Using a General Prior Knowledge Graph to Improve Data-Driven Causal Network Learning

Meghamala Sinha[a], Stephen A. Ramsey[a,b]

[a]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, United States
[b]Department of Biomedical Sciences, Oregon State University, Corvallis, OR 97331, United States

Abstract

We describe a method "Kg2Causal" for using a large-scale, general-purpose biomedical knowledge graph as a prior for data-driven causal network structure learning. Given a set of observed nodes in a dataset, and some relationship edges between the nodes derived from a knowledge graph, Kg2Causal uses the knowledge graph-derived edges to guide the data-driven inference of a causal Bayesian network. We tested Kg2Causal on several real-world biological datasets with known ground-truth networks and demonstrate improvement in network learning accuracy, relative to a baseline of an uninformative network structure prior. We also demonstrate the application of our method if data are collected under different experimental conditions including interventions on the observed variables.

Keywords

Causal inference, Structure learning, Knowledge graph, Informative prior

1. Introduction

Causal modeling is a useful analytical tool in various fields due to its applicability in action planning, prediction and diagnosis [1, 2, 3, 4, 5]. However, learning a causal Bayesian network (CBN) solely from data is a challenging task [6, 7, 8]. CBN learning can be thought of as model selection problem in which model is a directed acyclic graph (DAG), where problem is to find the graph G that maximizes some objective (score) function of dataset D. In some CBN learning methods, score function is likelihood $p(D \mid G)$ representing overall fit of G to D in context of a generative model for the data. For a dataset with n observables (features), the number of possible DAGs—and thus requirement for data—grows super-exponentially with n [9]. In most network learning applications, prior knowledge exists about causal (or suspected causal) relationships among the observables; such prior knowledge can be valuable resource for network structure learning [10]. Supposing that prior knowledge can be represented as prior probability $p(G)$ on network structure, one can alternatively choose as basis for CBN scoring function, the *posterior* probability $p(G \mid D) = p(D \mid G)\, p(G)/p(D)$. In contrast to substantial amount of work done on variety of marginal likelihood and scoring methods, less attention has been given to functional form (and associated parameterization) of the prior $p(G)$ for application contexts

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

sinham@oregonstate.edu (M. Sinha)

where structured prior knowledge is available. Without expert knowledge, standard network inference approaches, by default, assume uniform (uninformative) prior which can lead to erroneous relationships or relationship orientations both due to (i) size of space of networks and (ii) degeneracy of Markov-equivalent networks. Proper incorporation of informative priors can enhance model efficiency [11] and can also overcome weakness of smaller dataset.

For most applications of causal modeling, some prior knowledge is available. For example, in medicine, most cases have prior knowledge about etiology, symptoms, and treatment of underlying diseases or conditions which can be obtained from biomedical literature or knowledge-bases. Although there is in general large scale availability of structured prior knowledge (for example, ontologies) in various scientific domains, these mostly comprise disparate information sources in various standards and formats, which poses a challenge to integrate them into single structure. These problems motivated building of large multi-graphs called knowledge graphs (KG) [12] that incorporate structured knowledge from multiple sources within a consistent schema. Knowledge graph is a term of art to mean a large graph-structured model to store interlinked relationships between nodes representing concepts [13]. These large-scale networks accommodate structural information which can be leveraged for reasoning, recommendation or decision making. We hypothesized that combining information from structured databases of general prior knowledge with causal modeling based on context-specific multivariate measurements will improve accuracy of learned network compared to result of data-driven causal modeling without incorporating prior knowledge.

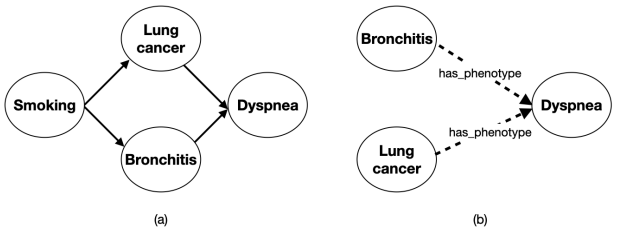

Figure 1: (a) A network containing known relationships between lungs condition and diseases, (b) corresponding sub-graph in a knowledge graph

In this work, we propose a method, "Kg2Causal", for extracting relations as pairs of nodes from a knowledge graph, and for incorporating them as priors on corresponding edges in a score-based, data-driven causal network learning method. In this study, prior edges from knowledge graph are accounted for in prior probability of the graph, using the method of Castelos and Siebes [11]. We used a large-scale biomedical knowledge graph (KG) [1] that we and collaborators (see Acknowledgments) had previously constructed (see Sec. 2.5) containing millions of nodes representing drugs, genes, diseases, or phenotypes (Fig. 1b), as well as edges between nodes representing various types of predicate relationships. For the measurement-based network learning component of KG2Causal, we used an optimizing method combining Tabu search algorithm [14] with Bayesian Dirichlet uniform (BDeu) [15, 16] network score. Using five different multivariate molecular biology datasets for which ground-truth networks were available (see Sec. 4), we empirically analysed network learning accuracy of "Kg2Causal" along with various types of uninformative prior to test the usefulness of adding KG-based priors. We provide a comparative benchmark of our methods performance over five real-world biological datasets and two synthetic datasets of varying sizes where we found

