causal atoms.

Given the independence tests over $\mathbf{V}$ as input, the goal of logical causal discovery is to find consistent assignments to the causal and ancestral output atoms.

3 Using Extractions in Causal Discovery

In the problem of fusing noisy extractions with causal discovery, in addition to the observations, we are given a set of variables $\mathbf{K} = \{K_{11} \dots K_{nn}\}$ of evidence from knowledge base (KB) $\mathcal{K}$, where K_{ij} is an affinity score of the interaction between V_i and V_j based on text extraction.

Extending previous logical causal discovery methods, we additionally represent $\mathbf{K}$ in the predicate set with TEXTADJ(A, B). TEXTADJ(A, B) corresponds to K_{AB} and denotes the absence or presence of an undirected edge, or adjacency, between V_A and V_B as extracted from text. Evidence of adjacencies is critical to inference of CAUSES(A, B). However, adjacencies in standard causal discovery are inferred from statistical tests alone. In our approach, we replace statistical adjacencies with TEXTADJ(A, B). The goal of fusing KB evidence in logical causal discovery is to find maximally satisfying assignments to the unknown causal atoms based on constraints over both independence and text-based signals. In section 4, we present a probabilistic logic approach defining constraints using statistical and KB evidence.

4 A Probabilistic Approach to Inferring Causal Knowledge

Our approach uses probabilistic soft logic (PSL) [1] to encode constraints for causal discovery. A key advantage of PSL is exact and efficient MAP inference for finding most probable assignments. We first review PSL and then present our novel encoding constraints that combine statistical and KB information.

4.1 Probabilistic Soft Logic

PSL is a probabilistic programming framework where random variables are represented as logical atoms and dependencies between them are encoded via rules in first-order logic. Logical atoms in PSL take continuous values and logical satisfaction of the rule is computed using the Lukasiewicz relaxation of Boolean logic. This relaxation into continuous space allows MAP inference to be formulated as a convex optimization problem that can be solved efficiently.

Given continuous evidence variables $\mathbf{X}$ and unobserved variables $\mathbf{Y}$, PSL defines the following Markov network, called a hinge-loss Markov random field (HL-MRF), over continuous assignments to $\mathbf{Y}$:

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\Big(-\sum_{r=1}^{M} w_r \phi_r(\mathbf{y}, \mathbf{x}) \Big) , \tag{1}$$

where $\mathcal{Z}$ is a normalization constant, and $\phi_r(\mathbf{y}, \mathbf{x}) = (\max\{l_r(\mathbf{y}, \mathbf{x}), 0\})$ is an efficient-to-optimize *hinge-loss* feature function that scores configurations of assignments to $\mathbf{X}$ and $\mathbf{Y}$ as a linear function l_r of the variable assignments.

An HL-MRF is defined by PSL model $\mathcal{M} = \{(R_1, w_1) \dots (R_m, w_m)\}$, a set of m weighted disjunctions, or rules, where w_i is the weight of i-th rule. Rules consist of logical atoms and are called *ground rules* if only constants appear in the atoms. To obtain the HL-MRF, we first substitute logical variables appearing in $\mathcal{M}$ with constants from observations, producing M ground rules. We observe truth values $\in [0, 1]$ for a subset of the ground atoms, $\mathbf{X}$ and infer values for the remaining unobserved ground atoms, $\mathbf{Y}$. The ground rules and their corresponding weights map to ϕ_r and w_r. To derive $\phi_r(\mathbf{y}, \mathbf{x})$, the Lukasiewicz relaxation is applied to each ground rule to derive a hinge penalty function over $\mathbf{y}$ for violating the rule. Thus, MAP inference minimizes the weighted rule penalties to find the minimally violating joint assignment for all the unobserved variables:

$$\arg \min_{\mathbf{y} \in [0,1]^n} \sum_{r=1}^{m} w_r \max\{l_r(\mathbf{y}, \mathbf{x}), 0\}$$

PSL uses the consensus based ADMM algorithm to perform exact MAP inference.

4.2 CAUSFUSE

CAUSFUSE extends constraints introduced by the PC algorithm [13]. Whereas PC infers adjacencies from conditional independence tests, CAUSFUSE uses text-based adjacency evidence in all causal constraints. The text-based adjacency evidence bridges domain-knowledge contained in KBs with statistical tests that propagate causal information.

