Multi-Source Causal Analysis: Learning Bayesian Networks from Multiple Datasets

Ioannis Tsamardinos and Asimakis P. Mariglis

Abstract We argue that causality is a useful, if not a necessary concept to allow the integrative analysis of multiple data sources. Specifically, we show that it enables learning causal relations from (a) data obtained over different experimental conditions, (b) data over different variable sets, and (c) data 'over semantically similar variables that nevertheless cannot be pulled together for various technical reasons. The latter case particularly, often occurs in the setting of analyzing multiple gene-expression datasets. For cases (a) and (b) above there already exist preliminary algorithms that address them, albeit with some limitations, while for case (c) we develop and evaluate a new method. Preliminary empirical results provide evidence of increased learning performance of causal relations when multiple sources are combined using our method versus learning from each individual dataset. In the context of the above discussion we introduce the problem of Multi-Source Causal Analysis (MSCA), defined as the problem of inferring and inducing causal knowledge from multiple sources of data and knowledge. The grand vision of MSCA is to enable the automated or semi-automated, large-scale integration of available data to construct causal models involving a significant part of human concepts.

1 Introduction

Unlike humans that continuously and synthetically learn from their observations modern data-analysis fields, for the greatest part, approach learning as single, isolated, and independent tasks. The data analyzed form a relatively

Ioannis Tsamardinos
CSD, University of Crete and BMI, ICS, FORTH, e-mail: tsamard at ics and forth and gr

Asimakis P. Mariglis
BMI, ICS, FORTH, and Physics Dept, University of Crete

homogenous group of observations in terms of observed quantities (i.e., variables), sampling methodology, experimental conditions, and typically, source: that is, they form a single dataset. The computation and inferences of the analysis of one dataset are rarely used in the analysis of other datasets. Consider the following scenario:

- **Dataset 1**: An experimenter is observing variables $\{A, B, C, D\}$ in an independently and identically distributed (i.i.d.) sample of a population with the intent to learn a predictive or diagnostic model for D based on the remaining variables. For example, the predictors $\{A, B, C\}$ could be medical quantities and D the presence or absence of a specific disease in the general population.
- **Dataset 2**: A randomized clinical trial is performed measuring $\{A, B, C, D\}$ where variable B is randomly set to different values (e.g. a medication is prescribed that directly controls the value of B) and the effect on disease D is observed. These data cannot be merged with Dataset 1 because the joint distributions of the data are different. For example, if B is caused by the disease (e.g., the disease increases the concentration of a protein in the blood) then B will be highly associated with D in Dataset 1; in Dataset 2 where the levels of B exclusively depend on the medication administered, B and D are not associated.
- **Dataset 3**: Variables $\{D, E, F\}$ for prediction of disease F. These data cannot be pulled together with Dataset 1 or 2 because they measure different variables.
- **Dataset 4**: Variables $\{A', B', C, D\}$ are observed in an i.i.d. sampling, where A' , B' are semantically similar but not identical to A, B respectively. These data may not be pulled together with Dataset 1 for a number of reasons. It could be the case for example, that A and B are continuous, while A' , B' are recorded as discrete (e.g., low, medium, high); or they measure the same quantity using different scales and methods with no apparent mapping from one to the other. This is common in psychology where quantities such as psychological improvement or degree of depression of a patient can be measured using several methods, sometimes not fully objective. Particularly, this is a common situation in gene-expression measurements where the gene-expression level A of a specific gene in one study is not directly comparable to the gene-expression level A' of the same gene in a different study due to multiple technology barriers for translating A to A' (see [4] for a more detailed explanation of these limitations).

State-of-the-art machine learning and statistical methods will typically be applied to identify a predictive or diagnostic model from the above datasets. No matter what kind of analysis is performed however, **each dataset is typically analyzed in isolation**. A growing number of datasets such as the above is made public, each developed with a specific hypothesis in mind or to build a specific predictive model. Modern machine-learning methods

have yet to fully address, or even focus, on the problem of synthesizing such information.