[1] https://github.com/RTXteam/RTX/code/reasoningtool/kg-construction

that Kg2Causal had superior network learning accuracy to methods that do not use general knowledge-base as network structure prior. Finally, we demonstrate (Sec. 4.3) the application of Kg2Causal if data are collected under different experimental conditions including *interventions* on the observed variables. We implemented "Kg2Causal" in the R programming language (leveraging the `bnLearn` package [17]) and provide the code as open-source software [2].

2. Related work and Background

In this section, we describe Kg2Causal's conceptual foundations including CBNs, score-based causal modeling, interventions, and knowledge graph-based priors in network learning.

2.1. Causal network: Brief Overview

A causal Bayesian Network [1, 2] is a DAG $G = (V, E)$, where $V = \{V_1, \ldots, V_n\}$ denotes the set of variables (nodes) and $E \subset V \times V$ denotes the causal relationships (edges). For an edge (V_i, V_j), we say that V_i is a parent (cause) of V_j, and V_j is a child (effect) of V_i. We will use $\mathrm{Pa}(V_i)$ to denote the set of parents of V_i. The conditional probability distribution p_i defines the probability of V_i given the states of its parents $\mathrm{Pa}(V_i)$. A causal network represents a joint distribution p over variables V as long as it satisfies two main assumptions:

a) *Causal Markov:* Any given variable V_i is independent of its non-descendants, conditioned on all of its direct causes. The assumption implies that the joint distribution $p(V)$ can be factored as: $p(V) = \prod_{i=1}^{n} p_i(V_i \mid \mathrm{Pa}(V_i))$.

b) *Faithfulness:* The joint distribution $p(V_1, \ldots, V_n)$ is *faithful* to G if every conditional independence relation in p is entailed by the causal Markov assumption applied to G [18].

2.2. Constructing a causal network

Let us assume, we have a dataset D having observations over set of n variables. One of the main classes of causal learning approaches is the *Score-based* approach which is derived from classic Bayesian method where a scoring function evaluates the fit of graph G to data D [16, 6] with a higher value indicating better fit. A search algorithm is used to explore the space of all possible graphs, to maximize the scoring function. Typical heuristic algorithms used for this purpose include hill-climbing or Tabu search approaches [14]. Other common score based methods are GDS [19] and Gies [6]. According to standard Bayesian rule, a causal graph G, is a *DAG* learned from given data D as $p(G \mid D) \propto p(G)p(D \mid G)$, where $p(G)$ is prior distribution over space of all possible DAGs reflecting prior knowledge and $p(D \mid G)$ is marginal likelihood of the data D. As described in Sec. 3, the Kg2Causal method incorporates a score-based approach.

2.3. Learning with interventions

Interventions—external manipulations of nodes *("targets")* in a network—are important to detect causal relations that can help disambiguate *Markov equivalent* sub-networks [16]. Let I_e represent set of target nodes that are altered in interventional experiment e and $O_e = V \backslash I_e$ be

[2] https://github.com/meghasin/Kg2Causal

the complementary set of observational variables. Each intervention can have one or more targets whose conditional probabilities are changed (so that, conditioned on intervention, target variable's distribution may depend only on a (possibly empty) subset of its parent observables). Hence, each intervention results in deletion of arrows pointing towards the intervened nodes. The joint distribution of p after intervention is $p(V_1, ..., V_n) = \prod_{V_i \in O_e} p_i(V_i \mid Pa(V_i)) \cdot \prod_{V_j \in I_e} p'_j(V_j \mid Pa(V_j))$, where $p(V_i \mid Pa(V_i))$ is conditional probability similar to V_i, given that V_i is not a target node, and $p'(V_j \mid Pa'(V_j))$ is post-intervention conditional probability of V_j given its new set of parents $Pa'(V_j)$. For a so-called "perfect" intervention, one would set $Pa'(V_j) = \emptyset$ [1]. Score-based approaches are well-suited to mixed interventional-observational datasets, in contrast to constraint-based approaches which are applicable to observational data.