Figure 1 shows all the rules used in CAUSFUSE. The first set of rules follow directly from the three constraints introduced by PC. We additionally introduce joint rules to induce dependencies between ancestral and causal structures to propagate consistent predictions. We describe below how CAUSFUSE rules upgrade PC to combine KB and statistical signals for causal discovery.

Figure 1: PSL rules for combining statistical tests and KB evidence in causal discovery.

Rule Type	Rules
PC-inspired Rules	C1) $\neg\textsc{TextAdj}(A, B) \rightarrow \neg\textsc{Causes}(A, B)$
	C2) $\textsc{Causes}(A, B) \rightarrow \neg\textsc{Causes}(B, A)$
	C3) $\textsc{TextAdj}(A, B) \wedge \textsc{TextAdj}(C, B) \wedge \neg\textsc{TextAdj}(A, C) \wedge \textsc{CondDep}(A, C, S) \wedge \textsc{InSet}(B, S) \rightarrow \textsc{Causes}(A, B)$
	C4) $\textsc{TextAdj}(A, B) \wedge \textsc{TextAdj}(C, B) \wedge \neg\textsc{TextAdj}(A, C) \wedge \textsc{CondDep}(A, C, S) \wedge \textsc{InSet}(B, S) \rightarrow \textsc{Causes}(C, B)$
	C5) $\textsc{Causes}(A, B) \wedge \textsc{Dep}(A, C) \wedge \textsc{CondIndep}(A, C, S) \wedge \textsc{InSet}(B, S) \wedge \textsc{TextAdj}(B, C) \rightarrow \textsc{Causes}(B, C)$
	C6) $\textsc{Causes}(A, B) \wedge \textsc{Causes}(B, C) \wedge \textsc{TextAdj}(A, C) \rightarrow \textsc{Causes}(A, C)$
Joint Rules	J1) $\textsc{Causes}(A, B) \rightarrow \textsc{Anc}(A, B)$
	J2) $\neg\textsc{Anc}(A, B) \rightarrow \neg\textsc{Causes}(A, B)$
	J3) $\textsc{Anc}(A, B) \wedge \textsc{Anc}(B, C) \rightarrow \textsc{Anc}(A, C)$
	J4) $\textsc{Anc}(A, B) \wedge \textsc{TextAdj}(A, B) \rightarrow \textsc{Causes}(A, B)$
	J5) $\textsc{TextAdj}(A, B) \wedge \textsc{TextAdj}(B, C) \wedge \textsc{Dep}(A, C) \wedge \textsc{CondIndep}(A, C, S) \wedge \textsc{InSet}(B, S) \wedge \textsc{Causes}(B, A) \wedge \neg\textsc{Anc}(C, A) \rightarrow \textsc{Causes}(B, C)$

PC-inspired Rules PC uses conditional (in)dependence and adjacency to rule out violating causal orientations. However, in CAUSFUSE, all adjacencies are directly mined from a KB. Rule C1 discourages causal edges between vertices that are not adjacent based on evidence in text. Rule C2 penalizes simple cycles between two vertices. Rules C3 and C4 capture the first PC rule and orient chain $V_i - V_j - V_k$ as $V_i \rightarrow V_j \leftarrow V_k$, a v-structure, based on independence criteria. Rule C5 orients path $V_i \rightarrow V_j - V_k$ as $V_i \rightarrow V_j \rightarrow V_k$ to avoid orienting additional v-structures. Rule C6 maps to the third PC rule, and if $V_i \rightarrow V_j \rightarrow V_k$ and $V_i - V_k$, orients $V_i \rightarrow V_k$ to avoid a cycle. PC applies these rules iteratively to fix edges whereas in CAUSFUSE, the rules induce dependencies between causal edges to encourage parsimonious joint inferences.

Joint Rules Joint rules encourage consistency across ancestral and causal predictions through constraints such as transitivity that follow from basic definitions. Rule J1 encodes that causal edges are also ancestral by definition and rule J2 is the contrapositive that penalizes causal edges to non-descendants. Rule J3 encodes transitivity of ancestral edges, encouraging consistency across predictions. Rule J4 infers causal edges between probable ancestral edges that are adjacent based on textual evidence. Rule J5 orients chain $V_i - V_j - V_k$ as a diverging path $V_i \leftarrow V_j \rightarrow V_k$ when V_k is not likely an ancestor of V_i. Joint rules give preference to predicted structures that respect both ancestral and causal graphs.

