Bridging Causal Theory and LLMs: A User-Centric Approach to Causal Graph Generation

Daniele Potertì

Università degli Studi di Milano-Bicocca, Department of Economics, Management and Statistics (DEMS), Italy

Abstract

Causality plays a significant role in understanding cause-and-effect relationships, particularly in the context of data-driven decision-making. This research introduces an approach to causal graph generation that combines the power of Large Language Models (LLMs) with classical algorithms from causal theory. In contrast to recent related works that seek to exclude the involvement of domain experts, our method places them at the forefront of the process. We present a novel pipeline that streamlines and enhances expert validation by providing robust causal graph proposals. These proposals are enriched with transparent reports that blend foundational causal theory reasoning with explanations from LLMs. This user-centric approach aims to foster a more comprehensive and reliable causal graph generation process.

Keywords

Causal Discovery, LLMs, Human-AI-Interaction

1. Introduction

In our data-driven world, understanding causality is crucial for informed decision-making. Causal inference, the practice of determining the cause-and-effect relationship between variables [1], and causal discovery, the identification of these relationships from observational data [2], have taken center stage.

Direct Acyclic Graphs (DAGs) are essential in modeling causal relationships and forming the basis of structural causal models [1]. Representing variables as nodes and causal relationships as edges, preclude feedback loops and distinctly illustrate causal effects. Leveraging DAGs facilitates the transparent expression of causal assumptions, the identification of confounding variables, and the creation of accurate and interpretable models. This underscores their essential role as a foundation for constructing decision models and delivering explanations [3].

Modeling causal knowledge is complex and challenging since it requires an actual understanding of the relations, beyond statistical evidence. Yet, a critical ingredient in this quest for causality is domain knowledge [1]. Nevertheless, in numerous real-world scenarios, the collaboration between non-technical domain experts and computer scientists in defining a causal graph can be a challenging and time-consuming process [4].

AIxIA 2023, 22nd International Conference of the Italian Association for Artificial Intelligence, Roma Tre University - ICITA Department, Via Vito Volterra, 62, 00146 Roma, 6 - 9 Nov, 2023
d.poterti@campus.unimib.it (D. Potertì)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3670/paper84.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

LLMs, such as GPT-3.5 and GPT-4, show promise in bridging the gap between computational efficiency and domain expertise [5]. They excel in causal reasoning tasks, translating between natural language and formal methods, generating causal graphs, and identifying background causal contexts [6]. LLMs can act as proxies for human domain knowledge, enhancing causal analysis while reducing human effort [7]. Their prowess, mainly when used alongside existing causal statistical methods [8] holds the potential to revolutionize how we approach and understand causality. However, despite their benefits, LLMs may misinterpret complex causal models without deep domain expertise, leading to erroneous decisions [1]. Deep domain knowledge remains crucial to ensure the validity of causal relationships in fields like epidemiology, economics, and social sciences, even as systems like LLMs can play a more significant role in causal discovery.

The central goal of this research is to integrate LLMs with causal discovery techniques to streamline the final human validation. In this validation process, domain experts are relieved from the task of discovering causal relationships and can focus solely on validating the proposed causal graph. This approach aims to save time and effort, ultimately making the validation process more efficient and accessible.

Contribution. We present a first attempt at formalizing a causal pipeline to design a causal graph requiring minimal domain knowledge, that seamlessly combines data-driven statistical causality techniques with the insights of LLMs, acting as proxies for domain expertise. Moreover, this comprehensive framework has been encapsulated into an open-source Python package — that will be available in open source — designed to ease its integration into real-world scenarios.

As a contribution, this framework aims to be domain-specific, statistically robust, transparent, and explainable to ensure trust and effective validation by the human-in-the-loop of the generated causal graph. More in particular:

- **domain-specific**: among the results generated by the framework is a set of probable DAGs that optimally depict the causal relationships derived from the given data. Notably, these graphs are tailored not just to the data, but also to the specific domain, thanks to the integration of the LLMs with the Causal Discovery theory.
- **statistically robust**: the framework includes a final statistical sensitivity assessment for each DAG. This assessment evaluates the causal effect of each edge in accordance with the literature on causal inference, ensuring the robustness of the results.
- **transparent and explainable**: for each DAG, a comprehensive report resulting from rigorous causal theory testing, along with explanations provided by LLMs, either reinforces or questions the presence of particular edges and directional relationships within the graph.

The process concludes with the domain expert making an informed decision to select the preferred causal graph from among the proposed options. In our implementation, we place significant emphasis on the aspect of prompt engineering within the overall processing of the proposed causal pipeline.

2. Related Works

In this section, we explore recent developments in the literature on LLMs that (i) focus on their capabilities in the realm of causal reasoning, and (ii) investigate strategies for integrating LLMs with causal discovery techniques.

