Scalable and Hybrid Ensemble-Based Causality Discovery

Pei Guo*, Achuna Ofonedu†, Jianwu Wang*
*Department of Information Systems, University of Maryland, Baltimore County, Baltimore, United States
{peiguo1, jianwu}@umbc.edu
†Department of Electrical Engineering and Computer Science, Catholic University of America, Washington, D.C., United States
ofonedub@cua.edu

***Abstract*—Causality discovery mines cause-effect relationships among different variables of a system and has been widely used in many disciplines including climatology and neuroscience. To discover causal relationships, many data-driven causality discovery methods, e.g., Granger causality, PCMCI and Dynamic Bayesian Network, have been proposed. Many of these causality discovery approaches mine time series data and generate a directed causality graph where each graph edge denotes a cause-effect relationship between the two connected graph nodes. Our benchmarking of different causality discovery approaches with real-world climate data show these approaches often generate quite different causality results with the same input dataset due to their internal learning mechanism differences. Meanwhile, there are ever-increasing available data in virtually every discipline, which makes it more and more difficult to use existing causality discovery algorithms to produce causality results within reasonable time. To address these two challenges, this paper utilizes data partitioning and ensemble techniques, and proposes a two-phase hybrid causality ensemble framework. The framework first conducts phase 1 data ensemble for partitioned data and then conducts phase 2 algorithm ensemble from data ensemble results. To achieve scalability, we further parallelize the ensemble approaches via the Spark big data analytics engine. Our experiments show that our proposed approaches achieve good accuracy through ensemble and high scalability through data-parallelization in distributed computing environments.**

***Index Terms*—Causality discovery, Ensemble learning, Data parallelism, Granger causality, Dynamic Bayesian Network**

I. Introduction

Causality [21] is a fundamental research topic studying cause-effect relationships among different components of a system and causality study can help explain why the system has certain behaviors. Causality learning/discovery has been widely studied and applied in many disciplines including climatology and neuroscience.

Many data-driven causality learning approaches have been proposed, such as Granger causality [12], PCMCI [24], Dynamic Bayesian Network [19], and Convergent Cross Mapping [32]. These approaches often mine time series data of two or more variables in a system and produce their predictions on cause-effect relationship among these variables. For instance, the work at [26] uses Granger causality to study cause-effect relationships among multiple climate variables and shows that sea surface temperature changes at pacific ocean near equator, an indicator of the El Niño-Southern Oscillation (ENSO) climate phenomenon [13] can cause abnormal surface temperature, pressure and precipitation remotely.

One challenge with the variety of different causal discovery approaches/algorithms is that these approaches often lead to divergent causality conclusions from the same dataset, which makes it difficult to explain and use data-driven causality discovery results. There have been some studies comparing different causality discovery methods [15], [33]. For example, the experiments on comparing three causality discovery algorithms show there are only 83% overlapping among the results on average [15]. Yet there is still a lack of comprehensive framework to effectively integrate these diverse algorithms.

The other challenge to be tackled by this paper is the ever-increasing volume and dimension of available data for causality discovery. For instance, total worldwide climate data volume is projected to increase from 5 PB in 2010 to 350 PB in 2030 [20]. It is more and more difficult to use existing causality discovery algorithms to handle the increasing dimensionality and resolution of these climate datasets. Meanwhile, data volume is just one factor for time complexity of many causality discovery algorithms. As an example, a popular Granger causality algorithm's execution time grows quadratically with the increase of either of the three factors: data record number, variable number and time lag number [4]. Parallel causality discovery is crucial as a solution to reduce computation time.

To address the above two challenges, this paper applies data partitioning and ensemble techniques to achieve scalable and accurate causality learning. Ensemble learning [23] is a meta machine learning algorithm which combines multiple base or individual learners in order to get better overall learning accuracy. In this paper, we propose a two-phase hybrid causality ensemble learning framework by first partitioning data into smaller sizes and conducting phase 1 data ensemble for each data partition and then conducting phase 2 algorithm ensemble from phase 1 ensemble results. The framework can be easily parallelized through big data engines like Spark [1] and is adaptable to different ensemble approaches. To the best of our knowledge, this study is the first supporting both scalable and ensemble learning for causality discovery. The implementations of our work is open-sourced at [2].

The contributions of this paper are as follows.

- We propose a two-phase hybrid causality ensemble framework by first conducting phase 1 data ensemble for partitioned data and then conducting phase 2 algorithm ensemble from phase 1 data ensemble results. The framework can combine learning results from different data partitions (namely data ensemble), and different algorithms (namely algorithm ensemble).
- Based on the above framework, we propose an approach for parallel causality ensemble learning via Spark [1] and the MapReduce programming model [10].
- We did experiments to evaluate our proposed scalable ensemble framework and approach, which shows that our approach can achieve both perfect accuracy and almost linear speedup.

The rest of the paper is organized as follows. The background is introduced in Section II. The two-phase hybrid causality ensemble learning framework is explained in Section III. Section IV contains the ensemble approach based on the scalable causality ensemble framework. Section V describes the parallelization of our implementation. The experiments and evaluations are in Section VI, with related work discussion in Section VII. Finally, Section VIII concludes our paper.

II. Background

A. Ensemble Learning

Ensemble learning [23] is a meta machine learning algorithm which uses multiple learning methods to obtain better predictive performance than learning from any of the constituent methods. Since 1990, ensemble learning methods have become a major learning paradigm because of both empirical good performances in real-world applications and theoretical proof on its advantages [25]. Many state-of-art data mining approaches/packages, e.g. random forest and XGBoost [8], are based on ensemble learning.

Many ensemble learning algorithms have been proposed and they mainly vary in the following three aspects: 1) what are base/individual learners, 2) how each base learner learns from input data, 3) how to combine results of base learners. For base learner selection, if base learners used in an ensemble learning belong to the same type, e.g. decision tree or neural network, the ensemble algorithm is called homogeneous ensemble. Otherwise, it is called heterogeneous ensemble. Regarding how each base learner learns, there are three main approaches and they mostly differ in how input data is fed to base learner. The first approach, called stacking ensemble [31], uses the same input data for all base learners. Bagging ensemble [6], as the second approach, uses different sampling results from the original input data for different base learners. The third approach is boosting ensemble [11] which uses multiple base learners iteratively and, in each iteration, assigns higher weight to data whose learning accuracy was low in previous iterations. On base learner combination, common methods are majority voting and weighted majority voting [23].