In our evaluation, we investigate the implications of using a noisy extraction-based proxy for adjacency and the benefits of joint modeling.

5 Experimental Evaluation

Our experiments investigate the two main claims of our approach:

1. We study whether the noisy extractions are a suitable proxy for latent adjacencies and give similar performance to a conventional logic-based approach that impute adjacency values using only observations.
2. We understand the role of joint ancestral and causal rules over observational data in mitigating noise from the extraction-based evidence.

We evaluate CAUSFUSE on real-world gene regulatory networks in yeast. We compare against CAUSOBS, the PSL model variant that performs prototypical causal discovery using only observational data. CAUSOBS replaces TEXTADJ with STANDARDADJ, adjacencies computed from conditional independence tests.

5.1 Data

Our dataset for evaluation consists of a transcriptional regulatory network across 300 genes in yeast with simulated gene expression from the DREAM4 challenge [8, 10]. We snowball sample 10 smaller subnetworks of sizes 20 with low Jaccard overlap to perform cross validation. The data contains 210 gene expression measurements simulated from differential equation models of the system. We

perform independence tests on the real-valued measurements which are known to contribute numerous spurious correlations. In addition to the gene expression data, we model domain knowledge based on undirected protein-protein interaction (PPI) edges extracted from the Yeast Genome Database:

$$\text{ANC}(A, B) \wedge \text{LOCAL}_{\text{PPI}}(A, B) \rightarrow \text{CAUSES}(A, B)$$

We obtain text-based affinity scores of interaction between pairs of yeast genes from the STRING database. STRING finds mentions of gene or protein names across millions of scientific articles and computes the co-occurrence of mentions between genes. As an additional step, STRING extracts relations between genes and increases the affinity score if genes are connected by salient terms such as "binds to" or "phosphorylates."

5.2 Results

Model Variant	F_1
CAUSFUSE	0.19 ± 0.08
CAUSOBS	0.20 ± 0.05
CAUSFUSES-PC	0.17 ± 0.07
CAUSOBS-PC	0.19 ± 0.05

Table 1: CAUSFUSE achieves comparable performance with CAUSOBS, suggesting that noisy extractions can approximate unseen adjacencies. Without joint rules, CAUSFUSES-PC shows worse performance, pointing to the benefit of sophisticated joint modeling in mitigating noisy extractions.

We evaluate CAUSFUSE and CAUSOBS using 10-fold cross validation on DREAM4 networks. CAUSOBS uses the same rules as our approach but computes STANDARDADJ as ground $\text{DEP}(A, B)$ atoms that never appear in groundings of $\text{CONDINDEP}(A, B, S)$, based on definition.

To evaluate the additional benefit of joint rules, we compare sub-models of CAUSFUSE and CAUSOBS run with causal orientation rules only, denoted CAUSFUSES-PC and CAUSOBS-PC respectively. Table 1 shows average F_1 scores of all model variants for the regulatory network prediction task on DREAM4.

Noisy Extractions Maintain Performance First, we see comparable performance between CAUSFUSE and CAUSOBS, answering our first experimental question on how closely noisy extractions approximate adjacencies. In table 1, there is no statistically significant difference between the F_1 scores of CAUSFUSE and CAUSOBS. The comparable performance between CAUSFUSE and CAUSOBSsuggests that the noisy extractions can substitute observational data computations without significantly degrading performance.

Joint Rules Overcome Noise Our investigation into model variants sheds light on the second experimental question around how logical rules overcome the noise from extractions. When comparing PC-only variants of each method, CAUSOBS-PC gains over CAUSFUSES-PC, suggesting that sophisticated joint rules are needed to mitigate the noise from KB extractions. The consistency across predictions encouraged by the joint rules bolsters the extraction-based adjacency signal.

Extractions Yield Higher Precision, Lower Recall To further investigate the extraction evidence mined from STRING, we compare both STANDARDADJ and TEXTADJ against gold-standard adjacencies, which we obtain from undirected regulatory links. Table 2 shows the average precision and

Adjacency	**Precision**	**Recall**
TEXTADJ	0.32	0.11
STANDARDADJ	0.27	0.3

Table 2: Extraction-based adjacencies achieve higher-precision but lower recall, further substantiating need for joint rules in recovering missing causal orientations.

recall of each adjacency evidence type across the DREAM4 subnetworks. Interestingly, TEXTADJ achieves higher precision than its statistical counterpart. However, STANDARDADJ gains over TEXTADJ in recall. The result further substantiates the benefit of joint modeling in recovering additional orientations under low-recall inputs. Nonetheless, the comparison points to the need for a deeper understanding of the role KBs play in causal reasoning.