The current practice is instead for humans to serve as the means of integrating the extracted knowledge. Researchers read the scientific literature and form inside their heads a (causal, arguably) model of the working mechanisms of the entity they study. A conscious effort of manual knowledge synthesis in biology is for example the KEGG PATHWAY database at http://www.genome.jp/kegg/pathway.html defined as "... a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks ". Obviously, the manual synthesis of information is severely limited by our mental capacities.

At a first glance, it may seem impossible that the above datasets can be analyzed simultaneously. However, modern theories of causality are gradually making this possible. We argue that the concept of causality is fundamental in achieving an automated, or semi-automated, combined analysis of different data-sources as in the above scenario. This is because causality (a) can model the effects of actions, such as setting different experimental conditions and sampling methodology and (b) it can be inferred by tests of conditional independence that can be performed on different datasets. We introduce the general problem and conceptual framework of **Multi-Source Causal Analysis** (MSCA) defined as *the problem of inferring and inducing causal knowledge from multiple sources of data and knowledge*. We consider as the main components of MSCA (i) a formal representation and modeling of causal knowledge, (ii) algorithms for inducing causality from multiple sources and for making causal inferences, and (iii) the ability to justify and explain the causal inferences to human experts. The vision of MSCA is to enable data and knowledge synthesis on a grand-scale were thousands of studies are simultaneously analyzed to produce causal models involving large parts of human concepts. We now present algorithms that allow the simultaneous inference of causal knowledge from the above datasets.

2 Causal Bayesian Networks as a Language for Causality

We assume the readers' familiarity with the standard Pearl [5] and Spirtes et al. [6] causality framework based on the concept of Causal Bayesian Network and only briefly review it. We consider the standard notions of probabilistic causality for "X is causing Y" and for defining direct causality. Let us consider a set of random variables $\mathbf{V}$. We represent the causal structure among the variables in $\mathbf{V}$ with a directed acyclic graph (DAG) G with the vertexes corresponding to the variables $\mathbf{V}$; an edge $X \rightarrow Y$ exists in G if and only if X directly causes Y relatively to $\mathbf{V}$. We define a Causal Bayesian Network (CBN) as the tuple $\langle G, P \rangle$, where G is a causal structure over $\mathbf{V}$ and P is the joint probability distribution of variables $\mathbf{V}$. We assume that for a CNB

$\langle G, P \rangle$ the Causal Markov Condition ($\mathcal{CMC}$) holds: every variable X is probabilistically independent of any subsets of its non-effects (direct or indirect) given its direct causes. A causal graph G is depicted in Figure 1(a).

We denote the independence of X with Y given $\mathbf{Z}$ as $I(X;Y|\mathbf{Z})$. We also denote the d-separation of two nodes X and Y by a subset $\mathbf{Z}$ as $Dsep(X;Y|\mathbf{Z})$ (see [5] for a formal definition). The d-separation criterion is a graphical criterion that determines all the independencies in the distribution P that are entailed by the graph and the $\mathcal{CMC}$: $Dsep(X;Y|\mathbf{Z}) \Rightarrow I(X;Y|\mathbf{Z})$. If $\neg Dsep(X;Y|\mathbf{Z})$ we say that X is d-connected to Y given $\mathbf{Z}$. For a broad class of distributions, called Faithfull distributions, the converse also holds, i.e., $I(X;Y|\mathbf{Z}) \Leftrightarrow Dsep(X;Y|\mathbf{Z})$ named as the Faithfulness Condition ($\mathcal{FC}$). The name faithful stems from the fact that the graph faithfully represents all and only the independencies of the distribution; another equivalent way of expressing faithfulness is that the independencies are a function only of the causal structure and not accidental properties derived by a fine tuning of the distribution parameters. In Pearl's terminology, faithful distributions and corresponding CBNs are called stable: under small perturbations of the distribution, the set of independencies remains the same.