2.4. Incorporation of Priors

In this subsection we introduce three types of uninformative priors on the network structure $p(G)$, uniform prior, marginal prior, and Bayesian variable selection prior (VSP). We then describe the knowledge graph-based prior that we use in Kg2Causal method. In cases for lack of prior knowledge, default choice for prior $p(G)$ is a **uniform prior** distribution, as follows:

$$\frac{p(E \cup \{V_i, V_j\} \mid D)}{p(E \mid D)} = \frac{p(E \cup \{V_i, V_j\})}{p(E)} \frac{p(D \mid E \cup \{V_i, V_j\})}{p(D \mid E)}$$

where nodes V_i and V_j can have three possible cases $V_i \Rightarrow V_j$ (representing $(V_i, V_j) \in E$),$V_i \Leftarrow V_j$ (representing $(V_j, V_i) \in E$) or $V_i \nRightarrow V_j$ (no arc) and each have equal probability of occurrence. So the probability for these edges are assigned as $p(V_i \Rightarrow V_j) = p(V_i \Leftarrow V_j) = p(V_i \nRightarrow V_j) = 1/3$, since we know that $p(V_i \Rightarrow V_j) + p(V_i \Leftarrow V_j) + p(V_i \nRightarrow V_j) = 1$. This implies $p(V_i \Rightarrow V_j) + p(V_i \Leftarrow V_j) = 2/3$, which means a higher promotion for the inclusion of new arcs and favouring the propagation of false positives in G. Hence, its not always a good idea to use uniform prior specially for cases where data is not too supportive of the DAG learned and where n is large. A better version of uniform prior is to use marginal probabilities instead, where an independent prior can be assumed for each arc with same independent marginal probabilities as uniform priors, also called **marginal uniform** [20]. In this case, the probability of inclusion of each edge is assigned as $p(V_i \Rightarrow V_j) = p(V_i \Leftarrow V_j) = 1/4$ and $p(V_i \nRightarrow V_j) = 1/2$. Compared to the uniform prior, the marginal uniform prior is less prone to false-positive edges in the posterior-probability-maximizing graph. The **Bayesian variable selection prior (VSP)** assigns a probability of inclusion of possible parent nodes, with the default being $1/n$.

The heart of Kg2Causal is the use of an informative prior based on a general-purpose knowledge graph; for this purpose we use an edge decomposition technique described by Castelo and Siebes [11]. For any pair of vertices (V_i, V_j) for which an edge $V_i \Rightarrow V_j$ exists in the general-purpose knowledge graph, we assign a prior probability ($\beta = 1/2$) on those edges, with probability 1/4 for $V_j \Rightarrow V_i$ and probability 1/4 for $V_i \nRightarrow V_j$, since the later two are alternate edges that have no corresponding edge in the general knowledge graph we use the uniform probability distribution as shown in Fig. 2. In this way we can create a complete prior probability (from partial knowledge) over the network G; on log scale, we define $p(G)$ as $\log\ p(G) = \sum_{V_i \Leftrightarrow V_j \in E,\ i \neq j} \log\ p(V_i \Leftrightarrow V_j) + \sum_{V_i \cdots V_j \in E,\ i \neq j} \log\ p(V_i \nRightarrow V_j)$.

2.5. Knowledge Graphs

A "knowledge graph" [13] is a multigraph consisting of nodes and edges (labeled by relationship type or description of instance attributes) between them. Although most relationships in knowledge graphs are between entities and context-based associations, these do not always imply causal relationship. Nevertheless, such links are strong association that can strengthen causal relationships that we seek to discover. The key idea of Kg2Causal is to use links from large knowledge graphs as generalised prior information to aid in data-driven network learning in highly specific application contexts. For this work, we leveraged a general biomedical knowledge graph that we and collaborators (see Acknowledgments) had constructed, KG1 [3]. KG1 has 130,443 nodes, 3.5M edges, 11 node semantic types, and 17 edge relation types, and was compiled from 20 different biomedical knowledge-bases (Monarch, COHD, ChEMBL, DGIdb, DisGeNet, Disease Ontology, GeneProf, HMDB, KEGG, miRBase, miRGate, mychem.info, mygene.info, NCBI Gene, OMIM, Pathway Commons, Pharos, PubChem, Reactome, and UniprotKB). We hosted KG1 in a Neo4j database (ver. 3.5.13) and used the Cypher query language to search for concept mappings between ground-truth network variables and concept nodes in the KG1 knowledge graph, and for edge connections between mapped concepts within KG1.

