Visual Causality Analysis

Jun Wang*

Stony Brook University

Abstract

In my doctoral study, I investigate a visual analytic approach for causal inference. The goal of my research is a visual interface that leverages inference algorithms but allows human experts endowed with domain knowledge and intuition to refute or propose causal links. In this research overview, I provide an introduction to my research topic, outline the challenges, present a timeline for my research progress, and discuss potential future works.

1 Introduction

Recognizing the exact causal relations governing the observed phenomena is a fundamental task in science. However, even with the emergence of big data, this task remains challenging because it requires a fundamental theory of how and why the observed phenomena occur. While a number of causal inference algorithms have been devised by modern philosophers and statisticians for identifying casual relations in multivariate datasets, these algorithms typically cannot encode existing domain knowledge, or even common sense, to guide their analyses. This in turn leads them to hold strong assumptions on data conditions which can rarely be satisfied in practice. A remedy to overcome this significant shortcoming is to insert a human into the casual inference loop as a synergist partner with a visual interface. Such a visual analytic approach is named *Visual Causality Analysis* and is the major focus of my doctoral research.

The pipeline of the visual analytics on causality established in my research is illustrated in Fig. 1. The process starts from deriving causal dependencies from observational data using automatic algorithms (Fig. 1-1), and then parameterizing the causal structure with statistical metrics (Fig. 1-2). The set of causal relations among variables of a multivariate dataset is usually represented as Directed Acyclic Graphs (DAGs) called *Causal Graphs/Networks*, where nodes stand for variables and edges denote causal relations pointing from the cause to the effect. A domain user can be involved by interacting with the input data to select interesting data block (Fig. 1-3), generating hypothesis by proposing and refuting causal relations to change the DAG structure (Fig. 1-4), and validating the hypothesis by observing the evolvement of the parameterized model (Fig. 1-5) long with each operation. The primary goal of my research is to develop a comprehensive visual analytic framework to fulfil this pipeline and can be generally applied to a wide scope of data.

Besides an in-depth understanding of the modern causal inference theory, several practical obstacles must be tackled in achieving such a visual analytic framework. These practical issues include an effective visualization of the causal graph, methods for handling heterogenous data in inference, and the capability to infer and manage multiple causal models underlying different ranges of data. During my doctoral research, preliminary solutions for these issues have been proposed, based on which a prototype of the visual interface for causality analysis is developed. In this paper, I shall provide a brief overview of my research work, as well as propose some potential future development for discussion.

In the remainder of this research overview, I will first give a brief introduction of the background knowledge and the prior work most relevant for my research in Section 2. Then I will outline the main challenges in Section 3. Section 4 will provide a timeline of accomplished and planned projects and research. Outlook of my future research will be discussed in Section 5. And Section 6 closes with conclusion.

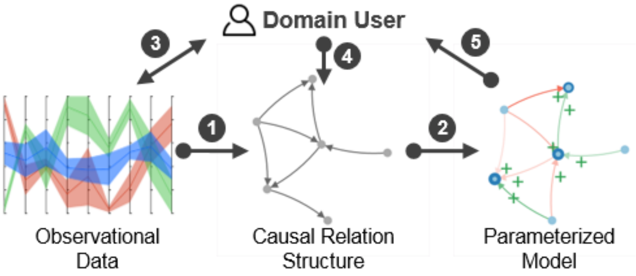


Figure 1: Pipeline of visual causality analysis.

2 Background and Related Work

My research builds on previous work of causation modeling and inference theories, as well as visual analytics research on causality analysis interfaces.

2.1 Causality Modeling and Inference

At the most basic, a causal relation is defined as a counter-fact, so that "A causes B" implies "if A did not happen B would not happen". Following the seminal work of Pearl [1] and Spirtes[2], theories of causality modeling and discovery on multivariate datasets have been widely studied. As mentioned, causal relations of a dataset can be depicted as causal graphs. A causal graph can be parameterized by several causal modeling methods. The two most common choices are Bayesian Networks [1] and Structural Causal Models (SCM) [3][4]. The former quantifies causal relations with conditional probability tables, and the latter with linear functions, e.g. linear regressions and logistic regressions.

