This paper is © IEEE and appears reformatted in *IEEE Transactions on Visualization and Computer Graphics*, 30(1), 2016, DOI: 10.1109/TVCG.2015.2467931

The Visual Causality Analyst: An Interactive Interface for Causal Reasoning

Jun Wang and Klaus Mueller, *Senior Member, IEEE*

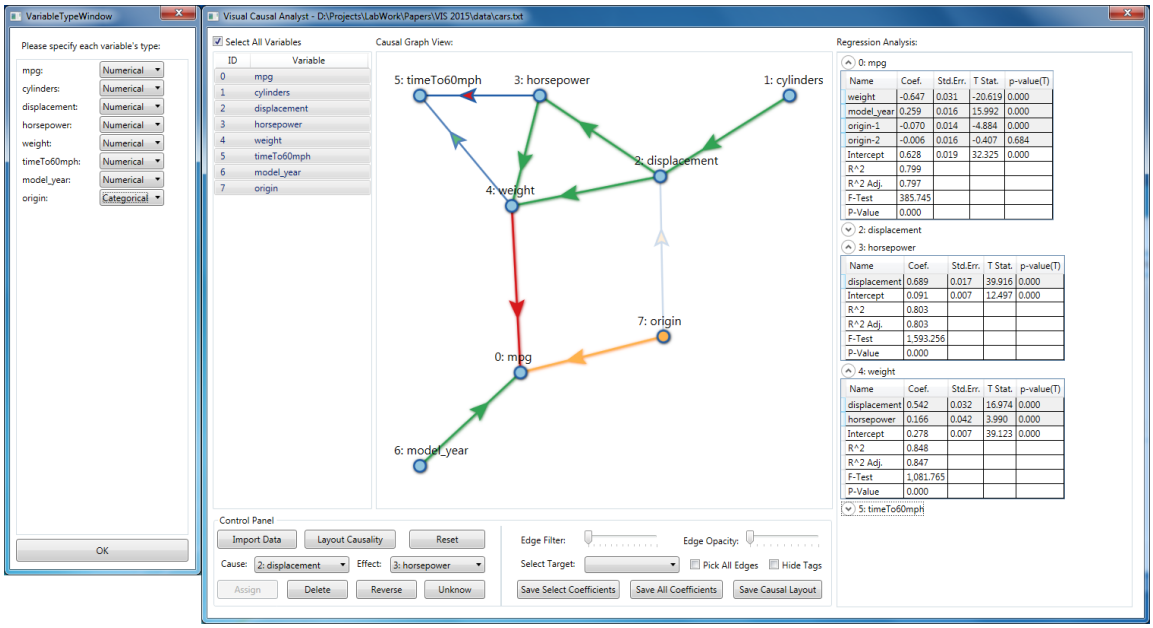


Fig. 1. An overview of the Visual Causality Analyst framework running on the auto MPG dataset [29].

Abstract—Uncovering the causal relations that exist among variables in multivariate datasets is one of the ultimate goals in data analytics. Causation is related to correlation but correlation does not imply causation. While a number of casual discovery algorithms have been devised that eliminate spurious correlations from a network, there are no guarantees that all of the inferred causations are indeed true. Hence, bringing a domain expert into the casual reasoning loop can be of great benefit in identifying erroneous casual relationships suggested by the discovery algorithm. To address this need we present the *Visual Causal Analyst* – a novel visual causal reasoning framework that allows users to apply their expertise, verify and edit causal links, and collaborate with the causal discovery algorithm to identify a valid causal network. Its interface consists of both an interactive 2D graph view and a numerical presentation of salient statistical parameters, such as regression coefficients, p-values, and others. Both help users in gaining a good understanding of the landscape of causal structures particularly when the number of variables is large. Our framework is also novel in that it can handle both numerical and categorical variables within one unified model and return plausible results. We demonstrate its use via a set of case studies using multiple practical datasets.

Index Terms—Visual knowledge discovery, Causality, Hypothesis testing, Visual evidence, High-dimensional data

1 Introduction

Recovering the causal relations from purely observational data is one of the ultimate goals for data analysts and a fundamental problem in science. After decades of efforts by many, causality research gained particularly strong attention when Judea Pearl, a long time pioneer of the field, won the Turing award in 2011 for the underlying mathematical framework of causal inference. The advantage of knowing the causal relations rather than just statistical associations, e. g., correlations, is that the former enables explicit guidance in predicting the effects of actions perturbing the observed system.

- *Jun Wang and Klaus Mueller are with the Visual Analytics and Imaging Lab at the Computer Science Department, Stony Brook University, Stony Brook, NY. E-mail: {junwang2, mueller}@cs.stonybrook.edu.*
- *Klaus Mueller is also with the Computer Science Dept. at SUNY Korea.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx xxx 2015; date of current version xx xxx 2015. For information on obtaining reprints of this article, send e-mail to: tvcg@computer.org.