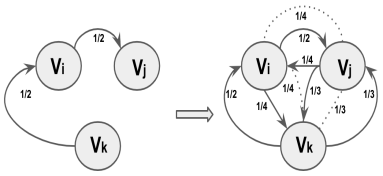

Figure 2: Complete prior by edge decomposition technique

3. Our Approach

We developed KG2Causal to leverage a general-purpose biomedical knowledge graph (see Sec. 2.5) in order to improve context-specific, data-driven network learning from multivariate observations; such observations could consist of gene expression measurements, proteomics measurements, or electronic health records. The key ideas of our approach are (i) mapping each variable in the dataset to a node in the knowledge graph, and querying relationships between them; (ii) extracting a subgraph containing the connected variables with edges between them; and (iii) use this edge set as our prior knowledge to guide the optimizing scoring step for inferring causal network. Mathematically, given a dataset D, with a set V of observable variables and given a general-purpose prior knowledge graph Γ as a multigraph, we want to learn a causal graph $G^f = (V, E)$ that approximately maximizes the posterior probability, i.e., $\text{argmax}_G(p(G \mid D, \Gamma))$, given a prior $p(G \mid \Gamma)$. As a comparison, we used three uninformative prior distributions, namely uniform, marginal and Bayesian variable selection priors with each dataset in order to understand whether or not—and to what extent—using an informative network prior improves accuracy of causal network learning in a biomedical context. The Kg2Causal network discovery workflow, illustrated in Figure 3, consists of the following steps:

- Map variables V to nodes in Γ, and extract a list β of edges from Γ among the nodes (collapsing same-direction multiedges to single edges).

[3] https://github.com/RTXteam/RTX/code/reasoningtool/kg-construction

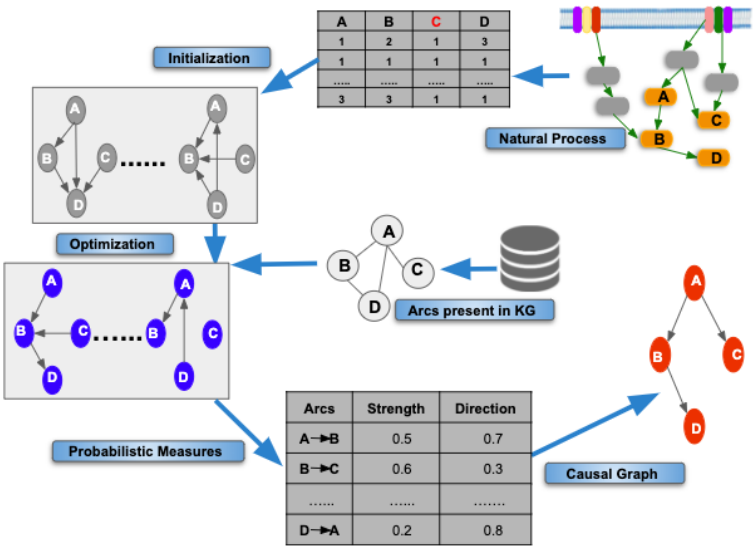

Figure 3: Workflow of Kg2Causal: **Step 1** - Data collection (can be a combination of observational and/or interventional studies). For example, in this figure, we show that node C (in red, top center) has been intervened. **Step 2** - Cleaning and discretizing the data. **Step 3** - Generating 100 random DAGs using the observed nodes. **Step 4** - Optimizing each of the 100 DAGs using Tabu search. In this step, we also extract edges present in the KG and incorporate them as a prior. **Step 5** - Calculating probability of occurrence for each possible arc in the 100 optimized DAGs. **Step 6** - Constructing the final network by selecting arc strengths above a threshold.

- Generate 100 random DAGs with nodes V. We empirically determined, based on our previous study [21], that this number is adequate for the medium-to-large datasets 4.
- In the score function, we include edge probability contributions from the prior knowledge graph (we assign probability 0.5 for every edge in β). For each DAG, we used the stochastic algorithm Tabu [14] to find a DAG that maximizes standard Bayesian Dirichlet equivalent uniform scoring function (BDeu) [15, 16].
- The previous step yields 100 optimized networks. Using these we compute the probability of each possible directed edge as its empirical frequency of occurrence among the DAGs. For example, if an edge (X, Y) appears in 80 out of 100 optimized DAGs, we assign it an empirical probability of 0.80. We store the edge probabilities in a list.
- We threshold the edge probabilities in order to obtain the set of edges E for G^f. Based on empirical studies, we chose a threshold of 0.85.

We chose Tabu for its robustness, simplicity (uses few parameters) and history-dependent ("memory"), although Kg2Causal is in principle compatible with any optimizing method.

3.1. Observational experiment

In the case where the dataset D is purely observational (i.e., no interventions) from a single experiment, Kg2Causal can be implemented algorithmically as described above; we provide a pseudocode description of the "observational" formulation of Kg2Causal in Algorithm 1.

