Causal Discovery in Temporal Domain from Interventional Data

Peiwen Li
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University, Shenzhen, China
Alibaba Group, Hangzhou, China
lpw22@mails.tsinghua.edu.cn

Yuan Meng*
Department of Computer Science and Technology
Tsinghua University, Beijing, China
yuanmeng@mail.tsinghua.edu.cn

Xin Wang*
Department of Computer Science and Technology, BNRist
Tsinghua University, Beijing, China
xin_wang@tsinghua.edu.cn

Fang Shen
Yue Li
Alibaba Group
Hangzhou, China
ziru.sf@alibaba-inc.com
yueqian.ly@alibaba-inc.com

Jialong Wang*
Alibaba Group
Hangzhou, China
quming.wjl@alibaba-inc.com

Wenwu Zhu*
Department of Computer Science and Technology, BNRist
Tsinghua University, Beijing, China
wwzhu@tsinghua.edu.cn

ABSTRACT

Causal learning from observational data has garnered attention as controlled experiments can be costly. To enhance identifiability, incorporating intervention data has become a mainstream approach. However, these methods have yet to be explored in the context of time series data, despite their success in static data. To address this research gap, this paper presents a novel contribution. Firstly, a temporal interventional dataset with causal labels is introduced, derived from a data center IT room of a cloud service company. Secondly, this paper introduces TECDI, a novel approach for temporal causal discovery. TECDI leverages the smooth, algebraic characterization of acyclicity in causal graphs to efficiently uncover causal relationships. Experimental results on simulated and proposed real-world datasets validate the effectiveness of TECDI in accurately uncovering temporal causal relationships. The introduction of the temporal interventional dataset and the superior performance of TECDI contribute to advancing research in temporal causal discovery. Our datasets and codes have released at https://github.com/lpwpower/TECDI.

CCS CONCEPTS

• **Computing methodologies** → *Causal reasoning and diagnostics*; *Temporal reasoning*.

KEYWORDS

Causal Discovery, Time Series, Intervention

ACM Reference Format:
Peiwen Li, Yuan Meng, Xin Wang, Fang Shen, Yue Li, Jialong Wang, and Wenwu Zhu. 2023. Causal Discovery in Temporal Domain from Interventional Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615177

*Corresponding authors

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615177

1 INTRODUCTION

The exploration of causal knowledge has stood as a fundamental undertaking in scientific research. Causal knowledge aids in comprehending intricate systems, including cloud service systems, and guides further optimization decisions. Nonetheless, even for very powerful large-scale models, it is challenging to handle causal discovery and causal inference tasks [9, 21]. Conducting randomized controlled trials, which serve as the gold standard for causal discovery, often proves arduous due to factors such as experimental costs and ethical constraints. As a result, causal discovery algorithms relying on observational data have garnered substantial attention in recent years. However, from the observational data, the true causal graph is only identifiable up to a Markov equivalence class under the faithfulness assumption [6]. Encouragingly, identifiability can be enhanced by obsorbing additional information from interventional data [4, 19], which have been applied in the causal discovery from static data [3, 8].

Temporal data is prevalent across diverse fields, such as industry [10], finance [7], meteorology [11], and neuroscience [2], which need causal knowledge. Unlike static data, temporal data provides a dynamic perspective, enabling us to capture the evolution of causal relationships. Therefore, causal discovery within the temporal domain extends beyond the mere identification of causal links between variables (such as X causes Y), incorporating the discovering causal lags (such as X_{t-i} causes Y_t with lag i). The approaches to causal discovery from temporal data roughly fall into three categories, namely, Granger causality, constraint-based methods and score-based methods. Besides the sole work [5], there is currently no method available to utilize interventional data for temporal causal discovery. The main reason for this blank may be that real intervention time series data are difficult to obtain.

Inspired by recent advancements in anomaly detection on time series [20], this paper introduces a novel approach for constructing a temporal interventional dataset based on monitoring data from a data center operated by Alibaba. In our dataset, we consider a source anomaly, where no other anomalies were detected in the previous period, as the intervention. Building upon this dataset, we design a

method called temporal causal discovery based on interventional Data (TECDI). TECDI leverages the smooth, algebraic characterization of acyclicity of causal graphs [12, 22] to achieve differential causal structure learning. Powered by the theoretically-grounded method based on static interventional data [4], our method is able to capture both contemporaneous and time-lagged causal relationships simultaneously with identifiability guarantee. To validate the effectiveness of TECDI, we conducted experiments using both simulated and proposed real-world datasets. The results clearly demonstrate the superiority of our method in discovering temporal causal relationships accurately.