B. Causality Discovery Methods

Existing causal relationships discovery methods can be categorized into two types depending on the input datasets types: 1) learning from multivariate independent and identically distributed (i.i.d.) data and 2) learning from multivariate time-series data. The learning results from a multivariate causality approach can be denoted as a directed graph where each graph edge represents a cause-effect relationship conditioned on all other variables in the graph. In this subsection, we explain three multivariate causality discovery approaches towards time-series input data, namely multivariate (graphical) Granger causality [4], PCMCI [24] and dynamic Bayesian network [19] and their algorithm details. Because they all belong to the same casualty discovery category and their learning results can be modeled as directed graphs, we could conduct ensemble learning using these algorithms as base learners which will be explained in later sections.

1) Multivariate Granger Causality: Granger causality, as a predictive model in economics, was proposed in 1969 by Nobel Laureate Clive W. Granger. By definition, in Granger causality, one time series x Granger causes another time series y, if and only if the regression for y based on past values of both x and y is statistically significant than the regression of y only based on past values of y itself. To demonstrate the definition, let the lagged variable x be x_{t-i} for i from 1 to maximum lag P; and similarly, the lagged y is represented by y_{t-i}. To test Granger causality, in first step, the following two linear regressions functions are fitted as follows:

$$y_t = a_{11} \cdot y_{t-1} + a_{12} \cdot y_{t-2} + ... + a_{1P} \cdot y_{t-P} + \varepsilon_1 \quad (1)$$

$$y_t = a_{21} \cdot y_{t-1} + ... + a_{2P} \cdot y_{t-P} + b_{21} \cdot x_{t-1} + ... + b_{2P} \cdot x_{t-P} + \varepsilon_2 \quad (2)$$

Next, the accuracy of predicting y_t using Equation (1) and Equation (2) are compared to check which regression works better. In most cases, statistical hypothesis test methods such as F-test or Chi-squared (χ^2) test are utilized to get a p-value to determine statistical significance.

The above pairwise Granger causality is proven to work well on discovery between each pair of variables. However, most datasets in research contain more than two variables. When the scientists intend to discover the causality among a subset or the whole set of a multivariate dataset, pairwise Granger causality ignores the causalities with other untested variables, which could generate spurious causal relationships such as confounding variable [22] and indirect causal relationship [18].

To address the limitations of the pairwise Granger causality method, multivariate Granger causality discovery, a.k.a. graphical Granger causality discovery, fits a vector autoregressive model (VAR) to time series data [16], compared to linear regression models in pairwise Granger causality. To demonstrate multivariate Granger causality model, we denote $X_{l=1}^{P}$ as lagged variables of time series variable x from time lag 1 to maximum lag P, and similarly $Y_{l=1}^{P}$ from y, $Z_{l=1}^{P}$ from

z. The joint VAR model for multivariate Granger causality is shown as as follows:

$$\begin{cases} y_t = A_1 \cdot Y_{l=1}^P + B_1 \cdot X_{l=1}^P + \varepsilon_{1t} \\ x_t = C_1 \cdot X_{l=1}^P + D_1 \cdot Y_{l=1}^P + \varepsilon_{2t} \end{cases} \tag{3}$$

with the prediction error covariance matrix being:

$$CovMatrix = \begin{bmatrix} var(\varepsilon_{1t}) & cov(\varepsilon_{1t}, \varepsilon_{2t}) \\ cov(\varepsilon_{2t}, \varepsilon_{1t}) & var(\varepsilon_{2t}) \end{bmatrix} \tag{4}$$

Besides lagged variables $X_{l=1}^P$ and $Y_{l=1}^P$, when a new variable z is taken into account, the new VAR model is:

$$\begin{cases} y_t = A_2 \cdot Y_{l=1}^P + B_2 \cdot Z_{l=1}^P + C_2 \cdot X_{l=1}^P + \varepsilon_{3t} \\ z_t = D_2 \cdot Y_{l=1}^P + E_2 \cdot Z_{l=1}^P + F_2 \cdot X_{l=1}^P + \varepsilon_{4t} \\ x_t = G_2 \cdot Y_{l=1}^P + H_2 \cdot Z_{l=1}^P + I_2 \cdot X_{l=1}^P + \varepsilon_{5t} \end{cases} \tag{5}$$

Correspondingly, the prediction error covariance matrix of VAR model in (5) is:

$$\Sigma = \begin{bmatrix} var(\varepsilon_{3t}) & cov(\varepsilon_{3t}, \varepsilon_{4t}) & cov(\varepsilon_{3t}, \varepsilon_{5t}) \\ cov(\varepsilon_{4t}, \varepsilon_{3t}) & var(\varepsilon_{4t}) & cov(\varepsilon_{4t}, \varepsilon_{5t}) \\ cov(\varepsilon_{5t}, \varepsilon_{3t}) & cov(\varepsilon_{5t}, \varepsilon_{4t}) & var(\varepsilon_{5t}) \end{bmatrix} \tag{6}$$

The next step, similar to the pairwise Granger causality testing, is to test whether introducing z can improve the prediction of y and how significant the improvement is. From the VAR model in Equation (3) of variable y and x, and the VAR model in Equation (5) of variable y, z, and x, the conditional Granger causality test from z to y conditioned on x, denoted as $(z \to y|x)$, is:

$$F\text{-}test(var(\varepsilon_{1t}), var(\varepsilon_{3t})) \tag{7}$$

From F-test in Equation (7), a p-value can be used to compare with a threshold to conclude whether z Granger causes y conditioned on x.