Algorithm 1 Kg2Causal

Input: Dataset D
Output: Final causal network *DAG* $G^f = (E, V)$

procedure
  V = nodes or columns in D
  Int = Intervened nodes in D
  β = edges existing the the knowledge graph
  randomDAG = *createRandomDAG*(V, 100)
  for i from 1 to 100 **do**
    optimisedDAG[i] = *Tabu*(randomDAG[i], Int, β)
  edgeProb = *edgeStrength*(optimisedDAG)
  G^f = *learnCausalDAG*(edgeProb, Threshold)

Algorithm 2 Kg2Causal for Observational and Interventional data

Input: Dataset D_1,D_2...D_k each collected from experiments 1,...,k
Output: Final causal network *DAG* $G^f = (E, V)$

procedure
  for j from 1 to k **do**
    V = nodes or columns in D
    Int = Intervened nodes in D
    β = edges existing the the knowledge graph
    randomDAG = *createRandomDAG*(V, 100)
    for i from 1 to 100 **do**
      optimisedDAG[i] = *Tabu*(randomDAG[i], Int, β)
    edgeProb[j] = *edgeStrength*(optimisedDAG)
  averageEdgeProb = averageNetwork(edgeProb)
  G^f = *learnCausalDAG*(averageEdgeProb, Threshold)

Figure 4: Algorithm 1: Kg2Causal for observational data. We start by creating n random starting DAGs using procedure `createRandomNet` and store them as `randomDAG`. Next, from each DAG in `randomDAG`, we learn an optimized network and store the n networks in a list `optimizedDAG`. In this step, we also pass β to `createRandomDAG`. Next, using the list of networks in `optimizedDAG`, we compute the probabilistic arc strength for each ordered pair of nodes as its empirical frequency, using procedure `edgeStrength` and store them as `edgeProb`. Finally, we use `learnCausalDAG` where we select the edges with weight above a predefined `Threshold`. **Algorithm 2: Kg2Causal for mixed observational-interventional data**.

3.2. Mix of Observational and Interventional experiments

With causal network learning based on a single observational dataset, it is difficult to differentiate between compatible Markov equivalent models [22]. In the simple case of three variables V_i, V_j and V_k, there are three possible causal models $V_i \Longrightarrow V_j \Longrightarrow V_k$, $V_i \Longleftarrow V_j \Longleftarrow V_k$, and $V_i \Longleftarrow V_j \Longrightarrow V_k$; all three structures are Markov equivalent. This ambiguity can be resolved by incorporating measurements from interventional experiments, causing the Markov equivalent structures to have different likelihoods. However, in real-world settings, it is difficult to obtain such interventional measurements as compared to observational measurement [23]. Even when interventional datasets are available, learning a causal network from mixed observational and interventional data is challenging, for two reasons: (i) datasets collected from different experiments under different environmental conditions or batches are not identically distributed, in which case their underlying causal structures may differ leading to errors if network inference is applied to the combined set of measurements; and (ii) in real-world settings interventions are not "perfect" but rather "uncertain" (i.e., "imperfect" or "fat-hand"), meaning that the interventions have other unknown targets, which if ignored would likely yield spurious interactions in network discovery. To deal with such cases, based on our previous study demonstrating the effectiveness of the Learn and Vote algorithm [21, 24], we extended Kg2Causal to include learning from a multi-experiment dataset using a voting-based integration method where experiment-specific causal networks are learned and combined by weighted averaging into a consensus causal network. The additional steps in Algorithm 2 are as follows:

1. Let there be k experiments (can be observational and/or interventional) that produced k datasets with observed variables as (V) and known intervention targets as INT, if any.
2. Repeat steps 1-4 (from Sec. 3) for all k experiments.
3. From the k arc-weight lists, average arc strengths and directions over all the k experiments in which the given arc is not intervened.
4. Per our earlier work [21], we used a threshold of 0.5 for the average arc probability.

4. Analysis and Results

In this section, we describe the observational datasets and ground-truth networks (Sec. 4.1) and the simulated mixed interventional-observational datasets (Sec. 4.2) that we analyzed. We present (Sec. 4.3) the results of empirical studies of network learning performance of Kg2Causal on these datasets in comparison to other types of network structure priors.

4.1. Observational datasets that were analyzed

To assess performance of Kg2Causal on biological network inference problems, we empirically analyzed five real world datasets for which published ground-truth networks were available:

Hepatic encephalopathy: This is a clinical study about a serious liver complication called hepatic encephalopathy (HE) [25] with conditions like electrolyte disorders, infections, poor spirits. It is a categorical dataset with eight nodes and ground-truth containing ten edges.

Sachs et al. T cell signaling: This is a study on mixed observational and interventional experiments to infer causal connections between eleven protein and phospholipids in the intracellular signaling network of individual human CD4+ T-cells [26]. The dataset contains measurement of gene expressions with ground truth network containing twenty edges.

Hematopoietic Stem Cell Differentiation (HSC): This is a real-world gene regulatory network to study underlying myeloid differentiation from multipotent myeloid progenitors to megakaryocytes, erythrocytes, granulocytes and monocytes [27] in mammals [28]. The dataset contains measurement of gene expressions with ground-truth network having thirty edges.