This paper makes two significant contributions. Firstly, to the best of our knowledge, TECDI is the first attempt to employ interventional data for real-world temporal causal discovery. This novel approach opens up new avenues for research in this field and expands the possibilities for understanding causal relationships in practical complex system. Secondly, we construct the first-ever real-world temporal interventional dataset, which we believe will greatly facilitate future investigations and advancements in this area of research.

2 RELATED WORK

The theoretical evidence confirms that the availability of interventional data can greatly enhance the identification of the underlying causal structure [4, 18]. However, the research progress in this direction has been limited due to the substantial challenges associated with designing intervention experiments and obtaining the necessary data. The study conducted in [8] addressed the challenge of learning causal graphs with latent variables from a combination of observational and interventional distributions, where the interventional targets are unknown. The proposed approach utilized a Ψ-Markov property to tackle this problem. In a different context, the work presented in [1] introduced a randomized algorithm that recovers the complete causal graph while minimizing the intervention cost. This algorithm relied on a novel characterization based on p-colliders. Additionally, DCDI [3] is a versatile method for causal discovery that operates under continuous constraints. It can effectively leverage different types of interventional data and incorporate expressive neural architectures like normalizing flows. These works, however, do not address the problem of temporal causal discovery using interventional data. The only study available on temporal data, as presented in [5], does not include an evaluation on real-world datasets.

3 METHODS

3.1 Definitions

3.1.1 Dynamic causal graphic models. We extend the formulation of causal graphic models to temporal data and form a dynamic causal graphic model, defined by a distribution P_Y over the vector $Y = (Y_0, Y_1, \ldots, Y_p)$ and a DAG $\mathcal{G} = (V, E)$. $Y_0 = (X_{0,1}, \ldots, X_{0,d})$, where d is the number of different contemporaneous variables, and $Y_1, \ldots, Y_p$ are time-lagged versions of Y_0, i.e. $Y_k = (X_{k,1}, \ldots, X_{k,d}), k \in \{1, \ldots, p\}$. Therefore, $Y = (X_{0,1}, \ldots, X_{0,d}, \ldots, X_{p,1}, \ldots, X_{p,d})$, abbreviated to $Y = (X_1, \ldots, X_{(p+1)d})$. Besides, each node $i \in V = \{1, \ldots, (p+1)d\}$ is related with a variable $X_i, i \in \{1, \ldots, (p+1)d\}$, and each edge $(i, j) \in E$ represents a direct causal relation from variable X_i to X_j. Under the Markov assumption of the distribution P_Y and graph $\mathcal{G}$, the joint distribution can be factorized as

$$p(x_1, \ldots, x_{(p+1)d}) = \prod_{j=1}^{(p+1)d} p_j(x_j | x_{\pi_j^{\mathcal{G}}}), \tag{1}$$

where $\pi_j^{\mathcal{G}}$ is the set of parents of the node j in the graph $\mathcal{G}$, and $x_{\pi_j^{\mathcal{G}}}$ denotes the entries of the vector x with indices in $\pi_j^{\mathcal{G}}$. We also assume *causal sufficiency*, which means there is no hidden common cause that is causing more than one variable in Y [13].

3.1.2 Intervention. An intervention on a variable x_j is corresponding to replacing its conditional density $p_j(x_j | x_{\pi_j^{\mathcal{G}}})$ by a new one. Apart from that, we define the *interventional target*, a set $I \subseteq V$ consisting of the variables been intervened simultaneously, and the *intervention family* $\mathcal{I} := (I_1, \ldots, I_Q)$, where Q is the number of interventions. To be specific, the observational setting, where no variables were intervened, is always known and denoted by $I_1 := \emptyset$. The qth interventional joint density can be represented as

$$p^{(q)}(x_1, \ldots, x_{(p+1)d}) := \prod_{j \notin I_q} p_j^{(1)}(x_j | x_{\pi_j^{\mathcal{G}}}) \prod_{j \in I_q} p_j^{(q)}(x_j | x_{\pi_j^{\mathcal{G}}}). \tag{2}$$

Note that, in temporal domain, only time-lagged variables $X_{k,j}, k \in \{1, \ldots, p\}$ and other variables $X_{k,i}, k \in \{0, \ldots, p\}, i \in \{1, \ldots, d\} \backslash \{j\}$ can affect a variable $X_{0,j}$, so only the contemporaneous variables $X_{0,1}, \ldots, X_{0,d}$ can be intervened and appear in interventional targets. Meanwhile, we mainly consider imperfect (or soft, parametric) interventions, and the setting of perfect interventions (or hard, structural) [8] is a special case that can be easily extended.