A large class of causal discovery algorithms performs statistical tests in the data to determine whether $I(X;Y|\mathbf{Z})$; subsequently, since in faithful distributions this is equivalent to $Dsep(X;Y|\mathbf{Z})$, the result of the test imposes a constraint on the data-generating graph. By combining and propagating these constraints these algorithms, named *constraint-based*, can determine the causal graphs that exactly encode (are consistent with) the independencies observed in the data distribution.

In practice, there are typically many latent variables. We can think of the variables $\mathbf{V}$ partitioned into observed variables $\mathbf{O}$ and hidden variables $\mathbf{H}$: $\mathbf{V} = \mathbf{O} \cup \mathbf{H}, \mathbf{O} \cap \mathbf{H}$. The data are sampled from the marginal $P_{\mathbf{O}}$ of the observed variables only, i.e., we can only test independencies involving variables in $\mathbf{O}$. A prototypical, constraint-based causal discovery algorithm is the FCI [6]. The output of FCI is what is called a Partial Ancestral Graph (PAG) containing common features of all causal graphs (including ones with hidden variables) that could faithfully capture the marginal data distribution $P_{\mathbf{O}}$. A PAG is shown in Figure 1(b). The edges have the following semantics [1]:

- $A \rightarrow B$ means that A is a direct cause[2] of B relatively to $\mathbf{O}$.
- $A \leftrightarrow B$ means that neither A directly causes B relatively to $\mathbf{O}$ nor vice-versa, but A and B have a common latent cause.

[1] Due to space limitations and for clarity of presentation, we do not discuss here the possibility of selection bias in sampling the data that can be addressed with the FCI algorithm.

[2] This is a simplification for purposes of removing some technical details from the presentation, not necessary for conveying the main ideas. The exact semantics of an edge is that there is an inducing path from A into B relative to $\mathbf{O}$, where the concept of inducing path is defined in [6].

- $A \diamond - \diamond B$ with the $\diamond$ denoting the fact that there is at least one causal graph consistent with the data where $\diamond$ is replaced by an arrowhead and at least one graph where there is no arrowhead (e.g., an edge $A \diamond - \diamond B$ means that $A \rightarrow B$, $A \leftarrow B$, and $A \leftrightarrow B$ are all possible).

3 Learning from Data Obtained over Different Experimental Conditions

We now argue that causality could be used to make inductions about the data-generating process of samples obtained under different experimental conditions. *This is because a causal model directly encodes the effects of manipulations of the system.* Assume for example that G represents the causal structure of the system without any intervention. Now assume that in i.i.d. datasets $\{D_i\}$ a set of variables $\mathbf{M}_i$ is being manipulated, i.e., obtains values set by an external agent performing an experiment. Then, the causal graph $G_{\mathbf{M}_i}$ of the system under manipulations $\mathbf{M}_i$ is derived from G by removing all incoming edges into any $V_j \in \mathbf{M}_i$. The intuitive explanation is that the value of V_i now only depends on the external agent and has no other causal influence [5, 6]. Assuming an unmanipulated graph G and known performed manipulations $\{\mathbf{M}_i\}$, graphs $\{G_i\}$ can be constructed; the fitness of each one to the corresponding data D_i can be estimated. This in turn allows us to estimate the overall fitness of the assumed model G to the set of datasets $\{D_i\}$. For example, the algorithm in [3] greedily searches the space of CBNs to find the best-fitting graph G to the set of datasets. The key-point is that unlike causality-based formalisms, correlation-based ones do not allow us to predict the effect of manipulations $\mathbf{M}_i$ in order to fit standard predictive or diagnostic models or even to perform feature selection. The algorithm in [3] could jointly analyze Datasets 1 and 2 of the scenario in the introduction.