Gonadal Sex Determination (GSD): This a real-world model which represents the gonadal differentiation circuit which monitors the transformation of the bipotential gonadal primordium (BGP) into either female or male gonads [29]. The network consists of eighteen genes and one node for the urogenital ridge. The dataset contains measurement of gene expressions with ground-truth network containing seventy nine edges [28].

Yeast cell cycle: This is a dataset derived from a network model of thirty genes participating in cell-cycle regulation of yeast [30]. The dataset was created by integrating gene expression data with transitive protein-protein interaction. The ground-truth network has 317 edges.

4.2. Mixed observational-interventional datasets

We tested Kg2Causal using Sachs et al. interventional dataset and simulated observational and interventional measurement data from synthetic networks using the `bnlearn` package. For observational data, we drew random samples and for interventional data, we set some target nodes in the network to fixed values in order to create mutilated networks [31] before drawing samples from them. To simulate an uncertain intervention (or "fat-hand") [32] we intervened one or more of child nodes of the intervention's target node.

Cancer: This is a synthetic network [33] on causes and consequences of lung cancer. We simulated data from one observational and one interventional experiment with equal number of samples (500) from each experiment to avoid bias. For interventional experiment we generated a mutilated network: cancer_mut with one intervention (node Smoker).

Asia: This is a synthetic network [34], about occurrence of lung disease and their epidemiological connection a prior visit to Asia. We simulated one observational and two interventional

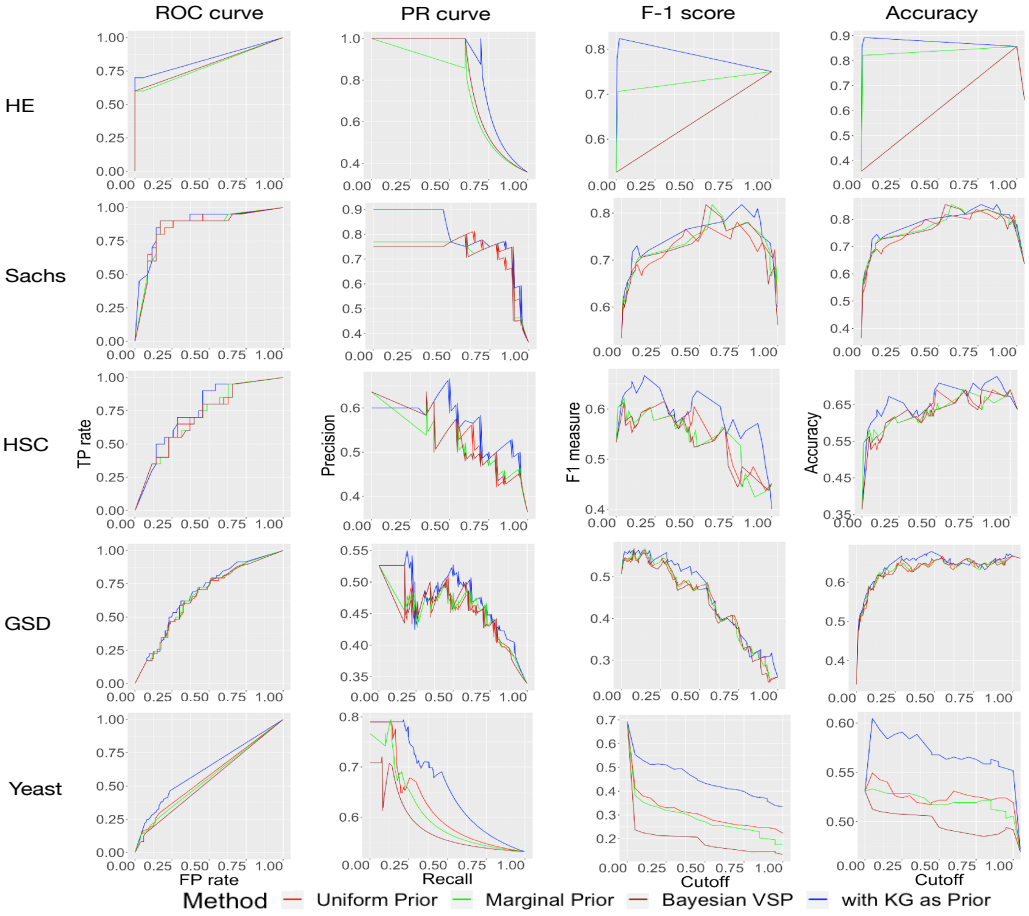

Figure 5: Empirical performance (ROC, precision-vs-recall, F1 vs. cutoff, and accuracy vs. cutoff) of Kg2Causal in each of five datasets compared to learning with uninformative priors (uniform, marginal, and Bayesian VSP).

experiment from the synthetic network with equal number of samples (500) from each experiment to avoid bias. For the interventional experiments we conducted experiments to generate two mutilated networks: asia_mut1 with one intervention (node "Lung Cancer") and asia_mut2 with two intervention (at nodes "Lung Cancer" and "Tuberculosis").