2) PCMCI: PCMCI is a causal discovery method described in [24] which identifies relevant variables for conditioning and estimates causality graph from time series data. The method makes use of a "time series graph" made of nodes representing the state variables at different time-lags. If the time lag is denoted by l, a causal link is notated $x_{t-l} \to y_t$, and this link exists if x_{t-l} is not conditionally independent of y_t given the past of all variables. Assuming the causal structure does not change over time, the same links are present at each time step.

The parents $\mathcal{P}(x)$ of a variable x are defined as the set of all nodes with a link towards x. However, estimating these parents directly by testing for conditional independence on the whole past is problematic due to high-dimensionality and because conditioning on irrelevant variables leads to biases.

PCMCI estimates causal links by a two-step procedure [24]:

1. Condition-selection: For every variable α, estimate a superset of parents $\tilde{\mathcal{P}}(\alpha_t)$ with an iterative Markov discovery algorithm [27] such as PC_1 algorithm. The condition-selection step reduces the dimensionality and avoids conditioning on irrelevant variables.

2. Momentary conditional independence (MCI): To test whether $x_{t-l} \to y_t$ with MCI, it evaluates:

$$x_{t-l} \perp y_t \mid \tilde{\mathcal{P}}(y_t), \tilde{\mathcal{P}}(x_{t-l}) \tag{8}$$

Equation (8) checks momentary conditional independence conditions between x_{t-l} and y_t, and checks whether or not x_{t-l} and y_t are not conditionally independent given $\tilde{\mathcal{P}}(y_t)$ and $\tilde{\mathcal{P}}(x_{t-l})$.

3) Dynamic Bayesian Network: Bayesian network [5] is one of many probabilistic graphical models which consists of a directed acyclic graph (DAG) and conditional probability distributions (CPDs) associated with each node in the model. A Bayesian network can be used to make predictions and decisions under uncertainty. A dynamic Bayesian network [19] is similar to a Bayesian network but with a temporal extension, making it an appropriate graphical model to use for temporal datasets. The two main steps to creating a probabilistic graphical model are structure learning and parameter learning.

In this paper, we adopt the approach in [33] for dynamic Bayesian network learning. The approach first expands variable set by adding new variables for each original variable through time lagging. For instance, P new variables can be created from original variable x: x_{t-i} for i from 1 to maximum lag P. With the expanded variable set, the K2 algorithm [9] is used to search through all possible causality graph structures and identify which structure has the highest possibility to produce the data. In this score-based structure learning approach, Bayesian information criterion (BIC) scoring is used. Next, after causality graph is generated for expanded variable set, the causality graph is simplified by removing lagged variable and combining the causality edges. For instance, two edges $x_{t-2} \to y_{t-1}$ and $x_{t-3} \to y_t$ are combined to one edge $x \to y$ in the final graph.

Moreover, for the sake of computational time, the time series data is partitioned into bins. Each bin defines a set of sub ranges, then the data is assigned to each labeled bin. For example, if the lowest value of the dataset is -5, and the highest value is 5. With the total bin number 10, a value of 1.2351 can be placed in a bin labeled 7, whose range is [1, 2). This approach increases the state counts of each variable and allows for faster computation.

III. A Two-Phase Hybrid Causality Ensemble Framework

To deal with both increasing volume of available input data and increasing variety of available causality discovery algorithms, we propose a two-phase hybrid causality ensemble framework to achieve ensemble of both multiple causality discovery algorithms as base learners and multiple data partitions as base learner's input data. Before diving into the details of this two-phase ensemble framework, we first explain how ensemble could be done with only data ensemble and algorithm ensemble. We note most causality discovery algorithms generate not only cause-effect relationships, but also time lag and probability of each relationship. In this paper, we only focus on structure causality ensemble, namely how multiple directed graphs can be combined into one, and leave the time lag and probability ensemble for future work.

A. Algorithm Ensemble for Causality Discovery

Algorithm ensemble approach deals with algorithm variety by applying different causality discovery algorithms as base learners with the same input data and later combining all base learner results. Each causality discovery algorithm mines the same time series dataset and produces its own directed graph where nodes denote time series variables and each directed edge denotes a cause-effect relationship between the two connected variables. Because each base learner works on the same input data, the nodes of result graphs are the same for different base learners. But different base learners could produce different causality edges. Then by applying a certain base learner combination method, such as majority voting, we can derive a new directed graph as ensemble result. The nodes in the ensemble graph are the same with the results in each base learner. For graph edges, we can iterate all possible edges of the graph and decide whether this edge should be in the ensemble graph by combining corresponding edges in base learner graph result. If we use majority voting as combination method, an edge will be in ensemble graph only if the edge appears in more than half of base learner graphs.

By applying algorithm ensemble, the ensemble result is often more accurate than utilizing only one single causality discovery algorithm. However, when the size of input time series dataset gets larger, the execution time of algorithm level ensemble increases dramatically because every base learner will take longer time to finish. Thus, a non-scalable algorithm ensemble approach is not enough to meet the challenge of dealing with the increasing data size.

B. Data Ensemble for Causality Discovery

Data ensemble approach deals with data volume challenge by first partitioning data into smaller datasets, then using the same causality discovery algorithm as base learners with data partitions, and later combining all base learner results. Data partitioning is often done horizontally, not vertically, so that each data partition can still have all variables needed for multivariate causality learning. Because input data are often time series, data partitioning can be easily done by splitting the overall time ranges into smaller time ranges. Similar to algorithm ensemble, the nodes of resulting causality graph are the same for different base learners and edges of the graphs might be different. Then we can derive ensemble graph using the same base learner combination method in the previous subsection. The limitation of this approach is that it does not deal with variety of causality learning algorithms.

C. Two-Phase Hybrid Data-Algorithm Ensemble for Causality Discovery

To address the challenges of diverse causality discovery results and increasing data size, we further integrate data ensemble and algorithm ensemble into one framework as illustrated in Figure 1, which conducts two-phase hybrid ensemble. We implement this generic framework as ***data-algorithm ensemble***, which means it conducts data ensemble first in phase 1 and then algorithm ensemble in phase 2.