3.2 The score for imperfect interventions

3.2.1 Model conditional densities. To begin with, we use neural networks to model conditional densities. Firstly, we encode the DAG $\mathcal{G}$ with a binary adjacency matrix $M^{\mathcal{G}} \in \{0, 1\}^{(p+1)d \times (p+1)d}$ which acts as a mask on the neural network inputs. Similarly, we encode the interventional family $\mathcal{I}$ with a binary matrix $R^{\mathcal{I}} \in \{0, 1\}^{Q \times (p+1)d}$, where $R_{qj}^{\mathcal{I}} = 1$ means that X_j is a target in I_q. Then, following equation (2), we further model the joint density of the qth intervention by

$$f^{(q)}\left(y; M^{\mathcal{G}}, R^{\mathcal{I}}, \phi\right) := \prod_{j=1}^{(p+1)d} \tilde{f}\left(x_j; \mathrm{NN}\left(M_j^{\mathcal{G}} \odot x; \phi_j^{(1)}\right)\right)^{1 - R_{qj}^{\mathcal{I}}} \tilde{f}\left(x_j; \mathrm{NN}\left(M_j^{\mathcal{G}} \odot x; \phi_j^{(q)}\right)\right)^{R_{qj}^{\mathcal{I}}}, \tag{3}$$

where $\phi := \{\phi^{(1)}, \ldots, \phi^{(Q)}\}$, the NN's are neural networks parameterized by $\phi_j^{(1)}$ or $\phi_j^{(q)}$, the operator $\odot$ denotes the Hadamard product (element-wise) and $M_j^{\mathcal{G}}$ denotes the jth column of $M^{\mathcal{G}}$, which enables selecting the parents of node j in the graph $\mathcal{G}$.

3.2.2 Maximize the score. Finally, we form the regularized maximum log-likelihood score:

$$\mathcal{S}_{\mathcal{I}^*}(\mathcal{G}) := \sup_{\phi} \sum_{q=1}^{Q} \mathbb{E}_{Y \sim p^{(q)}} \log f^{(q)}\left(Y, M^{\mathcal{G}}, R^{\mathcal{I}^*}, \phi\right) - \lambda |\mathcal{G}|, \tag{4}$$

where the ground truth interventional family $\mathcal{I}^* := (I_1^*, \ldots, I_Q^*)$ is known and $p^{(q)}$ stands for the qth ground truth interventional distribution. By maximizing it, we can get an estimated DAG $\hat{\mathcal{G}}$ that is $\mathcal{I}^*$-Markov equivalent to the true DAG $\mathcal{G}^*$ [3]. Then, we take $M^{\mathcal{G}}$ as a random matrix, where $M_{ij}^{\mathcal{G}} \sim B(1, \sigma(\alpha_{ij}))$, σ is the sigmoid function and α_{ij} is a scalar parameter. We group these α_{ij}s into a matrix $\Lambda \in \mathbb{R}^{(p+1)d \times (p+1)d}$. After that, we rely on augmented Lagrangian procedure [22] to maximize the following score:

$$\hat{\mathcal{S}}_{\mathcal{I}^*}(\Lambda) := \sup_{\phi} \mathop{\mathbb{E}}_{M \sim \sigma(\Lambda)} \left[\sum_{q=1}^{Q} \mathop{\mathbb{E}}_{Y \sim p^{(q)}} \log f^{(q)} \left(Y; M, R^{\mathcal{I}^*}, \phi \right) - \lambda \|M\|_0 \right], \tag{5}$$

under the acyclicity constraint: $\sup_{\Lambda} \hat{\mathcal{S}}_{\mathcal{I}^*}(\Lambda)$, s.t. $\operatorname{Tr} e^{\sigma(\Lambda)} - d = 0$.

Since we only focus on influences on $X_1, \ldots, X_d$ from other variables, we set $\Lambda[:, d+1 : (p+1)d]$, i.e. the meaningless part $M^{\mathcal{G}}[:, d+1 : (p+1)d]$, to zero before training.