4 Learning from Data over Different Variable Sets

Let us consider Dataset 1 of the example scenario measuring variables $\mathbf{V}_1 = \{A, B, C, D\}$ and Dataset 3 measuring $\mathbf{V}_1 = \{D, E, F\}$. Standard statistics and machine learning methods would analyze the datasets in isolation (e.g., build predictive models). Why cannot these methods make any additional inferences? Current state-of-the-art Machine Learning is arguably for the most part correlation-based and aims at identifying the set and type of such correlations. Given this observation, it may come as no surprise the difficulty found in combining knowledge and models; for one thing, correlation transitivity does not hold. If A is correlated with B and B is correlated with C, nothing can be said about the correlation between A and C. There

are no further inferences about the joint $P(A, B, C)$ other than the observed correlations. Is there anything more to infer from the above datasets? Unlike pairwise correlations, pairwise causal relations are transitive: if A is causing B and B is causing C, then A is causing C. These and other more complicated inferences allow us in some cases to induce more causal knowledge from the combined data, than from each dataset individually.

We will present an example of such inferences on the structure of the union set of variables $\mathbf{V} = \mathbf{V}_1 \cup \mathbf{V}_2$. The variables of each dataset are by definition latent when we infer structure from the other dataset. Thus, we will use the FCI algorithm. Let us assume that the true (unknown) causal structure is the one shown in Figure 1. From dataset on $\mathbf{V}_1$ we expect to observe the independencies $I(A; B|\emptyset)$, $I(A; D|C)$, $I(A; D|C, B)$, $I(B; D|C)$, $I(B; D|C, A)$ and only (because the graph is assumed faithful) and from the dataset $\mathbf{V}_2$ we will observe $I(D; F|\emptyset)$ and only. Running FCI on each dataset independently (assuming enough sample so that our statistical decisions about conditional independence are correct) will identify these independencies and obtain the two PAGs shown in Figure 1(b) and (c) named G_1 and G_2. The edges with a square in one of their ends-points denote that an arrowhead could or not substitute the end-point. For example, in Figure 1(b) the edge $A \circ\!\!\rightarrow C$ denotes the fact that FCI cannot determine whether the true edge is $A \rightarrow C$ (i.e., A directly causes C) or $A \leftrightarrow C$ (i.e., there is no direct causation between A and C but the observed dependencies and independencies are explained by the existence of at least one common hidden ancestor H, a.k.a. confounder $A \leftarrow H \rightarrow C$).

The models are informally combined as shown in Figure 1(d). An algorithm that formalizes and automates the procedure combining the PAGs stemming from different datasets is presented in [7] independently discovered at the same time our group was designing a version of such an algorithm. Due to space limitations, we only present some key inferences to illustrate our argument. There is nothing we can infer about the causation between E and $\{A, B, C\}$. We have no evidence for or against the existence of such causal relations so the corresponding edges are shown in dashed in the figure. In addition, there are CBNs that are compatible with any possible direction of such edges. Similarly, there is no evidence for or against edges between F and $\{A, B\}$. However, one can rule out the case that $A \rightarrow F$ because then at least one of the paths $D \leftarrow C \leftarrow A \rightarrow F$ or $D \leftarrow C \leftarrow H \rightarrow A \rightarrow F$ (if there is a latent variable H between C and A) would exist in the data-generating DAG. That is, there would be a d-connecting path from F to D given the emptyset, which contradicts the observed independence $I(D, F|\emptyset)$. With a similar reasoning, the possibility $B \rightarrow F$ is ruled out. *Even more impressive is the inference that there is no edge between F and C, a pair of variables that we have never measured together in the available data!* No matter what kind of edge we insert in the graph, it would create a d-connecting path between F and D. For example, if we insert an edge $F \leftrightarrow C$, it would imply the path $F \leftarrow H \rightarrow C \rightarrow D$, which contradicts the