The most reliable way to determine causation is by controlled experiments. However, controlled experiments are often either impossible or associated with high cost and thus impractical in real world. A loose detour usually taken is trying to express causation by correlations calculated from observational data. One typical example of this is the website Google Correlate [1] which can provide visitors with endless hours of entertainment by entering any search term and then browsing a long list of spurious correlations the term has with either time or US states. But while users of Google can easily tolerate the many irrelevant links the search engine typically generates, for other applications, such as healthcare diagnosis and financial prediction, blindly inferring causation from spurious correlations can have severe consequences.

To infer a precise model describing and measuring causal relations embedded in observational data, the theory of causal inference and analysis, started with the work of Pearl [2, 3], Spirts [4, 5], and others, has become a hot topic in recent years. While many

modern causal discovery algorithms claim that they can generate causal models with enough accuracy, they usually hold very strong assumptions on data distributions (e.g. Multinomial, or continuous with Gaussian sample error) that are hard to keep in practice, and make algorithms unstable when error relations are generated in early stages. Thus none can guarantee an answer that is accurate in the sense of being completely consistent with the real world. Even with the emergence of big data the automated derivation of a consistent causal model remains challenging because it requires a fundamental theory of how and why the observed phenomena occur. This in turn requires creativity with the human expert in the inference process. This is feasible when the model is sufficiently small, that is, the number of variables in the model is manageable. However, big data not only increases the number of observations, it also typically gives access to a greatly increased number of variables. These can help users build a consistent theory of the real world phenomena but the process is difficult to manage without visual support.

The system we describe in this paper, the *Visual Causality Analyst,* is a first step into creating such a visual support system. It offers various interactive and automatic tools for visual causal discovery. Following previous work on correlation maps [6] and Pearl's depicting of the causality structure as a directed acyclic graph (DAG) [2, 3, 7] our framework visualizes the causal relations as an interactive spatial 2D layout, in which each edge connecting two variables implies a causal relation and the direction of an edge identifies the effect from the cause. Our interface also offers real time visual interactions where users are allowed to arbitrarily change the relations between variables and the impact of each modification is visualized simultaneously on the graph. Mathematical measurements of causal relations in the form of either linear regression analysis (targeting numeric variables) or logistic regression analysis (targeting categorical variables) are calculated and fed back along with the interactive operations, enabling users to explore potential causalities with statistical proof. Subsequently, these measurements are then also visualized in the spatial layout in terms of edge colors and opacities.

The main utility for computational causal inference lies in the conditional independence (CI) test, which is usually conducted via G-test or partial correlation. The former applies to discrete (categorical) data, while the latter applies to continuous (numerical) variables. None can handle both. We choose the partial correlation approach since our motivating domain application has mostly numerical variables and discretizing numerical variables into bins causes loss of detail [8, 9] which is undesirable. To go the other way, we are inspired by recent work of Zhang et al. [6] which for each pair of categorical and numerical variables reorders and repositions the levels of the categorical variable such that Pearson's correlation between the pair is optimized. To accommodate the computational causal inference process we extend Zhang's level reordering and repositioning mechanism from a single numerical variable to sets of numerical variables. This global optimization mechanism enables the causal inference algorithms to return plausible results on datasets containing both continuous and discrete data.

Our paper is structured as follows. Section 2 discusses related work. Section 3 introduces theoretical background and contributions. Section 4 introduces our novel Visual Causality Analyst interface. Case studies on multiple datasets are given in Section 5, and Section 6 ends with conclusions and an outlook on future work.

2 Related Work

As mentioned, causality has been an active research topic and research on the visualization of causal networks has also emerged. In the following we shall briefly review this work.

2.1 Causality Visualization

A number of methods have been developed for the visualization of causality. The Hasse diagram is one of the earliest systems that have the ability to represent causal relations. It was originally introduced in order theory and has been adopted for demonstrating distributed systems [10], parallel processes [11], and many other information structures that contain causal events. However, since Hasse diagrams typically produce layouts with a large amount of intersecting edges and lack the ability to represent causal semantics, it can be difficult to comprehend them, especially when the number of variables is large and causal relationships are complex.

Growing-squares [12] and its enhancement *growing-polygons* [13] are both animated techniques that focus on visualizing sets of connected causal events called *processes*. The latter uses *n*-sided polygons to represent *n* processes and the gradual change of processes is visualized by animating the polygon's change of size. However, the growing-polygons can only illustrate causality at the process level with very limited abilities of signifying causal relation strengths. Kadaba et al. [14] address these problems by depicting causal relations by node-link arrows and glyphs, leveraging simple animation of node sizes to indicate interactions between the factors and the target. While such a graph design can be effective for causality visualization, it only feeds back brief semantics of a causal relation, e.g. positive or negative. However, when explicit causal measurements need to be demonstrated on the graph, no existing causal visualization approach can give a plausible result. Wongsuphasawat and Gotz [15] describe a system that visualizes alternative pathway chains of temporal event sequences. While these chains suggest causal effects they are not casual networks. Also, their system focuses mainly on event flow visualization and has no support for interactive statistical causal reasoning.