Exploring the Causal Reasoning Abilities of LLMs. Recent advancements indicate that models like ChatGPT are progressing towards artificial general intelligence (AGI), notably enhancing causal reasoning and high-precision tasks [9]. There is the potential for a paradigm shift in machine learning that could harmonize the strengths of AI with human capabilities, leading to transformative solutions and enhancing human decision-making [10]. Kıcıman et al. [5] further explores LLMs' capabilities in causal reasoning, illustrating their prowess in tasks like code generation and complex reasoning. These models demonstrate high performance in causal discovery, achieving up to 97% accuracy on the Tubingen benchmark and showcasing versatility across different domains. Despite this, they can occasionally falter in basic logic tasks, raising reliability concerns. Kıcıman et al. [5] emphasize the importance of incorporating human domain knowledge into causal analysis, suggesting that LLMs could serve as powerful tools to enhance this process through dynamic conversational interfaces. Yet, it is vital to remain cautiously optimistic about their capabilities due to potential erratic performances. Future research is poised to delve deeper into the capacities and boundaries of LLMs in this domain.

Integrating LLMs in Causal Discovery. Long et al. [11] introduced a novel approach to the challenges of causal discovery by formalizing the use of imperfect experts as an optimization problem, aiming to minimize the Markov equivalence class (MEC) size while ensuring the true graph remains included. They proposed a greedy approach reliant on Bayesian inference to achieve this, incrementally integrating expert knowledge. Empirical evaluations on real data revealed the effectiveness of their method, especially when the expert consistently provided correct orientations. However, when using LLMs as the experts, the results were mixed, suggesting both the potential and the challenges of integrating LLMs into causal discovery. Ban et al. [12] explore the role of LLMs in Causal Structure Learning (CSL), focusing on utilizing LLMs to pinpoint direct causal relations in observed data. This two-stage framework first uses LLMs to identify potential causal connections based on textual data and then applies these insights as constraints in data-driven CSL algorithms. The aim is to merge the intuitive causal understanding of LLMs with the detailed causal analysis found in CSL, potentially increasing its efficiency and accuracy. However, the study acknowledges the potential errors in the causal statements generated by GPT-4, indicating room for future improvements in both understanding and quality of causal relationships.

In contrast to existing approaches, our work distinguishes itself by prioritizing human validation. Unlike conventional methods that aim to autonomously generate causal graphs, we actively engage domain experts. Our pipeline simplifies their tasks with robust causal graph proposals, accompanied by a transparent report featuring both causal theory reasoning and LLM-based explanations.

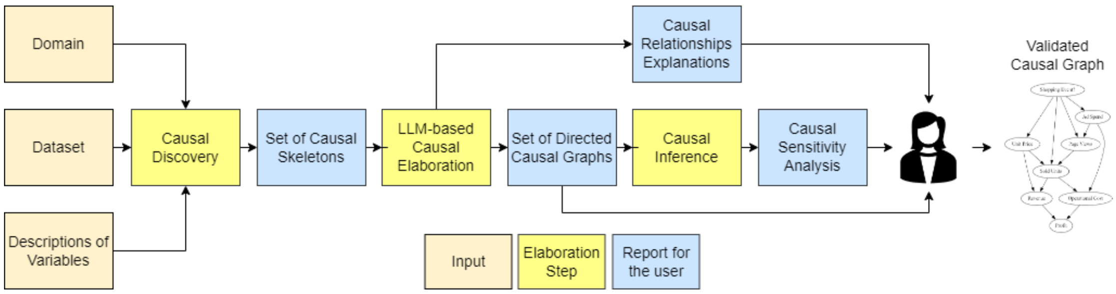

Figure 1: The proposed pipeline.

3. The causal pipeline with LLM and *Human-in-the loop*

Our proposed framework endeavors to deliver DAGs tailored to specific domains, leveraging an explanation-centric approach and undergoing rigorous statistical testing. This framework progresses through three essential phases: Causal Discovery, LLM-based Causal Elaboration, and Causal Inference, as illustrated in Figure 1. Notably, human involvement remains a crucial element throughout this process. This framework is designed to guide users in selecting the appropriate Causal Graph, drawing from the Set of Directed Causal Graphs, Causal Relationship Explanations, and Causal Sensitivity Analysis.

Causal Discovery Step. The causal discovery step begins with a dataset containing the variables of interest and their observed interactions. We process this dataset using four distinct causal discovery algorithms: PC (a constraint-based method) [2], GES (a score-based method) [13], FCI (an extension of the PC algorithm) [2], and LiNGAM (a functional causal model) [14]. Each of these algorithms represents a distinct method within causal discovery. The output of this process is a set of four Causal Skeletons, each one derived from a causal discovery algorithm. The LLM will further process each skeleton in the LLM-based Causal Elaboration Step.