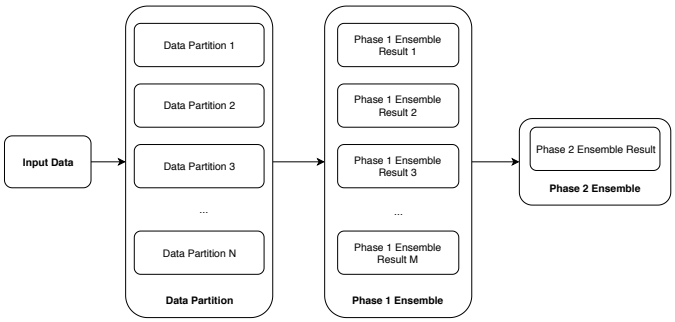

Fig. 1. Two-phase hybrid ensemble framework for causality discovery.

In the causality ensemble framework, the input data is first partitioned into different data slices from 1 to N. Then, phase 1 causality computation is executed to get N phase 1 ensemble result for each causality method. Next, all the phase 1 data ensemble results are combined into one final output through phase 2 algorithm ensemble.

IV. Approach of Two-Phase Hybrid Causality Ensemble Framework

Based on the two-phase hybrid causality ensemble framework explained in previous section, the data-algorithm causality ensemble approach is developed as illustrated in Figure 2. This data-algorithm ensemble approach is designed to effectively learn causal relationships from three data-driven causality learning approaches: multivariate Granger causality (MGC), $PCMCI$ and dynamic Bayesian network (DBN).

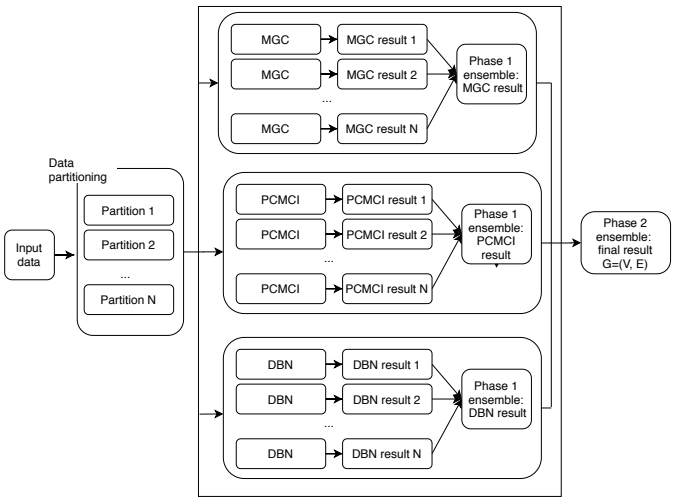

Fig. 2. Illustration of data-algorithm ensemble learning approach.

The data-algorithm ensemble approach (see Figure 2) denotes that data ensemble happens in phase 1, then algorithm ensemble happens in phase 2. In this approach, the input data is first partitioned into N slices. Then, each of the causality discovery method (MGC, $PCMCI$ and DBN) is executed on all the partitioned data to get one causality output directed graph for each data slices. For example, MGC outputs MGC_Result_1, MGC_Result_2, ... MGC_Result_N.

Different methods are executed in serial in the order of MGC, $PCMCI$, DBN. The outputs from all partitioned data slices corresponding to each causality method are collected for phase 1 data ensemble. The phase 1 ensemble results are computed by majority voting. In the following step, phase 1 ensemble results of each causality method ($MGC\_Ensemble$, $PCMCI\_Ensemble$ and $DBN\_Ensemble$) are combined using ensemble methods again into get a phase 2 algorithm ensemble causality result graph as final output.

Algorithm 1: Data-Algorithm Ensemble *(Data-Algorithm_Ensemble)*

Input: Different causality discovery methods: Multivariate Granger causality: MGC, PCMCI: $PCMCI$, Dynamic Bayesian Network: DBN,
Time series data: D,
Number of data partitions: N
Output: A directed causality graph: $G = (V, E)$

1: Partition data D into N partitions as $\{d\}$ = d_1, d_2, ..., d_N
2: Get E_{MGC} = $Data\text{-}Algorithm\_Phase\_1(MGC, \{d\})$
3: Get E_{PCMCI} = $Data\text{-}Algorithm\_Phase\_1(PCMCI, \{d\})$
4: Get E_{DBN} = $Data\text{-}Algorithm\_Phase\_1(DBN, \{d\})$
5: ## Phase 2 edge ensemble:
6: **for** unique edges $\{e_i\}$ in E_{MGC}, E_{PCMCI} and E_{DBN} **do**
7: Count e_i appearance in E_{MGC}, E_{PCMCI} and E_{DBN} as n_i
8: **if** $n_i >= 2$ **then**
9: Add e_i to final graph G
10: **end if**
11: **end for**
12: Output $G = (V, E)$

The data-algorithm ensemble approach includes two algorithms: Algorithm 1 (*Data-Algorithm_Ensemble*) for the two-phase hybrid ensemble approach, which regards to the full process in Figure 2 and Algorithm 2 (*Data-Algorithm_Phase_1*) for phase 1 data ensemble corresponding to each phase 1 ensemble block in Figure 2.

The input of the *Data-Algorithm_Ensemble* (Algorithm 1) includes different causality discovery methods, which are multivariate Granger causality (MGC), PCMCI ($PCMCI$) and Dynamic Bayesian Network (DBN), time series input data D, and the number of data partitions N. The logic of Algorithm 1 for the whole ensemble process is as follows. In line 1, the input dataset D is first partitioned into N slices by its timestamp as $\{d\}$ = d_1, d_2, ..., d_N where the time interval of each slice is only $1/N$ of the original time series. Then it calls Algorithm 2 *(Data-Algorithm_Phase_1)* to execute each causality discovery method to get phase 1 ensemble causality edge set E_{MGC}, E_{PCMCI} and E_{DBN} from all the data partitions in lines 2-4. Finally, in lines 6-11, phase 2 ensemble result is computed by majority voting on edge set of all causality mining methods, E_{MGC}, E_{PCMCI} and E_{DBN}, that if two or more causality ensemble edge sets contain the same edge, this edge is added into final output graph $G = (V, E)$ with V denoting nodes and E as edges in line 12.