According to Pearl's DAG patterns of causal structures [2], a 2D spatial graph layout of the network is a natural fit. Spatial graph layouts have been widely used in information visualization in various contexts. A related example is the visualization of Bayesian belief networks [16], in which the layout is guided by a temporal order, and multiple visual variables like color, node size, and proximity are used to represent network semantics. More recently, Zhang et al. [6] demonstrated an interactive correlation map with spatial representations. By ways of slider bars, users can filter edges corresponding to weak relations. Our work is inspired by these methods and we extend them for the visualization of causal relations, providing a suite of interactive utilities to manipulate the graph.

2.2 Causality Representation and Inference

Our framework provides automatic discovery of causalities in the data, thus causality representation and inference algorithms are closely related to our work. The causality system is often represented as Bayesian Networks (BN) [17, 18], in which causal relations are represented as dependencies measured by conditional probabilities. Algorithms recognizing BN structures usually require knowledge of the data distributions, which is difficult to achieve in practice especially with continuous data. For this type of data, Structural Causal Models [3, 2, 7, 19] which assume effects are linear functions of their causes plus Gaussian noise are better suited. The structure of this model is typically built via a multi-phase process involving a number of CI tests using partial correlations [20, 21, 22, 4]. Unfortunately, these algorithms often fail when categorical variables are included in the data. Introducing dummy variables is a standard technique in statistics [23], but the resulting significant increase in the number of variables may lead to an exponential increase in the number of CI tests needed, and also the mutual exclusions among dummies from the same variable can be very difficult to guarantee.

For the purpose of handling both categorical and numerical variables in a correlation network, Zhang et al. [6] recently proposed an algorithm that uses a pairwise optimization approach to reorder and reposition the levels of each categorical variable with respect to each numerical variable. The new levels are computed by maximizing Pearson's correlation with the pair's numerical variable. This approach is superior to other encoding methods like [24, 25] in that it provides both reordered and optimized distances between categories. However, in contrast to correlation networks, causal inference requires a global frame and so the algorithm proposed by

Zhang et al. is not directly applicable. But it served as an inspiration for the global optimization approach we developed which computes the new level values of a categorical variable with respect to all numerical variables in the system.

3 Theoretical Background and Contributions

Our causality analysis framework comprises the following three steps: (1) find all true CI relations embedded in the observational data (see Section 3.1), (2) build a DAG that is consistent (termed *faithful*) to all of these conditions (see Section 3.2), and (3) compute the causal strengths of the relations coded by the DAG (see Section 3.4). Steps 1 and 2 make use of correlation analysis where we require a novel transformation of categorical to numerical variables which we introduce in Section 3.3. Conversely, step 3 uses dedicated regression analyses where no such transformation is needed. Our treatment of steps 1-3 is necessarily brief and the reader is referred to the tutorials by Pearl [7] and Spirtes [26] for more detail.

3.1 Causality Analysis and CI Test

The idea of causality analysis is based on counterfactual theory, which can be explained in terms of the form "If A had not occurred, B would not have occurred". Although counter facts cannot be treated equally as causation on a philosophical level [27], causality analysis serves uniquely in telling us how a distribution would differ if external conditions were changed by treatments or interventions [7]. To achieve such functionality, CI tests are used as core instruments. The goal of a CI test is to find out whether two variables are related when the rest of the system is controlled, i.e., test the dependency of two variables while eliminating the impact of all other variables or at least a subset of them. This can be interpreted as a simulation of a controlled experiment on observed data.

In statistics, for some variables X and Y in a numerical dataset, a CI test is equivalent to a test for zero partial correlations between X and Y given a set of other variables $\boldsymbol{Z}$ in the dataset. This is called *conditioning* on $\boldsymbol{Z}$ [2, 3]. The partial correlation between X and Y given $\boldsymbol{Z}$ is defined as the correlation of the residuals from regressions of X on $\boldsymbol{Z}$ and of Y on $\boldsymbol{Z}$. In a dataset, the partial correlations from each pair of variables conditioned on all remaining variables form a partial correlation matrix. Such a matrix can be efficiently computed based on the correlation matrix $\mathbf{R}$, so that with $\mathbf{R^{-1}} = (r^{ij})$, we have

$$\rho_{X_i X_j \cdot \boldsymbol{V} \setminus \{X_i X_j\}} = -\frac{r^{ij}}{\sqrt{r^{ii} r^{jj}}} \quad (1)$$

where X_i and X_j are two variables, and $\rho_{X_i X_j \cdot \boldsymbol{V} \setminus \{X_i X_j\}}$ is the partial correlation of X_i and X_j given all other variables in the dataset. Then with the partial correlation matrix, we are able to find all potential causal relations. This process as a whole is often called *feature extraction*, which is the first step in many causal discovery algorithms [22, 20, 28].