3.3 Extracting weight matrix

Eventually, we obtain the estimated full weight matrix $\hat{F} = \sigma(\Lambda)$ of graph $\mathcal{G}$. Then, we can extract the intra-slice matrix $\hat{W} = \hat{F}[1 : d, 1 : d]$ and inter-slice matrix $\hat{A}_k = \hat{F}[kd + 1 : (k+1)d, 1 : d]$ for each time lag $k = 1, \ldots, p$. They reflect causal relations of these d variables in both a contemporaneous and time-lagged manner.

4 EXPERIMENTS

4.1 Baselines

To evaluate the effectiveness of our method, we compare with the following models:

- **DYNOTEARS** [12]: As DYNOTEARS is a method focused on fitting the exact values of time series, it outputs quantitative weight values. We set the threshold value of each W and A to be 0.5.
- **PCMCI** [16]: We used the results of PCMCI with a significance level of 0.01.
- **NeuralGC** [17]: Since NeuralGC only learns contemporaneous relationships, we used W_{full}, representing the overall relations between variables, to compare our method with it. W_{full} is defined as:

$$W_{\text{full}}(i,j) = \begin{cases} 1, & W(i,j) + \sum_{k=1}^{p} A_k(i,j) > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Since the baseline models are unable to utilize interventional data, we have ensured a fair comparison to some extent by keeping the overall sample size consistent across all models, while using only observational data (or normal data in real datasets) for the baseline models, and both observational and interventional data (or abnormal data in real datasets) for the proposed method. Therefore, we validate that by making use of extra information from interventional data, our proposed method outperforms other models that only apply observational data.

4.2 On simulation datasets

4.2.1 Datasets. We generate temporal data in two steps:

- Sample intra DAG and inter DAG following the *Erdős-Rényi* scheme, then sample parameters in weighted adjacency matrix, where elements in intra-slice matrix W are uniformly from $[-1.0, -0.25] \cup [0.25, 1.0]$ and elements in inter-slice matrixes A_k are uniformly from $[-1.0\alpha, -0.25\alpha] \cup [0.25\alpha, 1.0\alpha]$, $\alpha = 1/\eta^k, \eta \geq 1, k = 1, \ldots, p$.
- Generate time series consistent with the sampled weighted graph following the standard structural vector autoregressive (SVAR) model[15]: $Y_0 = Y_0 W + Y_1 A_1 + \cdots + Y_p A_p + Z$, where Z is random variables under the normal distribution. Then, sample interventional targets from nodes in Y_0, and generate imperfect interventional data by adding a random vector of $\mathcal{U}([-0.5, -0.25] \cup [0.25, 0.5])$ to W_{ij} and A_{kij}, where x_j is the variable in interventional targets and $x_i \in x_{\pi_j^{\mathcal{G}}}$. Besides, if $W_{ij}, A_{kij} > 0$, add the change to it; if $W_{ij}, A_{kij} < 0$, minus the change to it.

Before training, all data are normalized by subtracting the mean and dividing by the standard deviation. We experimented on two simulated datasets: Dataset 1 contains 5 nodes, their 1 time-lagged variables and 5 different interventional targets, each of which covers a single different node. Dataset 2 contains 10 nodes, their 1 time-lagged variables and 10 different interventional targets, each of which covers a single different node.

4.2.2 Evaluation metrics. We leverage the following two main metrics to evaluate the performance of the proposed method on learning causal graph: i) the structural Hamming distance (SHD), which calculates the number of different edges (either reversed, missing or redundant) between two DAGs; ii) the structural interventional distance (SID), which represents the difference between two DAGs according to their causal inference conditions [14].

4.2.3 Results. The results on simulation data are reported in Table 1. We can find that on both datasets, our method achieves significantly better results than baseline models on all four metrics: SHD, SID of the overall structure, and SHD of W (intra-slice structure) and A (inter-slice structure). Figure 1 shows results on dataset 1.

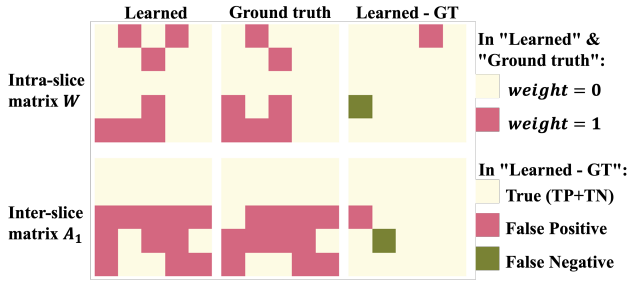

Figure 1: A show case of the result on simulation data.