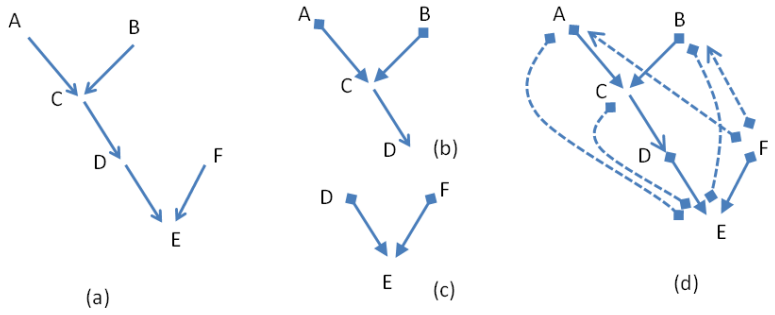

Fig. 1 (a) Presumed true, unknown, causal structure among variables $\{A, B, C, D, E, F\}$. (b)-(c) The causal structure identified by FCI when run on any large-enough dataset over $\{A, B, C, D\}$ and $\{D, E, F\}$ respectively. (d) Informally combining all the causal knowledge together and inferring new knowledge. Edges in dash are possible but have no direct evidence. Particularly notice, F cannot be a cause of A or B and there is no edge between C and F even though they are never measured together; this is because such edges would lead to a contradiction with the observed independencies in the data of (b) and (c).

observed independence $I(F, D|\emptyset)$ in D_2. In general, we can rule out edges or edge directions that create possible d-connecting paths that contradict the observed independencies in any dataset that we have available. *In essence, the $\mathcal{CMC}$ and the $\mathcal{FC}$ entail new constraints on the distribution of the union of the variables in all datasets.*

At this point we cannot help but wonder whether a similar modeling goes on in the brain. How do we know that the bus delays have no correlation with the corn price? Most of us have never measured them together (i.e., measured their correlation) and certainly do not check the corn price before deciding when to run to the bus station. We assume independence since we consider as improbable the cases of bus delays causing changes in corn price, or vice versa, as well as the existence of any common cause. There are probably an astronomical number of such independencies implicitly stored in our brains. We argue that they may be explained by causal inferences and presumptions about the causal structure of the natural world; they are not based on reasoning about correlations alone.

5 Learning from Data Obtained over Semantically Similar Variables

It is often the case that a common set of variables (i.e., variables that semantically correspond to the same quantity) is observed in different datasets, but for technical reasons the data cannot be pulled together in one dataset. For example, the variables may be measured by different equipment and so it may

be hard to translate the values from all datasets to a common scale. Such a situation is typical in gene-expression studies: for various technical reasons measurements corresponding to the gene-expression of a specific gene are not directly comparable among different studies [4]. In psychology and social sciences different and incomparable methods may be used to measure a quantity, such as social-economical status, degree of depression, or mental capacity of patients. When constructing a predictive model it seems difficult to combine the data together, without first finding a way to translate the values to a common scale. This rules out most machine-learning methods. However, certain inferences in constraint-based causal discovery (as in the previous sections) are possible using only tests of conditional independence.

We now develop a multi-source test of conditional independence that employs all available data without the need to translate them first. We denote with $T(X;Y|\mathbf{Z})$ the test of conditional independence of X with Y given $\mathbf{Z}$. $T(X;Y|\mathbf{Z})$ returns a p-value assuming the null hypothesis of the independence. Constraint-based algorithms then use a threshold t rejecting the independence if $T(X;Y|\mathbf{Z}) < t$ (i.e., they accept $\neg I(X,Y|\mathbf{Z})$) and accepting the independence $I(X,Y|\mathbf{Z})$ otherwise. Since, we cannot pull all the data together, we perform the test of independence $T(X;Y|\mathbf{Z})$ individually in each available dataset D_i obtaining the p-values $\{p_i\}$. Fisher's Inverse χ^2 test can then be used to compute a combined statistic $S = -2\sum \log p_i$. S follows a χ^2 distribution with $2n$ degrees of freedom, where n is the number of datasets contributing data to the test, from which we can obtain the combined p-value p^* for the test $T(X;Y|\mathbf{Z})$ employing all available data. Other methods to combine p-values exist too [4].