4.3. Analysis of results

In this section we present results of empirical studies of network learning performance on the five observational datasets (see Sec. 4.1) and three mixed observational-interventional datasets (see Sec. 4.2), for Kg2Causal in comparison with three other types of network structure priors. To quantify the performance, we considered presence of an edge in the ground truth network as a "true positive" and absence of an edge as a "true negative" causal arc. For the observational datasets, we used Algorithm 1 with indicated prior (KG, uniform, marginal, or Bayesian VSP) as described in Sec. 3. For mixed interventional-observational datasets, we used Algorithm 2 with the indicated prior. For each dataset, we found (Fig. 5) that using general knowledge graph

Table 1
Performance of "Kg2Causal" versus three uninformative priors, for network learning from observational data: Each row represents a specific real-world dataset (see Sec. 4.1) with corresponding ground-truth network and is split by performance metric (AUROC, AUPR). Rows are ordered by network size (# of nodes).

Dataset	Size	Metric	Uniform	Marginal	Bayesian VSP	Kg2Causal
		AUROC	0.800	0.789	0.800	0.842
HE	8	AUPR	0.810	0.799	0.812	0.854
		AUROC	0.857	0.854	0.849	0.879
Sachs	11	AUPR	0.732	0.729	0.718	0.800
		AUROC	0.705	0.702	0.701	0.745
HSC	11	AUPR	0.547	0.542	0.543	0.564
		AUROC	0.656	0.660	0.656	0.676
GSD	18	AUPR	0.457	0.460	0.458	0.473
		AUROC	0.569	0.556	0.537	0.623
Yeast	30	AUPR	0.619	0.606	0.581	0.662

Table 2
Performance of "Kg2Causal" versus three uninformative priors, for network learning from mixed interventional-observational data: Rows correspond to datasets and columns correspond to types of priors, split into analyses where data are pooled (per Algorithm 1) or voting is used (per Algorithm 2).

Dataset	Size	Metric	Uniform		Marginal		Bayesian VSP		Kg2Causal	
			Pooling	Voting	Pooling	Voting	Pooling	Voting	Pooling	Voting
		AUROC	0.750	0.700	0.750	0.740	0.750	0.780	0.813	0.833
Cancer	5	AUPR	0.776	0.677	0.776	0.690	0.776	0.720	0.809	0.738
		AUROC	0.878	0.944	0.878	0.944	0.887	0.884	0.903	0.956
Asia	8	AUPR	0.711	0.902	0.711	0.905	0.736	0.852	0.817	0.940
		AUROC	0.857	0.873	0.854	0.867	0.849	0.855	0.879	0.883
Sachs	11	AUPR	0.732	0.777	0.729	0.739	0.718	0.728	0.800	0.812

as prior improves performance, by ROC, precision/recall, F1, and accuracy. Quantitatively, Kg2Causal had higher area under ROC curve (AUROC) and area under precision-recall curve (AUPR) scores than network learning with three non-KG priors tested, for the five observational (Table 1) and three mixed interventional-observational (Table 2) datasets. Moreover, the results of comparative analysis of Kg2Causal performance on mixed datasets (Table 2) show effect of pooling data from different experiments (Algorithm 1) as compared to voting (Algorithm 2) for such cases: pooling is better for small network (Cancer) (consistent with our previous findings [21]), whereas voting is better for medium-sized networks (Asia and Sachs).

5. Discussion and Conclusion

A limitation of this study is that due to lack of availability large ground-truth causal networks, all datasets analyzed in this work are for small to medium sized networks (8-30 nodes); due to scalability issue of score-based methods, Kg2Causal method as described here would be challenging to apply to larger networks (many hundreds to thousands of nodes and beyond), which is an area of future work. Further, we plan to explore ways to incorporate a network structure prior in constraint based algorithms (for example, PC algorithm [2]), given (in general) more

favorable scalability of constraint-based algorithms and given the overwhelming preponderance of observational-only datasets that are available. We also want to evaluate alternative methods (other than the method [11] that we are using) for incorporating priors and compare them. Present work clearly demonstrates, for the case of causal network learning from small- to medium-sized biomedical or biological datasets, the importance of aggregating and leveraging structured prior knowledge in order to maximize network learning accuracy.

6. Acknowledgments

This work was supported in part by the National Center for Advancing Translational Sciences (NCATS) through the Biomedical Data Translator program (OT2TR002520 & OT2TR003428 to SAR). We thank David Koslicki, Eric Deutsch, Yao Yao, Zheng Liu, Deqing Qu, Finn Womack, and Ujjval Kumaria for their work on constructing the KG1 knowledge graph.