5.3 Experiment Details

To obtain marginal and conditional (in)dependence tests, we use linear and partial correlations with Fisher's Z transformation. We condition on all sets up to size two. We set rule weights for both PSL models to 5.0 except for rule C2 which is set to 10.0, since it encodes a strong acyclicity constraint. Both models use an α threshold on the p-value to categorize independence tests as CONDDEP,DEP or CONDINDEP,INDEP. We select α with 10-fold cross validation. We hold out each subnetwork in turn and use the best average F_1 score across the other subnetworks to pick $\alpha \in \{0.1, 0.05\}$ raised to powers $\{1, 2, 3, 4, 5\}$. CAUSOBS selects two different α values for binning independence tests and computing adjacencies, and CAUSFUSE requires a single α for tests only. We also select rounding thresholds for both PSL models within the same cross-validation framework. Since α is typically small, we rescale truth values p for CONDINDEP,INDEP by $\sqrt[3]{p}$ to reduce right-skewness of values. We rescale all STRING affinity scores to be between 0 and 1.

6 Related Work

Our work extends constraint-based methods to causal discovery, most notably the PC algorithm [13], which first infers adjacencies and maximally orients them using deterministic rules based on conditional independence. PC only supports external evidence in the form of fixed edges or non-edges. Our work is motivated by recent approaches that cast causal discovery as a SAT instance over conditional independence statements [6, 7, 5]. SAT-based approaches are based on logical representations that more readily admit additional constraints and relations from domain knowledge. However, so far, logical causal discovery methods use external evidence to identify probable edges.

In a separate vein, prior work has extended text-mining to identify regulatory networks and genetic interactions only from scientific literature [11, 12, 9]. In contrast, our goal is to propose techniques that leverage both statistical test signals and text evidence. The work most similar to ours combines gene expression data with evidence mined from knowledge bases to infer gene regulatory networks [3]. However, the regulatory network inference orients edges using hard-coded knowledge of transcription factors instead of reasoning about causality. In our approach, we propose a principled causal discovery formulation as the basis of incorporating KB evidence.

7 Discussion and Future Work

In this work, we present an initial approach for reasoning with noisy extraction-based evidence directly in a logical causal discovery system. We benefit from a flexible logical formulation that supports replacing conventional adjacencies computed from observational data with cheaply obtained extractions. Our evaluation suggests that the noisy KB-based proxy signal achieves comparable performance to conventional methods. The promising result points to future research in exploiting KBs for causal reasoning, greatly mitigating the need for costly observational data. We see many directions of future work, including better extraction strategies for mining scientific literature and finding text-based proxies for additional statistical test signals. KBs could provide ontological constraints or semantic information useful for causal reasoning. We additionally plan to study knowledge-based constraints for causal discovery.

8 Acknowledgments

This work is supported by NSF grants CCF-1740850 and NSF IIS-1703331.

References

[1] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research (JMLR)*, 2017. To appear.

[2] David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[3] Panagiotis Chouvardas, George Kollias, and Christoforos Nikolaou. Inferring active regulatory networks from gene expression data using a combination of prior knowledge and enrichment analysis. *BMC bioinformatics*, 17(5):181, 2016.

[4] Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *UAI*, 2011.

[5] Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Jarvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *UAI*, 2013.

[6] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, 2014.

[7] Sara Magliacane, Tom Claassen, and Joris M Mooij. Ancestral causal inference. In *NIPS*, 2016.

[8] Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107:6286–6291, 2010.

[9] Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. Literome: Pubmed-scale genomic knowledge base in the cloud. *Bioinformatics*, 30(19):2840–2842, 2014.

[10] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5:e9202, 2010.

[11] Carlos Rodríguez-Penagos, Heladia Salgado, Irma Martínez-Flores, and Julio Collado-Vides. Automatic reconstruction of a bacterial regulatory network using natural language processing. *BMC bioinformatics*, 8(1):293, 2007.

[12] Yong-Ling Song and Su-Shing Chen. Text mining biomedical literature for constructing gene regulatory networks. *Interdisciplinary Sciences: Computational Life Sciences*, 1(3):179–186, 2009.

[13] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9:62–72, 1991.

[14] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, 2017.