Algorithm 2: Phase 1 Ensemble for Data-Algorithm Ensemble *(Data-Algorithm_Phase_1)*

Input: Causality discovery method: $Causality$,
Data partition set: $\{d\}$
Output: A set of directed edges in Graph corresponding to causality discovery method: $E_{causality}$

1: **for** each data partition d_i in $\{d\}$ **do**
2: Get causality edge set from causality computation: E_i = $Causality(d_i)$
3: **end for**
4: ## Phase 1 edge ensemble:
5: **for** unique edges $\{e_j\}$ in all E_i **do**
6: Count e_j appearance in all E_i as n_j
7: **if** $n_j > N/2$ **then**
8: Add e_j to $E_{causality}$
9: **end if**
10: **end for**
11: Output $E_{causality}$

The phase 1 data ensemble in the data-algorithm ensemble approach, namely *Data-Algorithm_Phase_1* is shown in Algorithm 2. Its inputs include the specific causality discovery method $Causality$, and the partitioned time-series dataset $\{d\}$. In lines 1-3, the causality discovery method executes for each data partition d_i in $\{d\}$ to output a causality edge set E_i from $Causality(d_i)$. Since this causality edge set contains edges from each partition, in lines 5-10, phase 1 ensemble method loops to check if the number of a given edge e_j appears in more than half of the partition edge set. For instance, if there are 10 partitions, and a causality edge (x_1, x_2) appears 6 times in all the partition edge set, it is added to the phase 1 ensemble output $E_{causality}$ as in line 8 then be output as in line 11.

V. Parallel Two-Phase Hybrid Causality Ensemble Learning via Spark Big Data Engine

The above two-phase hybrid causality ensemble approach is further implemented in parallel via Spark [1] to achieve scalability to deal with big data in two aspects: 1) automatic data partitioning and 2) parallel function mapping.

Regarding the data partitioning part in our parallel implementation, the data is first load into Spark as resilient distributed dataset (RDD); then it is automatically partitioned by timestamp of each record, as in the phase 2 algorithm ensemble of data-algorithm ensemble, in Algorithm 1 line 1. More specifically, every data partition, as a chunk of the large distributed dataset, is assigned an index i for phase 1 ensemble in next step.

For parallel function mapping, the parallelization of data-algorithm ensemble is implemented in its phase 1 data ensemble, as in Algorithm 2 lines 1-3. With Spark RDD partitioning, now each data partition d_i becomes an RDD partition. Next, these RDD partitions are mapped to be transformed by the causality discovery method $Causality$ in parallel, then be reduced as the edge set E_i for later phase 2 ensemble computation.

VI. Experiments

The experiments were conducted on top of the HPCF2018 cluster at the University of Maryland, Baltimore County [3], where each computing node containing two 18-core CPUs and 384 GB memory. For our experiment environment, one cluster contains one master node and several worker nodes. Moreover, the Spark programs are managed by Slurm workload manager in standalone cluster mode. For software, Python (version 3.6.8), Spark (version 2.4) are used. For Spark configurations, each node contains one executor, each driver/executor's memory is 200GB, and partition number is set as 48.

For test data, we created four synthetic datasets to evaluate our proposed algorithms' performance. One important reason for synthetic dataset generation is to know causality ground truth so we could evaluate learning result accuracy. Similar to the synthetic dataset generation approach for Granger causality and DBN evaluation in [33], we generated our synthetic dataset based on linear and nonlinear causal dependency Equation (9) and Equation (10), where εs are random noises. The causality graph for the equation can be found at Figure 3 and Figure 4. The linear and nonlinear datasets with different sizes (namely 1 million and 10 million for row numbers) were generated using the same equations correspondingly.

$$\begin{cases} x_1(t) = 0.95 \cdot \sqrt{2} \cdot x_1(t-1) - 0.90 \cdot x_1(t-2) + \varepsilon_1 \\ x_2(t) = 0.5 \cdot x_2(t-1) + \varepsilon_2 \\ x_3(t) = -0.5 \cdot x_1(t-1) + 0.25 \cdot \sqrt{2} \cdot x_3(t-1) \\ \quad + 0.25 \cdot \sqrt{2} \cdot x_2(t-1) + \varepsilon_3 \\ x_4(t) = -0.95 \cdot x_4(t-1) - 0.25 \cdot \sqrt{2} \cdot x_3(t-1) + \varepsilon_4 \\ x_5(t) = 0.5 \cdot x_1(t-1) + 0.95 \cdot x_2(t-2) \\ \quad - 0.25 \cdot \sqrt{2} \cdot x_3(t-1) + 0.5 \cdot x_5(t-1) + \varepsilon_5 \end{cases} \tag{9}$$

$$\begin{cases} x_1(t) = 0.125 \cdot \sqrt{2} \cdot \exp(-x_1(t-1)^2/2) + \varepsilon_1 \\ x_2(t) = 1.2 \cdot \exp(-x_1(t-1)^2/2) + \varepsilon_2 \\ x_3(t) = -1.05 \cdot \exp(-x_1(t-1)^2/2) \\ \quad + 0.2 \cdot \sqrt{2} \exp(-x_2(t-2)^2/2) + \varepsilon_3 \\ x_4(t) = -1.15 \cdot \exp(-x_1(t-2)^2/2) \\ \quad + 0.2 \cdot \sqrt{2} \cdot \exp(-x_4(t-1)^2/2) \\ \quad + 1.35 \cdot \exp(-x_3(t-1)^2/2) + \varepsilon_4 \\ x_5(t) = -1.15 \cdot \exp(-x_2(t-1)^2/2) + \varepsilon_5 \end{cases} \tag{10}$$

A. Baseline Approaches and Parameter Setting

We employed seven baseline approaches in our experiments. The first three were single causality discovery approaches: Multivariate Granger causality (MGC), $PCMCI$ and Dynamic Bayesian Network (DBN). The next three were corresponding data ensemble approaches for each of the three single causality discovery approaches following the way described in Section III-B. The last one was an algorithm ensemble approach by combining all the three single causality discovery approaches following the way described in Section III-A. For experiment parameter settings, we set the maximum time lagging as 3 for synthetic data and the p-value threshold as 0.05 for both MGC and $PCMCI$ tests. Besides, the total bin number for DBN was set as 5 to reduce computation time. In $PCMCI$ method, we utilized its different conditional independence tests for linear and nonlinear causality discovery. For nonlinear conditional independence tests, as we had a large dataset, $RCOT$ test was applied.