4.3 On real datasets

4.3.1 Datasets. **Scence description.** In modern data centers, stable IT equipment operation is crucial. Advanced air conditioning systems are used to regulate the heat generated by the equipment and maintain a stable indoor temperature. In a typical data center

Table 1: Results on Simulation Data

Method	Dataset 1: 5 nodes, 1 lag				Dataset 2: 10 nodes, 1 lag			
	SHD ↓	SHD_W ↓	SHD_A ↓	SID ↓	SHD ↓	SHD_W ↓	SHD_A ↓	SID ↓
DYNOTEARS	20.40 ± 2.41	7.20 ± 1.55	13.20 ± 2.39	38.60 ± 3.72	36.00 ± 5.21	11.80 ± 3.36	24.20 ± 4.92	118.60 ± 20.66
PCMCI	18.10 ± 4.43	10.40 ± 2.17	7.40 ± 2.55	24.10 ± 3.11	62.00 ± 17.15	38.10 ± 11.16	23.90 ± 9.53	118.30 ± 24.08
Proposed	8.80 ± 3.65	3.20 ± 1.81	5.50 ± 2.01	12.00±8.21	27.80 ± 8.32	9.90 ± 3.67	17.70 ± 6.07	46.70 ± 23.12
NeuralGC	14.90 ± 2.28	-	-	20.00 ± 0.00	31.30 ± 4.72	-	-	85.50 ± 6.36
Proposed	4.80 ± 1.81	-	-	19.90 ± 0.32	21.50 ± 5.62	-	-	84.10 ± 6.51

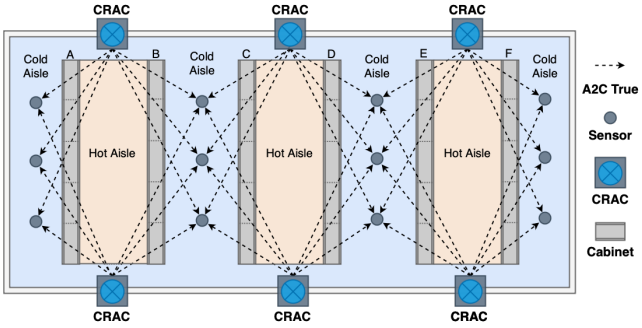

Figure 2: A typical data center cooling system diagram.

room, shown as figure 2, equipment rows (A, B, C, D, etc.) are arranged systematically in the room, with physical barriers isolating adjacent rows of cabinets to prevent hot and cold air from mixing. Computer room air conditioners (CRACs) supply cooling on both sides of the room, creating a closed loop structure that maintains a stable environment. Multiple sensors located in the cold aisle provide real-time temperature monitoring to ensure a continuous supply of cold air. This approach enables timely adjustments to maintain stable IT equipment operation.

Data acquisition. The data used in this study was obtained from a specific data center at Alibaba. It covers monitoring data from a cooling system of a particular room from January 1st, 2023 to May 1st, 2023, and includes 38 variables in total. These variables comprise *18 cold aisle temperatures* from sensors and *20 air conditioning supply temperatures* from CRACs. We collected several time series of these 38 variables during normal as well as abnormal states. For the latter, data was sampled within 20 minutes of the occurrence of the abnormality, with each sampling interval to be 10 seconds. Anomaly points were identified by learning the normal distribution range from historical data, using the n-σ method. Any data points that fall outside of the n-σ range (e.g., 3 to 5) of it selves are extracted as anomaly time points.

4.3.2 Evaluation metrics. Given the absence of ground truth DAGs in the real dataset, we employ two types of performance metrics based on prior knowledge to assess the algorithms' efficacy in learning causal graphs. These metrics mainly consider the physical location relationships within the room. Firstly, we reckon the edges from cold aisle temperatures (downstream) to air conditioning supply temperatures (upstream) is necessarily incorrect, and should be considered as a false negative. Therefore, we calculate i) **C2A False**, whether the model has learned those edges. Moreover, the closer the sensor is to a certain air conditioner, the greater the influence on this sensor. We assume that the edges from the air conditioning supply temperatures to the temperatures of the two adjacent cold aisles exist (shown in figure 2), and calculate ii) **A2C False**, the number of true edges the algorithm has not learned; and iii) **A2C True**, the number of true edges the algorithm has learned.