An important detail of the implementation of the test is the following. Constraint-based methods do not perform a test $T(X;Y|\mathbf{Z})$ if there is not enough statistical power. The statistical power is heuristically assumed adequate when there are at least k available samples per parameter to be estimated in the test (typically k equals 5 or 10 [8] in single-source analysis; in our experiments it was set to 15). When combining multiple datasets, each dataset may not have enough sample to perform the test, but their combination could have. So, we implemented a new rule: the test $T(X;Y|\mathbf{Z})$ is performed when the total average samples per parameter exceeds k, i.e., $\sum n_i/m_i \geq k$, where n_i the samples of dataset i and m_i the parameters estimated by the specific test in dataset i (these maybe different if for example a variable takes 3 possible values in one dataset and 4 in another). When all datasets have the same number of parameters for the test, the rule results in testing whether $n/m \geq k$ as in the single-dataset case. *The new multi-source test $T(X;Y|\mathbf{Z})$ can be used in any constraint-based algorithm for combining data from different sources provided: X, Y and $\mathbf{Z}$ are measured simultaneously and the data are sampled under the same causal structure, experimental conditions, and sampling conditions (e.g., case-control data cannot be combined with i.i.d. data using this test).*

As a proof-of-concept we have performed experiments using the ALARM [1] network, the above multi-source test of independence, and our implementation of the PC algorithm [6] (similar to FCI but in addition assuming no latent variables so there are no bidirectional edges output). First, we sample data from the network distribution for n different sources (datasets) and measure the difference of the reconstructed network with the data-generating one using the Structural Hamming Distance (SHD) measure, which corresponds roughly to the sum of missing and extra edges, and wrong orientations [8]. Figure 2 on the left shows the SHD vs. the number n of combined datasets, *assuming each dataset has a fixed number of training samples equal to 500.* Figure 2 on the right shows the SHD vs. the number n of equal-sized datasets, *assuming a fixed total available sample of 5000 cases.* Each point in the graph is the average of 5 different runs of the same experiment. There are two lines in each graph. The blue (bottom) one corresponds to datasets having no difference in the measurement of their variables, i.e., their values could be pulled together in a single dataset. This line serves as a baseline and to test the validity of the method and the implementation. The green (top) line in each graph stems from combining datasets that simulate the effect of measuring the same quantity with different methods or scales. For each dataset, a non-binary variable with probability 10% was binarized by grouping a set of consecutive values together (e.g., values Low and Medium would be grouped to a single value).

In general, the number of errors measured by the SHD is decreasing as the total available sample size is increasing (Figure 2 left) until it flattens out and reaches the limit of the method. In addition, Figure 2 right shows that the number of errors is increasing for a fixed sample size that becomes increasingly fragmented into different datasets. Notice that, the lines corresponding to datasets where the variables are measured in different scales are always above (exhibit more errors) than when the datasets are homogenous. Curiously, the green line in Figure 2 left seems to be increasing again with the number of datasets, which requires further investigation. This artifact seems persistent in other experiments we run and it seems unlikely it is due to statistical fluctuations.

We would also like to note an important detail of the method. The p-values produced by the individual tests $T(X;Y|\mathbf{Z})$ on a single dataset are typically based on the Pearson's χ^2 or the likelihood ratio test that are only asymptotically correct. When the sample size is low the approximations to the p-values are rough and the statistic $S = -2\sum \log p_i$ and corresponding compound p-value severely skewed. Thus, the method as implemented should not be used with small sample sizes. We are currently investigating Bayesian methods, exact independence tests and other techniques based on permutations and ranking [4] to overcome this problem.