Algorithms learning the structure of causal DAGs can be roughly categorized into two classes – score-based algorithms and constraint-based algorithms. The former typically associate a DAG with a score function, e.g. Bayesian Information Criterion [5][6], and performs, for instance, a greedy search in the space of all possible DAGs. Examples are the algorithms of GES [7] and K2 [8]. Since the number of possible structures is super-exponential in the number of variables, such algorithms often suffer from high search cost. In contrast, the constraint-based algorithms build the causal networks according to the constraints of dependencies and conditional dependencies in the data. These constraints are usually learned with conditional independence (CI) tests via partial correlation [9] or G^2 statistics [10]. Some well-known algorithms are SGS [2], PC [2][11], IC [12], Total Conditioning (TC) [13], and

* junwang2@cs.stonybrook.edu

others, differing in the ways the CI tests are arranged so that the algorithm can be more efficient.

As the knowledge of data distribution required in BN is usually hard to acquire in practice, the SCM is adopted in my work. Also, considering both accuracy and efficiency, the algorithms of PC and TC are used to infer the causal graph structure. However, it is important to note that these automated algorithms are commonly based on several strong assumptions of data that may hardly be satisfied by real-world applications. For example, they typically assume that the observed data is sufficient to recover all true causal relations (*Faithfulness*), and the set of measured variables contains all common causes of variable pairs in the set (*Causal Sufficiency*). When these assumptions are violated, false relations can be introduced and true relations may be missed. As a consequence, none algorithm can guarantee an exact model, which makes human involvement necessary.

2.2 Visual Analytics of Causality

Visual analytics of causality has become a popular topic in recent years, aiming to provide decision support in certain organization and to aid hypothesis generation and evaluation in a scientific investigation [14]. One of the earliest example is the *Growing-polygons* [15] scheme which captures causation in the *process* level, i.e. as a sequence of causal events. It uses animated polygon colors and sizes to signify causal semantics. The work of Vigueras and Botia [16] considers ordered events in a distributed system as causations and visualizes their dependencies as causal graphs. Focusing on the upstream-downstream relations of variables, *ReactFlow* [17] visualizes causal relations as pairwise pathways connecting duplicated variables in two columns. Some other efforts in visual mining of causation include *OutFlow* [18] and *EventFlow* [19]. Both visualize event sequences as alternative pathways in a temporal order and use event chains to explore embedded patterns. Liu et al. [20] visualize event streams as flows aligned by event types. However, none of these above systems leverages automated algorithms for causal discovery, thus they all require significant user input to acquire such knowledge.

3 Challenges

Besides a thorough understanding of the contemporary causality analysis theory, several practical difficulties must be tackled before achieving an effective visual analytics approach to causal inference, some of which are outlined as following.

3.1 Causal Graph Visualization

While a number of layouts are available for visualizing DAGs, the goal here is that the story of causal dependencies can be easily recognized by users. Although the widely-used force-directed approaches could be a feasible choice for demonstrating the overall structure of the graph, they often suffer from a dense and unpredictable layout. With such layouts, local structures in causal sequences can become difficult to observe especially when they are part of more complex graphs. However, these local structures can often be of great interest to domain users.

What's more, semantics of causal relations and their statistic measures need to be well encoded, so that users can get an intuitive understanding of the causal effects. As SCM is adopted, strength and significance of each causal relation as well as the goodness of fit of the whole model can be measured by regression coefficients and statistics.