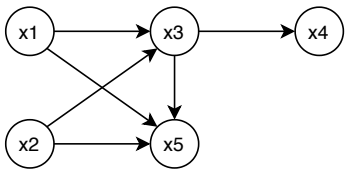

Fig. 3. Linear synthetic data ground truth causal graph.

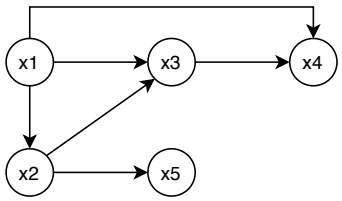

Fig. 4. Nonlinear synthetic data ground truth causal graph.

B. Accuracy Evaluation

We employ Structural Hamming Distance (SHD) metric [29] to compare accuracy of different approaches. SHD is a common metric to measure the difference between two directed graphs with the same node set. SHD value is defined as the total step count of three types of actions needed to transform from one direct graph to another direct graph: 1) reversing an edge's direction, 2) removing an extra edge, 3) adding a missing edge. We calculate SHD between ground truth graph and each learned graph. The lower SHD value means the more similarity between the two graphs, so the algorithm that generates the learned graph is more accurate.

We measured the accuracy of single causality discovery method of MGC, $PCMCI$ and DBN and the three data ensemble baseline approaches and the algorithm ensemble causality ensemble approach. The results were shown columns 2-8 of Table I. For linear datasets, we could see from the table that both data ensemble and algorithm approach could achieve the same or better accuracy than single causality discovery approaches. For nonlinear datasets, data ensemble approaches still performs better in accuracy; however, algorithm ensemble could perform a little bit worse due to two algorithm making the same wrong prediction on certain edges.

The accuracy of two-phase hybrid causality ensemble approach was shown in column 9 of Table I. Compared to

TABLE I
STRUCTURAL HAMMING DISTANCE (SHD) COMPARISON OF DIFFERENT CAUSALITY DISCOVERY APPROACHES

	MGC	PCMCI	DBN	Data-level Ensemble GC	Data-level Ensemble PCMCI	Data-level Ensemble DBN	Algorithm-level Ensemble	Two-phase Data-Algorithm Ensemble
Linear 1M	1	1	4	1	1	4	0	0
Linear 10M	1	1	3	1	1	4	0	0
Nonlinear 1M	5	13	1	4	3	3	4	0
Nonlinear 10M	6	6	1	2	1	1	3	0

TABLE II
EXECUTION TIME TABLE: SERIAL BASELINE EXPERIMENTS(H:MM:SS.SS)

Synthetic Dataset	MGC	PCMCI	DBN	Data-level Ensemble MGC	Data-level Ensemble PCMCI	Data-level Ensemble DBN	Algorithm -level Ensemble
1M Linear	0:00:08.16	0:07:23.28	0:27:14.83	0:01:49.78	0:06:35.17	0:28:40.29	0:31:59.37
10M Linear	0:01:21.88	1:31:06.78	5:58:52.31	0:18:52.88	0:54:17.20	3:45:56.39	6:45:01.24
1M Nonlinear	0:00:07.83	0:24:07.11	0:24:39.91	0:01:13.25	0:28:00.59	0:27:17.02	0:45:26.36
10M Nonlinear	0:01:24.59	4:33:06.51	5:15:30.22	0:18:15.43	3:57:54.57	2:56:27.57	8:47:18.09

algorithm/data ensemble baseline approaches, our two-phase causality ensemble approach achieves perfect accuracy since their SHD values are all zero. In linear experiments, compared to data ensemble and algorithm ensemble baseline approaches, our two-phase hybrid causality ensemble approach could get the same or better results. In nonlinear experiments, two-phase hybrid ensemble approach achieves better accuracy than both data ensemble and algorithm ensemble. They both perform better than all the baseline approaches in accuracy for all the datasets.

C. Scalability Evaluation

We conducted scalability experiments for our proposed two-phase hybrid causality ensemble approaches given different sizes of datasets at a distributed computing environment mentioned above with 5, 7 and 9 compute nodes.

1) Execution Time: The execution times of all the baseline algorithms are shown in Table II for linear and nonlinear, 1M and 10M dataset testing. The execution times for parallel experiments are shown as in Table III and Table IV for 1M and 10M records of linear dataset. For nonlinear dataset, the execution times are recorded as in Table V and Table VI for 1M and 10M data correspondingly.

TABLE III
EXECUTION TIME: PARALLEL EXPERIMENTS ON 1M LINEAR DATA

Linear 1M	Data-level Parallel Ensemble MGC	Data-level Parallel Ensemble PCMCI	Data-level Parallel Ensemble DBN	Two-phase Ensemble Data-Algorithm
4 Worker Nodes	0m19.037s	10m51.091s	2m11.193s	3m55.372s
6 Worker Nodes	0m20.068s	8m57.787s	1m12.336s	2m54.335s
8 Worker Nodes	0m20.030s	6m46.573s	1m1.355s	2m4.477s

TABLE IV
EXECUTION TIME: PARALLEL EXPERIMENTS ON 10M LINEAR DATA

Linear 10M	Data-level Parallel Ensemble MGC	Data-level Parallel Ensemble PCMCI	Data-level Parallel Ensemble DBN	Two-phase Ensemble Data-Algorithm
4 Worker Nodes	2m02.216s	51m33.383s	10m46.239s	24m03.187s
6 Worker Nodes	1m46.703s	35m48.498s	7m55.188s	18m39.600s
8 Worker Nodes	1m35.964s	22m34.472s	6m46.441s	12m50.270s