However, variables actually independent to each other may still be found causally related when conditioned on certain variables. Suppose a graduate school admits students only by the sum of one's GPA and personal statement score. We may find these two scores negatively correlated within those who are admitted, as high GPA with low statement score or low GPA with high statement score is just enough for being admitted. But there is no apparent causal relation between the two scores in the real world. This means some variables (admission status in this example) cannot be conditioned on in finding the true CI relations between two variables. Such variables are called *colliders* and their descendants, and conditioning on them will generate false causations and introduce triangle patterns. The right set of variables to be conditioned on so that two variables can be deemed having a CI relation is called *d-separating* set. If no such set can be found in the dataset for a pair of variables, we can infer there is direct causation between them. All terms refer to [2].

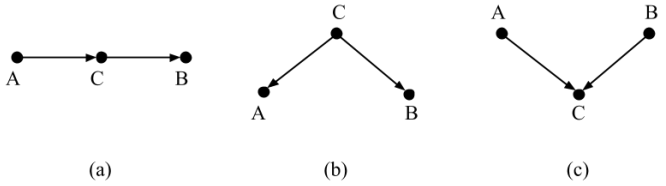


Fig. 2. Three basic patterns of causal relations among any tree observed variables related to each other: (a) a chain of causal relations from A to B via C; (b) a common cause C influencing both A and B; (c) a common effect caused by both A and B.

3.2 Causal Graph and Inference

The goal of causality analysis is to build a causal graph that is faithful to all the CI relations embedded in the observational data. A causal graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ is a DAG that consists of a set of vertices $\boldsymbol{V}$ denoting the variables and a set of directed edges $\boldsymbol{E}$ denoting the causal relations between two variables. Assuming there is no latent variable, the basic graph pattern of the causal relations among any three observed variables related to each other are: (1) a chain of causal influences (Fig. 2a), (2) a common cause influencing multiple variables (*confounding*, Fig. 2b), or (3) a common effect caused by multiple variables (*selection bias*, Fig. 2c). The first two patterns imply the same conditional independency which is "A is independent of B conditioning on C". But the third pattern, also called the V-structure, is different as C is just the collider of A and B as mentioned in the previous subsection, thus the true independency of A and B can be recognized only when C is NOT conditioned on.

However, in feature extraction we simply conditioned on all other variables and no causal relations are oriented, thus patterns as Fig. 2c would become an undirected triangle, and patterns in Fig. 2a and b would look the same. The resulting undirected graph is often called a Markov field or a moral graph depending on the author.

How to remove false links and orient the edges correctly is one of the major issues in modern causal inference researches. The usual procedure is to look at each pair of connected variables and conduct a subset search for colliders in variables forming triangles with them. If colliders are found then the two variables are disconnected and V-structures are recognized. This process costs a number of CI tests exponential to the number of variables forming triangles with each variable pair in the undirected graph. This is where the main computation cost lies. After all triangles have been processed a constraint propagation algorithm is run and a maximally but often partially oriented DAG is obtained [2, 4].

It is worth noting that partially oriented graphs returned by such a causal inference process only represent *observationally equivalent classes* [2] of true causal graphs, as there may be multiple DAG corresponding to the same set of CI relations. Expressed formally, for the generated $\boldsymbol{G}$ and some variables X, Y and Z in $\boldsymbol{G}$,

$$\begin{aligned} X, Y \text{ adjacent in } \boldsymbol{G} &\Leftrightarrow X \to Y \text{ or } Y \to X \text{ in reality} \\ X \to Z \leftarrow Y \text{ in } \boldsymbol{G} &\Leftrightarrow X \to Z \leftarrow Y \text{ in reality} \end{aligned} \quad (2)$$

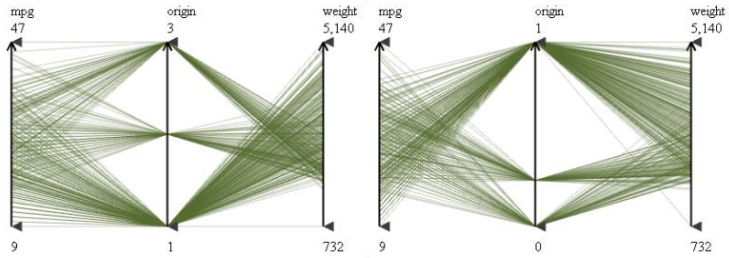
This means that we will always need further verification to obtain the perfect causal graph in practice. Our system is purposed to help analysts in this verification task, using visualization to allow them to maintain their bearings on all levels of scale.