3.2 Visual Model Refinement

When the decision of refuting or accepting certain causal relations cannot be made with users' domain knowledge, a scoring strategy that can be applied to each causal relation as well as the overall model is demanded so that the alternative models with or without certain relations can be quantitatively compared. Although common statistics calculated from regression residuals, e.g. F-statistics and r-squared, are capable to measure the model goodness of fit, they usually do not take model complexity into consideration. This implies that just by adding more relations into the model these statistics will mostly improve. However, this can potentially lead to overfitting, which means that the model is an extremely good fit for the dataset from which it was learned, but generates huge errors on any other dataset recorded from the same source. Hence, based on William of Occam's parsimony principle, models should be kept as simple as possible. The idea is that by adding new relations to a causal model we obtain an improvement in its fit to the data to some degree, but at the same time the model also becomes "worse" because it will be harder to fit to new data. So, the challenge is to select a scoring strategy respecting both the goodness of fit and the model complexity. A visual analytics scheme is also required so that models can be compared visually.

3.3 Handling Heterogeneous Data

Real-world problems often have a mix of numeric and categorical (ordinal, nominal) data. This stands at odds with current causal inference algorithms which can only handle either numeric or categorical variables, but not both. Simply binning all numeric variables and applying G^2 test can be a plausible solution. However, with this approach, not only is there a loss in variable value scales, but also the order of bins will be ignored in the G^2 tests, both of which can lead to huge decrease in result accuracy.

Another possible solution is to go the other way, which means to re-space and re-order the levels of categorical variables so that they can be treated as numeric ones and partial correlation can be used to do CI tests. Such a strategy has been proposed by Zhang et al. [21] applied in correlations studies, which, for each pair of categorical and numeric variables, reorders and repositions the levels of the categorical variable such that Pearson's correlation between the pair is optimized. Potentially, this can be extended and applied in causality analysis. The challenges are that, in a causality context, the mapped values (1) must be consistent regarding all other variables and (2) better be in a continuous space as this is assumed in partial correlation based CI tests.

3.4 Causal Models from Data Subdivisions

Another practical challenge is posed by *Simpson's Paradox* [22], which states that a relation found in the overall data may not hold in certain data subdivisions, and conflicting relations buried in some specific data ranges may cancel each other so that none can be observed in the general population. For example, by bracketing the price of a product to lower ranges one may see positive correlations with sales, while negative correlations come with a higher price range. What's more, causal relations with opposite directions may also exist as feedback loops. For instance, the price of a product affects its sales when the sales are low, but a large number of sales can also reduce the cost and so lower the price. As a result, it is often the case that multiple causal models differing in both structure and regression parameters can arise from data partitions. Ignoring such facts and always learning the model using the whole dataset will potentially lead to faulty relations returned by inference algorithms. It would be of great help if such disturbances can be reduced and different causal models hiding in the data can be revealed.

What's more, diagnosing these models by investigating their similarities can often reveal interesting knowledge, especially when the data is partitioned into a large number of subsets and a corresponding number of models are learned.

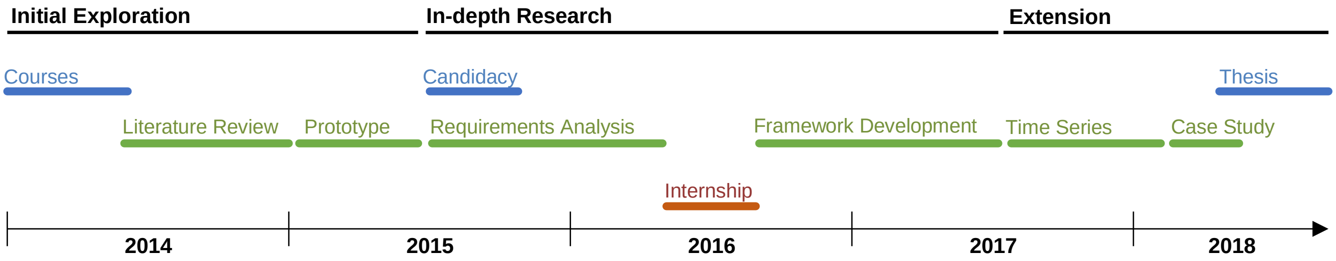


Figure 2: My doctor research can be divided into three main phases: Initial Exploration, In-depth Research, and Extension