TABLE V
EXECUTION TIME: PARALLEL EXPERIMENTS ON 1M NONLINEAR DATA

Nonlinear 1M	Data-level Parallel Ensemble MGC	Data-level Parallel Ensemble PCMCI	Data-level Parallel Ensemble DBN	Two-phase Ensemble Data-Algorithm
4 Worker Nodes	0m20.195s	13m3.590s	2m19.261s	7m25.565s
6 Worker Nodes	0m19.066s	11m5.580s	1m28.426s	5m31.534s
8 Worker Nodes	0m18.492s	8m1.691s	1m1.877s	4m13.503s

TABLE VI
EXECUTION TIME: PARALLEL EXPERIMENTS ON 10M NONLINEAR DATA

Nonlinear 10M	Data-level Parallel Ensemble MGC	Data-level Parallel Ensemble PCMCI	Data-level Parallel Ensemble DBN	Two-phase Ensemble Data-Algorithm
4 Worker Nodes	2m02.023s	39m37.026s	10m46.367s	24m15.382s
6 Worker Nodes	1m50.979s	28m45.792s	7m52.148s	18m1.137s
8 Worker Nodes	1m37.207s	25m41.998s	6m54.563s	16m0.709s

The Spark based parallel implementations of the three data-ensemble baseline approaches use the same techniques in Section V. We measured their execution times as in columns 2-4 of parallel experiments execution time tables. We also recorded the execution times of data-algorithm ensemble showing in column 5 of all execution time tables for parallel experiments.

We note Tables III, IV, V, VI show data-level parallel ensemble PCMCI is slower than our two-phase ensemble. By checking the execution logs, we found it is because at the runtime the Spark session encountered idle time for executors in the cluster, thus the computation time is fairly long. However, we did not see the same behavior in the two-phase ensemble experiments. The reason for this unexpected result will be further investigated.

2) Speed Up: By comparing the execution times our parallel hybrid approaches in Tables III, IV, V, VI with the execution times of our serial algorithm ensemble baseline approach in Table II, we evaluated the speed ups of our parallel hybrid ensemble approaches. The algorithm ensemble baseline was executed on a single node. As shown in Figures 5 and 6, both achieved near linear speed up. Figure 5 shows the speed ups of two-phase ensemble in comparison to algorithm ensemble baseline for 10M row linear dataset. With 8 worker nodes, the speed up is more than 32 times. Similarly, Figure 6 shows speed up of two-phase ensemble compared to algorithm ensemble baseline for 10M nonlinear dataset. Its speed up, when running with 8 worker nodes, reaches 17 times compared to the baseline. Our approaches can achieve better than linear speedup because the time complexity of each baseline algorithm is worse than $O(n)$. For instance, Granger causality algorithm's execution time grows quadratically with the increase of the data record number [4]. By splitting data into N partitions, the execution time for each data partition is less than $1/N$ of the baseline serial approach.

VII. RELATED WORK

There have been many studies on ensemble learning and scalable/parallel machine learning. But we believe our work is the first study dealing with both algorithm variety and data volume for causality discovery. We also did not find many studies directly on ensemble learning for causality. Because

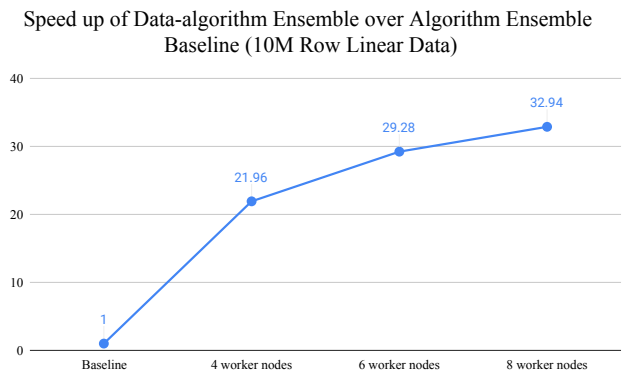

Fig. 5. Speed up of two-phase ensemble compared to algorithm ensemble baseline for 10M row linear dataset.

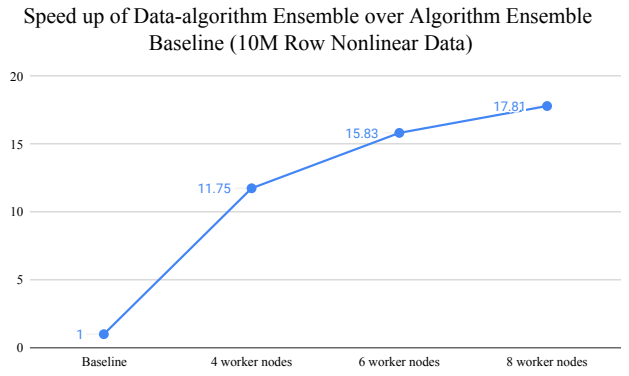

Fig. 6. Speed up of two-phase ensemble compared to algorithm ensemble baseline for 10M row nonlinear dataset.

causality graph can be categorized as a type of probabilistic graphic model, we first discuss and compare with related work on ensemble learning for probabilistic graphic models in the first subsection. We further discuss and compare additional big data parallel ensemble learning work beyond probabilistic graphic models.

To achieve probabilistic graphical model ensemble, using the three categories explained in Section II, existing ensemble learning approaches can also be categorized into 1) algorithm ensemble for work at [17], 2) data ensemble work at [14], [30], and 3) hybrid ensemble for both data and algorithm at [28] and [7]. In algorithm ensemble category, [17] supports parallel ensemble learning of multiple classifiers on the same data. As a data ensemble approach, [14] first splits the training data, then trains Bayesian sub-networks in parallel, finally does boosting as ensemble method on the trained sub-networks to get the learning result. [30] is also a data ensemble approach for Bayesian network learning from big datasets to achieve better scalability and accuracy. As a hybrid ensemble approach, [28] conducts two-phase (algorithm ensemble for each data partition and data ensemble for multiple data partitions) Bayesian network ensemble learning. The main differences of this work and [28] are: 1) this work first conducts data ensemble among all data partitions and then algorithm ensemble for different algorithms where [28] first conducts algorithm ensemble then data ensemble; 2) our algorithm-level ensemble belongs to heterogeneous ensemble because each learning algorithm uses its own causality discovery models, while [28] belongs to homogeneous ensemble with different learning algorithms of the same Bayesian network model; 3) this paper targets causality discovery instead of Bayesian network learning.