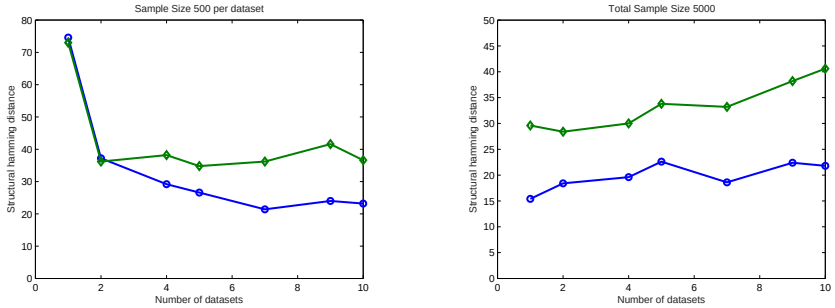

Fig. 2 Left: the Structural Hamming Distance (SHD) versus the number of datasets, when each dataset has 500 sample cases. Right: the SHD versus the number of datasets, assuming a fixed total sample size of 5000. The line with the circles in each graph corresponds to datasets having no difference in the measurement of their variables. The line with the diamonds in each graph stems from combining datasets where some non-binary variables have been binarized by grouping together consecutive values.

6 Discussion and Conclusions

We show that the concept of causality, causal theories, and recently developed algorithms allows one to combine data-sources under different experimental conditions, different variables sets, or semantically-similar variables to infer new knowledge about the causal structure of the domain. Omitted due to space limitations, is a similar discussion about data obtained under different sampling methods (e.g., case-control vs. i.i.d. data) and selection bias (see [6] for a discussion). If one examines closely the basic assumptions of the algorithms mentioned, causal induction as presented is based on the following assumptions: the Causal Markov Condition, the Faithfulness Condition and acyclicity of causal relations. Even though debatable, these assumptions are broad, reasonable, and non-parametric (see [5, 6] for a discussion). In addition, they could be substituted for other sets of assumptions depending on the domain, e.g., if one is willing to accept linearity and multivariate normality, Structural Equation Models can be employed for modeling causality that deal with cyclic (a.k.a. non-recursive) networks.

Other work in data analysis for merging datasets exists. This includes the field of meta-analysis [4] and multi-task learning [2]. The former mainly focuses on identifying correlations and effect sizes. It does not model causation so it can only deal with datasets sampled using the same method over the same experimental conditions and variables. Multi-task learning is limited to building predictive models for different tasks with a shared representation and input space (predictor variables) in order to extract useful common features [2]. It cannot deal with the range of different data sources for which causality provides the potential.

Our contribution in this paper is to draw attention to the various, recent, existing causal algorithms and techniques for merging datasets, to develop a new method for the case of learning networks from datasets over semantically similar variables, and to note some of the current limitations. For example, using the methods presented, Dataset 1 could be analyzed in conjunction with Dataset 2 or Datasets 3 and 4; however, Dataset 2 cannot be analyzed together with the rest because the algorithm in [3] is not constraint-based. In addition, the latter algorithm assumes there are no common hidden confounders (a.k.a. Causal Sufficiency). We are currently working on constraint-based methods for analyzing datasets obtained under different experimental conditions that will allow all the datasets in the scenario to be analyzed together. We also note that the algorithm for merging datasets over different variable sets [7] is impractical because of its time and memory requirements; further improvements are required before it becomes useful to the average researcher. We have named the effort of inducing causal knowledge from multiple data and knowledge sources as Multi-Source Causal Analysis (MSCA), with the vision of enabling the automation of the large-scale, co-analysis of available datasets; essentially MSCA aims in formalizing and automating the scientific process to some degree, since the result of the latter is typically causal knowledge.

References

1. I. Beinlich, G. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings 2nd European Conference in Artificial Intelligence in Medicine*, pages 247–256, 1989.
2. Richard A. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings 10th International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann, 1993.
3. Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *UAI*, pages 116–125. Morgan Kaufmann, 1999.
4. Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24:374–382, 2008.
5. J. Pearl. *Causality, Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K., 2000.
6. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
7. R. E. Tillman, D. Danks, and C. Glymour. Integrating locally learned causal structures with overlapping variables. In *NIPS*, 2008.
8. I. Tsamardinos, L.E. Brown, and C.F. Aliferis. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Machine Learning*, 65(1):31–78, 2006.