4 Research Progress

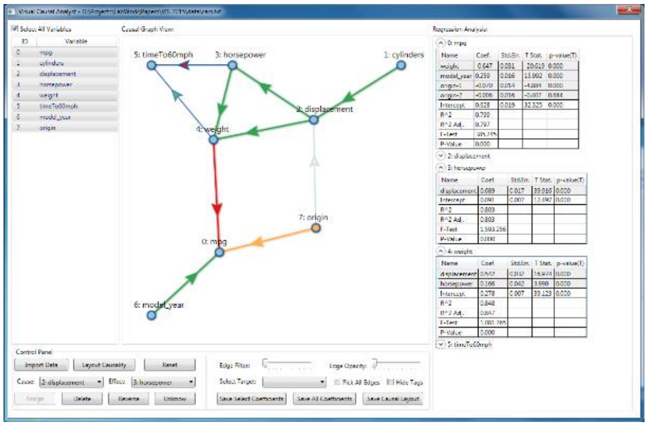
My doctoral research has been progressing along with identifying and tackling the challenges of visual causality analysis, and can be roughly divided into three phases (Fig. 2).

4.1 Initial Exploration (18 months)

At the beginning of my doctoral study, I focused on building the foundation for my research by conducting a comprehensive literature review on causal inference theories and related works in the field of visual analytics. The biggest challenge to me in this phase was to gain an accurate understanding of what causality is and how it is modeled and inferred statistically. With the accumulated theoretical knowledge, I selectively implemented the TC [13] and the PC-stable [11] algorithms, which have been used as part of my toolbox throughout my research.

What's more, based on initial exploration on required analysis of causality and referencing the existing works, a prototype of an interface was proposed, named *Visual Causality Analyst* [23](Fig. 3). The interface visualizes a causal network as an interactive force-directed 2D graph, where the type of a causal relation (positive, negative, or compound, measured by regression coefficients) is encoded as edge colors. The interface supports very basic interaction of creating, deleting, directing, and reversing causal links, as well as filtering edges by their strength. Other detailed regression parameters are listed as tables.

Along with the interface, a new strategy of mapping categorical variables' values regarding all numeric variables was proposed (and was latterly named Global Mapping, GM), so that heterogeneous data can be analyzed with the inference algorithm as well as the new interface. The framework was tested on a few illustrative datasets, e.g. the AutoMPG dataset [24].



ID	Variable
0	mpg
1	cylinders
2	displacement
3	horsepower
4	weight
5	timeTo60mph
6	model_year
7	origin

0: mpg

Name	Coef.	Std.Err.	T Stat.	p-value(T)
weight	0.647	0.031	20.619	0.000
model_year	0.259	0.016	15.992	0.000
origin=1	[illegible]	0.014	-4.284	0.000
origin=2	-0.006	0.016	-0.407	0.684
Intercept	0.628	0.019	32.325	0.000
R^2	0.799			
R^2 Adj.	0.797			
F-Test	[illegible]			
P-Value	0.000			

2: displacement

3: horsepower

Name	Coef.	Std.Err.	T Stat.	p-value(T)
displacement	0.689	0.017	39.916	0.000
Intercept	0.091	0.007	12.497	0.000
R^2	0.803			
R^2 Adj.	0.803			
F-Test	1,593.256			
P-Value	0.000			

4: weight

Name	Coef.	Std.Err.	T Stat.	p-value(T)
displacement	[illegible]	[illegible]	[illegible]	0.000
horsepower	0.166	0.042	3.990	0.000
Intercept	0.278	0.007	39.123	0.000
R^2	0.848			
R^2 Adj.	0.847			
F-Test	[illegible]			
P-Value	0.000			

5: timeTo60mph

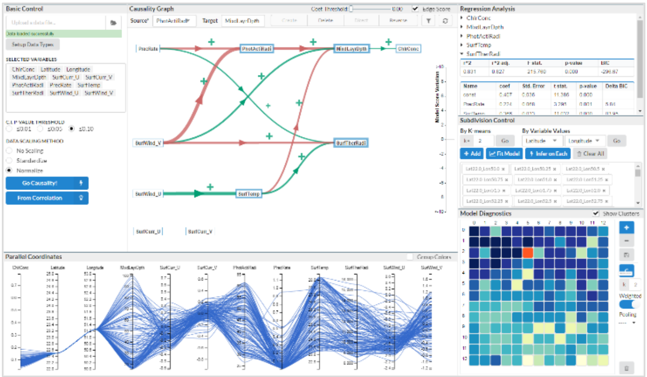
Figure 3: The Visual Causality Analyst visualizing the causal graph derived from the AutoMPG dataset [24].