Table 2: Results on Real Data

	C2A False ↓	A2C False↓	A2C True↑
DYNOTEARS	0.00 ± 0.00	240.00 ± 0.00	0.00 ± 0.00
PCMCI	32.90 ± 5.00	226.80 ± 2.86	13.20 ± 2.86
Proposed	4.00 ± 6.85	234.20 ± 3.16	5.80 ± 3.16
NeuralGC	25.70 ± 19.47	118.10 ± 4.01	1.90 ± 4.01
Proposed	3.90 ± 6.61	114.70 ± 2.50	5.30 ± 2.50

4.3.3 Results. The results on real data are presented in Table 2. It can be observed that our method outperforms other approaches comprehensively, not only achieving fewer C2A False and A2C False but also learning a higher number of A2C True relations.

5 CONCLUSION

In conclusion, this paper addresses the limitations of temporal causal learning from observational data. The novel contributions include the introduction of a temporal interventional dataset with causal labels and the proposed TECDI approach. The experimental results demonstrate that incorporating interventional data effectively improves the accuracy of temporal causal discovery. In the future, we plan to apply the learned causal graph to practical tasks, such as root cause localization.

ACKNOWLEDGMENTS

This work was supported by Alibaba Group through Alibaba Innovative Research Program. This work was supported in part by the National Key Research and Development Program of China No. 2020AAA0106300, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia. Peiwen would like to thank Yang Li (Tsinghua-Berkeley Shenzhen Institute, Tsinghua University) for her support.

REFERENCES

[1] Raghavendra Addanki, Shiva Kasiviswanathan, Andrew McGregor, and Cameron Musco. 2020. Efficient intervention design for causal discovery with latents. In *International Conference on Machine Learning*. PMLR, 63–73.

[2] Til Ole Bergmann and Gesa Hartwigsen. 2021. Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of cognitive neuroscience* 33, 2 (2021), 195–225.

[3] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. 2020. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems* 33 (2020), 21865–21877.

[4] Frederick Eberhardt. 2012. Almost Optimal Intervention Sets for Causal Discovery. arXiv:1206.3250 [cs.AI]

[5] Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, and Yue Yu. 2022. IDYNO: Learning Nonparametric DAGs from Interventional Dynamic Data. In *International Conference on Machine Learning*. PMLR, 6988–7001.

[6] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[7] Shawkat Hammoudeh, Ahdi Noomen Ajmi, and Khaled Mokni. 2020. Relationship between green bonds and financial and environmental variables: A novel time-varying causality. *Energy Economics* 92 (2020), 104941.

[8] Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems* 33 (2020), 9551–9561.

[9] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can Large Language Models Infer Causation from Correlation? arXiv:2306.05836 [cs.CL]

[10] Yuan Meng, Shenglin Zhang, Yongqian Sun, Ruru Zhang, Zhilong Hu, Yiyin Zhang, Chenyang Jia, Zhaogang Wang, and Dan Pei. 2020. Localizing failure root causes in a microservice through causality inference. In *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*. IEEE, 1–10.

[11] Peer Nowack, Jakob Runge, Veronika Eyring, and Joanna D Haigh. 2020. Causal networks for climate model evaluation and constrained projections. *Nature communications* 11, 1 (2020), 1415.

[12] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1595–1605.

[13] Judea Pearl. 2009. *Causality*. Cambridge university press.

[14] Jonas Peters and Peter Bühlmann. 2015. Structural intervention distance for evaluating causal graphs. *Neural computation* 27, 3 (2015), 771–799.

[15] Juan F Rubio-Ramirez, Daniel F Waggoner, and Tao Zha. 2010. Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies* 77, 2 (2010), 665–696.

[16] Jakob Runge. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 1388–1397.

[17] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. 2021. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2021), 4267–4279.

[18] Jin Tian and Judea Pearl. 2013. Causal discovery from changes. *arXiv preprint arXiv:1301.2312* (2013).

[19] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'Ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.* 55, 4, Article 82 (nov 2022), 36 pages. https://doi.org/10.1145/3527154

[20] Wenzhuo Yang, Kun Zhang, and Steven C. H. Hoi. 2022. A Causal Approach to Detecting Multivariate Time-series Anomalies and Root Causes. arXiv:2206.15033 [cs.LG]

[21] Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding Causality with Large Language Models: Feasibility and Opportunities. arXiv:2304.05524 [cs.LG]

[22] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems* 31 (2018).