3.3 Correlations of Categorical & Numerical Variables

To efficiently compute the partial correlation matrix for the CI tests, we need to calculate a correlation matrix first. While correlations between pairs of numerical variables and pairs of categorical variables can be achieved accordingly with Pearson's correlation coefficient and Cramer's V, traditional methods applicable for correlations between numerical and categorical variables, e.g.t-test, ANOVA, and MANOVA, have the problem that they are not normalized, so that one must consult significance tables to measure

Variable pair categorical/numerical	Pair-Opt Correlation	Global-Opt Correlation
origin/horsepower	0.488	0.476
origin/weight	0.595	0.561
origin/displacement	0.656	0.637
origin/mpg	0.576	-0.530
origin/timeTo60mph	0.272	-0.272

(a)



(b) (c)

Fig. 3. Effects of global optimization on the categorical variable *origin* from the auto MPG dataset [29]. (a) Comparisons between correlations of *origin* and several numerical variables under pairwise optimization and global optimization. The correlation value are similar in scale under two value mapping approaches. The different signs of correlations in the last two rows mean the level ordering after global value mapping is just the opposite to that after pairwise value mapping. (b) The parallel coordinate view of *mpg*, *origin*, and *weight* before global value mapping. (c) The parallel coordinate view of *mpg*, *origin*, and *weight* after global value mapping. We can see that both ordering and distances between categories are optimized, so that correlation is more visible than that in (b).

the association between two variables. However, the ability to handle mixed types of variables is often required in practical applications. Our solution to this problem makes use of Zhang et al.'s approach [6] which returns a maximized Pearson's correlation between a pair of numerical and categorical variables using the following equation:

$$v_c^i(j) = \mu\big(v_i(j)\big) \tag{3}$$

Here, $v_c^i(j)$ is the value assigned to level j of categorical variable v_c regarding to numerical variable v_i and $\mu(v_i(j))$ is the average of v_i corresponding to level j of v_c. This will bring an optimized ordering and distances of v_c's levels regarding v_i.

This method, however, is only partially useful for casual analysis because the level ordering and adjustment for a given categorical variable will be different for each numerical variable. This is fine for correlation analysis but causal reasoning requires global consistence of variable values in all CI tests. In the following we describe a novel generalization of the scheme of Zhang et al. that can achieve this.

An ideal globally consistent value mapping should be such that correlations between the categorical and all numerical variables are simultaneously maximized. A naïve idea would be to simply mediate all pairwise optimized values mapped from each numerical variable, setting up the target equation as,

$$\underset{v_c^{opt}}{arg\,\min} \sum_{i=1}^{D}\sum_{j=1}^{L} \left\| v_c^{opt}(j) - v_c^i(j) \right\|_F \tag{4}$$

in which we suppose there is a categorical variable v_c with L levels and D numerical variables $v_i, \mathrm{i} = \{1, 2, \dots, D\}$ in the dataset. $v_c^{opt}(j)$ is the global optimized value we require for level j of v_c, and $v_c^i(j)$ is the pairwise optimized value for level j with regards to numeric variable v_i.

However, with the empirical knowledge that strong causal relations typically lead to strong correlations (although this is not true reversely), the values of v_c's levels should more depend on numerical variables that are strongly correlated with it, but less on those are weakly correlated with it. This can be easily implemented by weighting the inner summation of equation (4) with the pairwise optimized correlation between v_c and v_i, namely ρ_i.

If two orderings of v_c's levels regarding two different numerical variables v_p and v_q are just opposite to each other, solving equation (4) will result in that all v_c's levels have similar values. The solution is to reverse one of two orderings so that the two become identical. This is equivalent to changing the sign of its pairwise optimized correlation weighting the inner summation. The mechanism to decide whether a level ordering should be reversed can be achieved by testing the sign of a *correlation of orderings* measurement, in which we consider v_c's level ordering $\boldsymbol{\lambda}_{c,p}$ regarding to v_p as a standard, then reverse the ordering $\boldsymbol{\lambda}_{c,q}$ with regards to v_q when the correlation of $\boldsymbol{\lambda}_{c,p}$ and $\boldsymbol{\lambda}_{c,q}$ is negative. The selection of v_p can be the one with largest pairwise correlation with v_c, call it v_{MAX}. Let $\Theta(v_p, v_q)$ be the decision function representing the process and $\rho(\boldsymbol{x}, \boldsymbol{y})$ be the correlation function, then,

$$\Theta(v_i, v_{MAX}) = sign\left(\rho\big(\boldsymbol{\lambda}_{c,i}, \boldsymbol{\lambda}_{c,max}\big)\right) \tag{5}$$

We are now ready to put together the final target equation, using $\Theta\big(v_p, v_q\big)$ and ρ_i as weights in the outer summation of equation (4):

$$\underset{v_c^{opt}}{arg\,\min} \sum_{i=1}^{D} \Theta_i \rho_i \sum_{j=1}^{L} \left\| v_c^{opt}(j) - v_c^i(j) \right\|_F \tag{6}$$

Here we use Θ_i to denote $\Theta(v_i, v_{MAX})$ for convenience

We found that satisfactory results can be achieved when $F = 2$. Then by making (6) equal to 0 and differentiating on $v_c^{opt}(j)$, we can solve the optimization problem and obtain a closed formula,

$$v_c^{opt}(j) = \frac{\sum_{i=1}^{D} \Theta_i \rho_i v_c^i(j)}{\sum_{i=1}^{D} \Theta_i \rho_i} \tag{7}$$

As $\sum_{i=1}^{D} \Theta_i \rho_i$ only serves as a normalization factor, also combining equation (3) we obtain,

$$v_c^{opt}(j) \propto \sum_{i=1}^{D} \Theta_i \rho_i \mu\big(v_i(j)\big) \tag{8}$$

With equation (8), we can now assign numerical values to v_c's levels, which can be used consistently in causal inference processes.