4.2 In-depth Research (24 months)

While effective, the proposed prototype is nevertheless relatively too simple. Real world scenarios, however, incur many practical difficulties that such a simple tool cannot handle. To devise a comprehensive visual analytic framework that can fulfil the visual causality analysis pipeline, an in-depth research on required analysis was conducted. This was carried out with further review of literatures on causality visual analytics and case studies with more complex real-world datasets provided by collaborating scientists and online resources.

During this period, the challenges in section 3 were outlined and established as the requirements of the new framework. To solve each of these problems, a new causal graph visualization is designed which renders a richer set of statistic parameters and exposes flow of causal sequences in a much more prominent way; the Bayesian Information Criterion is adopted to score each relation as well as the whole model so that model goodness of fit and model complexity are both considered when comparing alternative causal models; an improved version of GM is proposed and experimentally evaluated in causality contexts; strategies and interfaces for learning and diagnosing multiple models from data subdivisions are also designed and tested.

All these new techniques are implemented and integrated in a new visual interface named *Causal Structure Investigator* (CSI), which is demonstrated in Fig. 4. With the CSI interface, a user can observe potentially attractive data subdivisions with the parallel coordinates, and then partition the data by adjusting the brushed value range of variables or with clustering algorithms. Multiple models can be automatically inferred and labelled in the heatmap, with which the user can recognize credible causal models and extract reliable relations from all learned models.



r^2	r^2 adj.	F stat.	p-value	BIC
0.831	0.827	215.780	0.000	-296.87

Name	coef	Std. Error	t stat.	p-value	Delta BIC
const	0.407	0.036	11.386	0.000	
PrecRate	0.224	0.068	3.281	0.001	5.04

Figure 4: The interface of the Causal Structure Investigator, which is a comprehensive visual framework for causality analysis.

4.3 Extension (12 months)

During the final phase of my PhD study, I will further extend my research on visual causality analysis. A present limitation of my current work is that it does not support analysis on time series data, which would have many popular applications, such as finance, health, etc. A possible solution is to utilize the theory of logic-based causality, which can be capable to learn causes of certain events within time series.

I also plan to conduct more application studies applying my research on a wider scope of real-world datasets. I will continually deepen and renew my understanding and knowledge of causality theories as it is still an actively developing field. Finally, I shall summarize all my design experience and research findings into a conceptual and pragmatic framework of visual causality analysis which will become part of my thesis.

5 Future Work

Besides the planed work in my final phase of doctoral research, there are many future research on visual causality analysis I would like to do, as the topic is far from fully explored. For instance, one that might be easily done is a user study on causal graph visualization methods, as quite a few has been introduced and their ability in delivering causal semantics is still untested.

Another future work I would like to explore is to gain the ability to build causal models utilizing data from different measurements and sources but generated by the same causal mechanism, which is called the data fusion problem [25] or integrative causal analysis [26]. A visual interface supporting such analytical tasks would allow users to study scientific systems over a series of data collections.

Finally, causal graphs learned from causality analysis can serve as a starting point for prescriptive analytics. Automatic generation of such analytics, i.e. narration, is also a promising extension to the current work, where specific actions could be recommended given a user's request.

6 Conclusion

With this research, I intend to explore and broaden the topic of visual causality analysis. Based on a comprehensive literature review and case studies of several experimental datasets, a few specific challenges are identified and the corresponding visual analytic solutions are proposed. I will further extend the interface in developing and undertake several application studies to establish practical guidelines for designing and implementing such interfaces for visual causality analysis.