LLM-based Causal Elaboration Step. Once we retrieve the sets of Causal Skeletons from the previous step it is time to determine the direction of the relationships performing the task of Pairwise Causal Discovery [15]. In this task, the goal is to determine the causal relationship between two variables. In this stage of the framework, we undertake two primary tasks: initially, we delve into understanding the inherent nature of the relationships depicted in each Causal Skeleton, and subsequently, we explore potential additional relationships through conditional independence tests. One of the fundamental challenges arises from the presumption that the LLM can infer the domain context purely based on variable values[1]. In this stage, the task's success hinges largely on the quality of the prompt. We are in the process of creating a novel prompt that aligns with causal theory and maintains consistent performance across various LLMs, including GPT-3.5, GPT-4, and LLaMA2. The outcome of this stage is a set of DAG: a

[1] For the LLM to make meaningful inferences and not merely fabricate a domain, the expert needs to provide at least some foundational information about the domain.

directional graph where the sequence of cause-and-effect relations is so structured that it never loops back on itself. Furthermore, in this stage, we generate explanations using carefully crafted prompts to justify the outputs of the LLMs.

Causal Inference Step. The last phase of our pipeline is designed for the statistical validation of the obtained DAGs from the preceding step. In pursuit of this objective, we adhere to the four key steps of causal inference as outlined in Pearl et al. [1]. To facilitate this process, we leverage the capabilities of DoWhy, a widely recognized open-source Python library. What distinguishes DoWhy is its strong foundation in causal assumptions, firmly rooted in the well-established framework of causal graphs [16].

After the pipeline, we generate a transparent report for the user. This report includes all the proposed DAGs and the rationale behind their generation. It encompasses the results of causal discovery tests, LLMs prompt outcomes, and causal inference, empowering users to make an informed choice regarding their preferred causal graph.

4. Conclusion and Next Step

We are currently developing a pipeline for constructing a causal graph that effectively combines well-established causal theory algorithms from the literature with LLMs. While related works focus on generating causal graphs without the involvement of domain experts, our approach places human validation at the core of the process. Our pipeline is designed to streamline and enhance the expert's work by providing robust causal graph proposals. It accompanies these proposals with a transparent report that includes the underlying causal theory reasoning and explanations derived from LLMs. This approach enables informed decision-making and sets us apart from existing methodologies. Our next steps involve refining the prompts and releasing the code as open-source. Additionally, we are planning to apply this process to real-world cases and conduct user evaluations. These steps will help us gain a better understanding of the domain expert's requirements and foster trust in our approach.

References

[1] J. Pearl, et al., Models, reasoning and inference, Cambridge, UK: CambridgeUniversityPress 19 (2000) 3.

[2] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, MIT press, 2000.

[3] R. Crupi, A. Castelnovo, D. Regoli, B. San Miguel Gonzalez, Counterfactual explanations as interventions in latent space, Data Mining and Knowledge Discovery (2022) 1–37.

[4] X. Xie, F. Du, Y. Wu, A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications, IEEE Transactions on Visualization and Computer Graphics 27 (2020) 1448–1458.

[5] E. Kıcıman, R. Ness, A. Sharma, C. Tan, Causal reasoning and large language models: Opening a new frontier for causality, arXiv preprint arXiv:2305.00050 (2023).

[6] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, Sparks of artificial general intelligence: Early experiments with gpt-4, preprint arXiv:2303.12712 (2023).
[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog (2019).
[8] M. Hernán, J. Robins, Causal Inference: What If, Chapman & Hall/CRC, Boca Raton, 2020.
[9] Z. C. et al., Understanding causality with large language models: Feasibility and opportunities, arXiv preprint arXiv:2304.05524 (2023).
[10] C. T. Wolf, Reprogramming the american dream: from rural america to silicon valley—making ai serve us all by kevin scott and greg shaw, Information & Culture 56 (2021) 113–114.
[11] S. Long, A. Piché, V. Zantedeschi, T. Schuster, A. Drouin, Causal discovery with language models as imperfect experts, arXiv preprint arXiv:2307.02390 (2023).
[12] T. Ban, L. Chen, X. Wang, H. Chen, From query tools to causal architects: Harnessing large language models for advanced causal discovery from data, arXiv preprint arXiv:2306.16902 (2023).
[13] D. M. Chickering, Optimal structure identification with greedy search, Journal of machine learning research 3 (2002) 507–554.
[14] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, M. Jordan, A linear non-gaussian acyclic model for causal discovery., Journal of Machine Learning Research 7 (2006).
[15] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Schölkopf, Distinguishing cause from effect using observational data: methods and benchmarks, The Journal of Machine Learning Research 17 (2016) 1103–1204.
[16] A. Sharma, E. Kiciman, Dowhy: An end-to-end library for causal inference, arXiv preprint arXiv:2011.04216 (2020).