Fig. 3a illustrates how this global optimization performs for the auto MPG dataset [29] ($D = 7$, 392 data points). In the table, the first column gives the variable pairs, in which *origin* is a categorical variable and all others are numerical variables. The second column shows the correlation using Zhang et al.'s pairwise optimized assignment for each level of *origin*. The third column shows the (similar) correlations obtained with our global optimization method. For the last two variable pairs, the different sign of global from pairwise correlation means that the level ordering is just the opposite. Fig. 3b and c show two parallel coordinate tiles before and after the transformation, respectively. We observe that after the transformation, (1) categories (levels) that behave similarly are put close to each other; and (2) the correlation is more visible in the parallel coordinate plots.

3.4 Regression Analysis

After the structure of the causal graph model has been recovered, we need tools to model, measure, and test the causal relations statistically. In Pearl's theory of Structural Causal Models [3, 2, 7], linear regressions are used as such tools. Linear regression measures the linear relationships between a dependent variable y and one or more explanatory variables $x_k, k = \{1, 2, \dots, K\}$, taking the form

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (9)$$

In this equation, the subscript i indicates the i-th observation and ε_i represents the intercept, which is interpreted as causal effects from latent factors (e.g. unobserved variables, sampling noise) in causality theory; β_k is the regression coefficient for x_k, which is also taken as the main measurement of causal strength. If x_k is a L level categorical variable, it is turned into $L-1$ binary dummy variables, each standing for a level of x_k. Linear regression analysis can test the statistical significance of each explanatory variable via student's t-test, as well as test the goodness of fit of the whole model via F-test, R-squared coefficients, and many other statistical utilities.

Our framework uses logistic regression analysis to measure causal relations targeting categorical variables. Logistic regression analysis, although named "regression", is actually a model of classification probabilities, i.e. the probability of the categorical variable taking a certain level. It is a better fit than linear regression analysis in models targeting categorical variables. It takes the form of a logistic function as:

$$\sigma(t) = \frac{1}{1+e^{-t}},$$
$$\text{where } t = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (10)$$

in which β_k is the structural coefficient for x_k and measurement of causal strength, with error term ε_i representing the disturbance from latent factors. If x_k is a categorical variable with L levels, it is turned into $L-1$ dummy variables. Logistic regression analysis can also test for variable significances (via Wald statistics) and for goodness of fit (via deviance, likelihood ratio tests, and so on). Note here the optimized values of categorical variables are not used in either regression analyses, but only in causal structure inference processes.

4 The Visual Causality Analyst

In the following we use the Auto MPG dataset [29] to illustrate our interface and the interactions we defined on it. This dataset has eight variables – one of them categorical (*origin*) – and 392 instances. Fig. 1 shows all elements of our interface. The main window contains the 2D spatial layout of the causal graph in the center and the regression analysis view on the right. The variable type window on the left opens when a new dataset is read in, allowing the user to indicate which of the variables are numerical and which are categorical.

4.1 The Causal Graph Display

The causal graph is generated by the causal inference algorithm described in Section 3, using our global value mapping scheme for the categorical variables. The causal graph display provides an overview of all data dimensions in terms of their causal relations in variable space. In this display the vertices correspond to variables, laid out via a *Fruchterman-Reingold* force-directed model [30]. We set all edges to have the same natural strength so that vertices are uniformly spread on the canvas. The color of a vertex encodes the type of variable, blue for numerical and yellow for categorical. We use a different color for categorical variables since they will usually turn into dummy variables for the regression analysis. As such, each yellow vertex in the graph may correspond to multiple variables used in several regression analyses.

The edges of the graph link two variables in terms of their causal relationships. The direction icon on an edge encodes the direction of the causal relation, going from cause to effect. The colors of the direction icon encode the type of the causal relation. Green arrows encode positive relations, red arrows encode negative relations, and a yellow arrow emanating from a categorical variable corresponds to multiple relations between the target and dummies of the categorical variable. If the target variable is a categorical variable, the arrow will be yellow too. The reason to use yellow arrows is that complex causal relation involving categorical variables cannot be simply described as negative or positive.