The participation in the doctoral colloquium would be a great opportunity for me to get feedback on my research work and plan, and to get suggestions on the selection of my case studies.

References

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. New York, NY: Springer New York, 1993.

[3] J. Pearl, 'An Introduction to Causal Inference', *Int. J. Biostat.*, vol. 6, no. 2, pp. 1–62, 2010.

[4] S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij, 'Structural Causal Models: Cycles, Marginalizations, Exogenous Reparametrizations and Reductions', *arXiv*, Nov. 2016.

[5] K. P. Burnham and R. P. Anderson, 'Multimodel Inference: Understanding AIC and BIC in Model Selection', *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[6] G. Schwarz, 'Estimating the Dimension of a Model', *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[7] D. M. Chickering, 'Optimal structure identification with greedy search', *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2002.

[8] G. Cooper and E. Herskovits, 'A Bayesian Method for the Induction of Probabilistic Networks from Data', vol. 347, pp. 309–347, 1992.

[9] K. Baba, R. Shibata, and M. Sibuya, 'Partial correlation and conditional correlation as measures of conditional independence', *Aust. New Zeal. J. Stat.*, vol. 46, no. 4, pp. 657–664, 2004.

[10] R. E. Neapolitan, 'Chapter 10.3.1', in *Learning Bayesian Networks*, Pearson, 2003, pp. 600–603.

[11] D. Colombo and M. H. Maathuis, 'Order-independent constraint-based causal structure learning', *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3741–3782, 2014.

[12] J. Pearl and T. S. Verma, 'A theory of inferred causation', *Stud. Log. Found. Math.*, vol. 134, pp. 789–811, 1995.

[13] J. P. Pellet and A. Elisseeff, 'Using Markov Blankets for Causal Structure Learning', *J. Mach. Learn. Res.*, vol. 9, pp. 1295–1342, 2008.

[14] M. Chen *et al.*, 'From Data Analysis and Visualization to Causality Discovery', *Computer (Long. Beach. Calif).*, vol. 44, no. 10, pp. 84–87, 2011.

[15] N. Elmqvist and P. Tsigas, 'Animated visualization of causal relations through growing 2D geometry', *Inf. Vis.*, vol. 3, no. 3, pp. 154–172, 2004.

[16] G. Vigueras and J. A. Botia, 'Tracking causality by visualization of multi-agent interactions using causality graphs', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, pp. 190–204.

[17] T. Dang, P. Murray, J. Aurisano, and A. Forbes, 'ReactionFlow: an interactive visualization tool for causality analysis in biological pathways', in *Proceedings of the 5th Symposium on Biological Data Visualization: Part 2*, 2015, vol. 9, no. Suppl 6.

[18] K. Wongsuphasawat and D. Gotz, 'Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization', *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2659–2668, 2012.

[19] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, 'Temporal event sequence simplification', *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2227–2236, 2013.

[20] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson, 'Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths', *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 321–330, 2017.

[21] Z. Zhang, K. T. Mcdonnell, E. Zadok, and K. Mueller, 'Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map', *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 2, pp. 289–303, 2015.

[22] E. H. Simpson, 'The Interpretation of Interaction in Contingency Tables', *Source J. R. Stat. Soc. Ser. B*, vol. 13, no. 2, pp. 238–241, 1951.

[23] J. Wang and K. Mueller, 'The Visual Causality Analyst: An Interactive Interface for Causal Reasoning', *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 230–239, 2016.

[24] K. Bache and M. Lichman, 'UCI Machine Learning Repository', *University of California, Irvine, School of Information*. 2013.

[25] E. Bareinboim and J. Pearl, 'Causal inference and the data-fusion problem', *Pnas*, vol. 113, no. 27, pp. 7345–7352, 2016.

[26] I. Tsamardinos, 'Advances in Integrative Causal Analysis', in *Proceedings of the UAI 2015 Conference on Advances in Causal Inference*, 2015, pp. 90–91.