Besides the probabilistic graphic model related ensemble studies in the previous subsection, most other big data parallel ensemble learning algorithms are tree based where different trees can be trained in parallel with a data subset, then results from multiple trees are ensembled via majority voting (e.g., [7]) or tree boosting (e.g., XGBoost [8]). There are two main approaches of data partitioning: horizontal data partitioning based on rows and vertical data partitioning based on columns. [7] contains horizontal data partitioning and parallel learning among the data partitions. Input data is first partitioned vertically to divide training data features to independent subsets. Then each task loads the data from one feature subset to train an independent tree and multiple trees can be trained in parallel. For XGBoost [8], parallel training is done via horizontally partitioned data and they differ in how different trees are ensembled. As a comparison, parallelization in our hybrid ensemble approaches is done via horizontal data partitioning because all features are needed for each training and our data has time dependency. Further, multiple learning algorithms are employed in our data-algorithm ensemble while the above related works only employ the same learning algorithm for different data partitions.

VIII. Conclusions

Causality discovery is a fundamental research topic in many disciplines and discovered cause-effect relationships can help explain why a system has certain behavior or state. Nowadays, data-driven causality discovery faces two challenges: 1) the large volume of datasets to be learned from and 2) the variety of causality discovery algorithms. To deal with these two challenges, this paper proposes a two-phase hybrid ensemble causality learning framework and an implementation approach for scalable ensemble causality learning. Experiments show our algorithms outperform baseline ones in terms of both accuracy and execution time.

For future work, we will focus on the following aspects. First, we will extend the work to further enable ensemble of time lag and probability of causal edges. Second, we will study how to best select from many available causality learning algorithms, i.e., through diversity measurement, for better ensemble result accuracy. Further, we plan to apply the framework and algorithms to real-world climate applications and evaluate their effectiveness through the applications.

Acknowledgment

This work is supported by grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Math-

ematics and Atmospheric Sciences using Advanced Cyberinfrastructure Resources (OAC–1730250), and grant CAREER: Big Data Climate Causality Analytics (OAC–1942714) from the National Science Foundation. The execution environment is provided through the High Performance Computing Facility at UMBC.

REFERENCES

[1] Homepage — Apache Spark Project. http://spark.apache.org, 2020. Accessed: 2020-03-01.

[2] Scalable Ensemble Learning for Causality Discovery. https://github.com/big-data-lab-umbc/ensemble_causality_learning, 2020. Accessed: 2020-05-28.

[3] The UMBC High Performance Computing Facility (HPCF). https://hpcf.umbc.edu/, 2020. Accessed: 2020-03-01.

[4] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 66–75, New York, NY, USA, 2007. ACM.

[5] I. Ben-Gal. *Bayesian Networks*. Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons, 2007.

[6] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[7] J. Chen, K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li. A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Transactions on Parallel and Distributed Systems*, 28(4):919–933, 2017.

[8] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[9] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[10] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[11] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[12] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[13] N. J. Holbrook, J. N. Brown, J. Davidson, M. Feng, A. J. Hobday, J. M. Lough, S. McGregor, S. B. Power, and J. S. Riseby. El niño–southern oscillation. 2012.

[14] J. Hu, G. Wu, P. Sun, and Q. Xiong. A parallel bayesian network learning algorithm for classification. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 259–263. IEEE, 2016.

[15] S. Hussung, S. Mahmud, A. Sampath, M. Wu, P. Guo, and J. Wang. Evaluation of Data-driven Causality Discovery Approaches among Dominant Climate Modes. *Technical Report HPCF-2019-12, UMBC High Performance Computing Facility, University of Maryland, Baltimore County*, 2019.

[16] H. Luetkepohl. *The New Introduction to Multiple Time Series Analysis*. 01 2005.

[17] A. L. Madsen, F. Jensen, A. Salmerón, H. Langseth, and T. D. Nielsen. A parallel algorithm for bayesian network structure learning from large data sets. *Knowl.-Based Syst.*, 117:46–55, 2017.

[18] M. C. McGraw and E. A. Barnes. Memory matters: A case for granger causality in climate variability studies. *J. Clim.*, 31(8):3289–3300, Jan. 2018.

[19] K. P. Murphy and S. Russell. Dynamic bayesian networks: representation, inference and learning. 2002.

[20] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling. Climate data challenges in the 21st century. *science*, 331(6018):700–702, 2011.

[21] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[22] J. Pearl. Simpson's paradox, confounding, and collapibility. *Causality: models, reasoning and inference*, pages 173–200, 2009.

[23] R. Polikar. Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer, 2012.

[24] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting causal associations in large nonlinear time series datasets. https://arxiv.org/abs/1702.07007v2, 2018. Accessed: 2018-06-28.

[25] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[26] H. Song, J. Tian, J. Huang, P. Guo, Z. Zhang, and J. Wang. Hybrid causality analysis of enso's global impacts on climate variables based on data-driven analytics and climate model simulation. *Frontiers in Earth Science*, 7:233, 2019.

[27] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 10 2007.

[28] Y. Tang, J. Wang, M. Nguyen, and I. Altintas. Penbayes: A multi-layered ensemble approach for learning bayesian network structure from big data. *Sensors*, 19(4400), 2019.

[29] 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.

[30] J. Wang, Y. Tang, M. Nguyen, and I. Altintas. A scalable data science workflow ap-proach for big data bayesian network learning. In *Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing (BDC 2014)*, pages 16–25, 2014.

[31] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[32] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5:14750, 2015.

[33] C. Zou, K. J. Denby, and J. Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC Bioinformatics*, 10:122, Apr. 2009.