The opacity of the edge encodes the amount of change that is exerted by the cause onto the effect, which is measured by regression coefficients. A more visible edge has a stronger effect. However, edges with too low opacities are often difficult to observe on the graph. Thus gamma correction is introduced such that for an edge connecting variable v_i and v_i with regression coefficient β_{ij}, its opacity R_{ij} is

$$R_{ij} = \frac{|\beta_{ij}|^{\gamma} + \delta}{D} \quad (11)$$

where D is the normalizer to make all opacities lie in the range of $[0,1]$, γ is the gamma value, and δ is the offset to guarantee minimum opacity. Usually we set $\gamma = 0.8$ and $\delta = 0.1$ to avoid an edge to be rendered too weak to be observed, and at the same time rendering a strong edge evidently darker than a weak edge.

Below the graph in Fig. 1 is a control panel that allows users to run the causal inference algorithm – via the causal layout button – as well as add edges, give them cause-effect directions, and test these in the regression analysis after which a re-layout of the casual graph might be run. Slider bars allow the user to either filter away or enhance the opacity of weak edges. There is also a button to load new data which pops up the specification window on the left.

4.2 The Regression Analysis View

When a variable is selected the system computes the regression model for all variables with incoming casual edges to it. In statistics, the former variable is often called the response variable, while the latter are the predictor variables. The regression analysis view shows the regression coefficient for each predictor variable as well as the p-value to give an indication of. The fit of the overall model can be gauged by the R-square metric. It is 1.0 when the regression model fit perfectly. The R-square metric gains in meaning especially when it is used to compare regression models. If R-square decreases significantly when a predictor variable is dropped from the model then there is a good chance that this variable was required. Another test statistics our system reports is the F-value gauged by the F-statistics. The F-statistics is also particularly useful for comparing two competing models. We can write

$$F = \left(\frac{SSE_1 - SSE_2}{SSE_2}\right) \Big/ \left(\frac{DF_1 - DF_2}{DF_2}\right) \quad (12)$$

where SSE is the residual sum-of-squares (RSS) of a model and DF is its degrees of freedom which is the number of observations minus the number of predictors minus 1. Let's assume that SSE_1 is the RSS for the model with fewer predictor variables. Assume SSE_1 is higher than SSE_2 which is the RSS of the model with more predictors, and DF_1 is higher than DF_2 since there are fewer predictors. Now, If the more complicated model is correct, we can expect the relative increase in RSS (going from the complicated to the simple model) to be greater than the relative increase in the degrees of freedom, or $(SSE_1 - SSE_2)/SSE_2 > (DF_1 - DF_2)/DF_2$. The significance of this increase can be tested via the F-statistics, but even informally, when F is large when a predictor variable is included in the model, we know that this predictor was valuable.

4.3 Illustrative Example #1

In the graph of Fig. 1 the user has selected a few edges to highlight the causal flow they are part of. For example, the user marked all edges that link the miles per gallon (*mpg*) rating of a car to the factors that might cause this rating (and the physical process behind it). These factors are *weight*, *origin*, and *model_year*. The regression analysis window gives the statistical measurements and proofs for the identified causal models by means of linear regression and logistic regression analyses. Here we learn that *weight* has a strong negative effect (the regression coefficient is -0.647), while the effects of the other factors are rather mild. All but one effect is statistically significant – their p-values are less than 0.05.

4.4 Interaction with the Causal Graph Interface

It is often the case that true causalities may be missing or are wrong in a causal graph built from observational data. For this reason further verifications (hypothesis tests) on identified causal relationships are always needed. These verifications usually require modification of the causal graph by means of connecting two vertices and assigning the causal direction, reversing the direction of an edge, deleting an edge, or marking an edge as unknown (of direction). Furthermore, especially in the presence of large causal graphs, users will wish to focus on certain variables and their casual relationships while hiding all others.

Our causal inference interface provides interactive utilities capable to perform all of the above functions with visual feedback. That is to say, whenever the causal graph is modified, the impact of the modification on the rest of the graph, e.g. direction icon colors and edge opacities, will be updated immediately. Vertices of variables of interest can be selected either in the graph or by marking them in the variable list. Edge selection is achieved by either clicking on them in the graph view or choosing them in the control panel. We note, however, that any deselected (hidden) variable should still be taken into consideration in the causal structure learning process as we need to condition on them in CI tests. If hidden variables are not considered then erroneous causal relationships might be inferred. This is similar to the case when important variables have not even been observed. In both cases our visual interface provides a helpful medium for human experts to recognize these false relationships and seek their resolution.

Causality is subtle, and to test and measure it, we make use of the statistics and regression analyses tools described in Sections 3.4 and 4.2. We show the results of these analyses, such as coefficients and others in the regression analysis view whenever a causal graph is generated. The analysis view also provides automatic update on the analysis results whenever the graph is modified by the user. Finally, if an edge on the graph is selected, all regression analysis results involving it will be highlighted to made salient for the user.