References

[1] J. Pearl, Causality: models, reasoning, and inference, Econometric Theory 19 (2003) 46.
[2] P. Spirtes, C. Glymour, R. Scheines, Causation, prediction, and search. Adaptive computation and machine learning, MIT Press, Cambridge, MA, 2000.
[3] B. Chakraborty, M. Sinha, Student evaluation model using bayesian network in an intelligent e-learning system, Journal of Institute of Integrative Omics and Applied Biotechnology (IIOAB) 7 (2016).
[4] D. Chatterjee, A. Sinha, M. Sinha, S. K. Saha, A probabilistic approach for detection and analysis of cognitive flow., in: BMA@ UAI, 2016, pp. 44–53.
[5] D. Chatterjee, A. Sinha, M. Sinha, S. K. Saha, Method and system for detection and analysis of cognitive flow, 2020. US Patent 10,722,164.
[6] D. M. Chickering, Learning equivalence classes of Bayesian-network structures, J Mach Learn Res 2 (2002) 445–498.
[7] P. Giudici, R. Castelo, Improving Markov chain Monte Carlo model search for data mining, Machine Learning 50 (2003) 127–158.
[8] N. Friedman, D. Koller, Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks, Machine learning 50 (2003) 95–125.
[9] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, M. West, Experiments in stochastic computation for high-dimensional graphical models, Statistical Science (2005) 388–400.
[10] S. Mukherjee, T. P. Speed, Network inference using informative priors, Proc Nat Acad Sci USA 105 (2008) 14313–14318.
[11] R. Castelo, A. Siebes, Priors on network structures. biasing the search for Bayesian networks, Int J Approx Reason 24 (2000) 39–57.
[12] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition and applications, arXiv preprint arXiv:2002.00388 (2020).
[13] L. Ehrlinger, W. Wöß, Towards a definition of knowledge graphs., SEMANTiCS (Posters, Demos, SuCCESS) 48 (2016) 1–4.

[14] F. Glover, Future paths for integer programming and links to artificial intelligence, Computers & operations research 13 (1986) 533–549.
[15] D. Heckerman, D. Geiger, D. M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) 197–243.
[16] G. F. Cooper, C. Yoo, Causal discovery from a mixture of experimental and observational data, in: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 116–125.
[17] M. Scutari, Learning Bayesian networks with bnlearn, arXiv:0908.3817 (2009).
[18] M. J. Druzdzel, The role of assumptions in causal discovery (2009).
[19] A. Hauser, P. Bühlmann, Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs, J Mach Learn Res 13 (2012) 2409–2464.
[20] M. Scutari, On the prior and posterior distributions used in graphical modelling, Bayesian Analysis 8 (2013) 505–532.
[21] M. Sinha, P. Tadepalli, S. A. Ramsey, Voting-based integration algorithm improves causal network learning from interventional and observational data: an application to cell signaling network inference, Plos one 16 (2021) e0245776.
[22] D. Koller, N. Friedman, F. Bach, Probabilistic graphical models: principles and techniques, MIT press, 2009.
[23] Y. Hagmayer, S. A. Sloman, D. A. Lagnado, M. R. Waldmann, Causal reasoning through intervention, Causal learning: Psychology, philosophy, and computation (2007) 86–100.
[24] M. Sinha, Causal structure learning from experiments and observations (2019).
[25] Z. Zhang, J. Zhang, Z. Wei, H. Ren, W. Song, et al., Application of tabu search-based Bayesian networks in exploring related factors of liver cirrhosis complicated with hepatic encephalopathy and disease identification, Scientific Reports 9 (2019) 1–8.
[26] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, G. P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, Science 308 (2005) 523–529.
[27] J. Krumsiek, C. Marr, T. Schroeder, F. J. Theis, Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network, PLOS ONE 6 (2011).
[28] A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, T. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Nature Methods 17 (2020) 147–154.
[29] O. Ríos, S. Frias, A. Rodríguez, S. Kofman, Merchant, et al., A boolean network model of human gonadal sex determination, Theor Biol Medic Model 12 (2015) 26.
[30] W. Liu, J. C. Rajapakse, Fusing gene expressions and transitive protein-protein interactions for inference of gene regulatory networks, BMC Systems Biology 13 (2019) 37.
[31] J. Pearl, Graphical models for probabilistic and causal reasoning, in: Quantified representation of uncertainty and imprecision, Springer, 1998, pp. 367–389.
[32] D. Eaton, K. Murphy, Exact Bayesian structure learning from uncertain interventions, in: Artificial Intelligence and Statistics, 2007, pp. 107–114.
[33] K. B. Korb, A. E. Nicholson, Bayesian artificial intelligence, CRC Press, 2010.
[34] S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, J Roy Stat Soc B (1988) 157–224.