4.5 Illustrative Example #2

Fig. 4a shows another example for the Auto MPG dataset. Here, the user has decided to focus on the causal graph of *weight*, *horsepower*, and *timeTo60mph*, hiding all other variables and relationships. The graph implies that a light car with high horsepower tends to have short acceleration time, which is consistent with real world knowledge. However, we also see in the graph that high *horsepower* increases *weight* which would bring down *timeTo60mph*. To research this conflict we delete the edge from *horsepower* to *timeTo60mph* and observe (in Fig. 4b) that *weight* and *timeTo60mph* are now negatively related (the visualization updated accordingly). A visual indicator is that the edge opacity dropped compared to the opacity in Fig. 4a. This new relationship is inconsistent with common knowledge and it likely means that only considering weight cannot explain acceleration well. To explore this argument more deeply, a detailed statistical proof is needed. This proof can be provided by the regression analysis view of our framework.

Fig. 4c and d are two screen shots of the regression analysis view showing linear models of *timeTo60mph*, corresponding to the graph models in Figs. 4a and b, generated and updated automatically. We observe from Fig. 4c that when taking both *horsepower* and *weight* as causes, *horsepower* plays a much greater role (with regression coefficient -1.049) in effecting *timeTo60mph* than *weight* (with regression coefficient 0.632). When only regressing on *weight*, its regression coefficient is indeed negative (-0.343, Fig. 2d). However, the R-square coefficient in Fig. 4d is only 0.161, which is much lower than that in Fig. 4c where it was 0.622. This means that the linear model described in Fig. 4d is much worse than that in Fig. 4c, and we verified our previous guess that only considering *weight* will not explain acceleration well. Likewise the F-value drops by a large amount which also indicates that *horsepower* is a significant casual variable should not be dismissed.

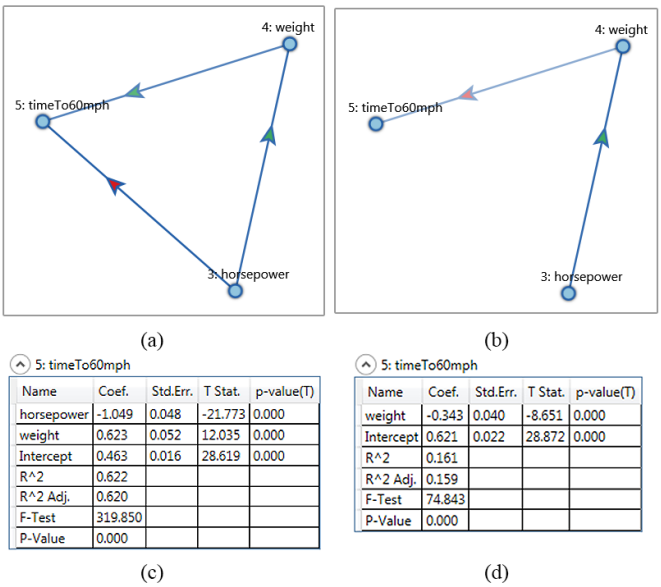


Fig. 4. The visual feedback and statistical analysis provided by the Visual Causality Analyst. (a) A visualized causal graph structure of variable *weight, horsepower,* and *timeTo60mpg* from the auto MPG dataset, generated by our framework. (b) The visual feedback after deleting the edge from *horsepower* to *timeTo60mpg*. Here *weight* becomes a negative cause of *timeTo60mpg*, which is inconsistent with common knowledge. (c) A screen shot of linear regression analysis targeting *timeTo60mpg* in line with the causal model of (a). (d) A screen shot of linear regression analysis targeting *timeTo60mpg* in line with the causal model of (b). Comparing (c) and (d) we can see that only regressing on *weight* cannot explain *timeTo60mpg* well, due to the dropping of R-and F-Test value from (c) to (d).

We learn from this investigation that *horsepower* is indeed a good predictor for acceleration and that the apparent conflict due to the positive causal link between *horsepower* and *weight* might be related to the *weight* variable and not *timeTo60mph.* So we would continue our investigation there (see Section 5.1).

5 Case Studies

We demonstrate our framework with the following three datasets, the first of which we have already used in the previous example.

Auto MPG dataset: This dataset contains 392 complete records of cars with 8 attributes: *mpg, cylinders, displacement, horsepower, weight, timeTo60mph, model_year,* and *origin*, in which *origin* is a three-category nominal variable and all other variables are continuous. All car models in the dataset use gasoline and were built before 1983. This dataset was retrieved from the UCI Machine Learning Repository [29].

Sales campaign dataset: The dataset has been synthesized based on actual data describing the sales marketing and its effects on a company's financials. There are 600 data samples each representing one salesperson, and 10 numeric variables: *%Completed, #Leads, LeadsWon, #Opportunity, PipeRevn (pipeline revenue), ExpectROI (Return on Investment), Cost, Cost/WonLead, PlanRev (planned revenue)*, and *planROI*. This data set was previously adopted for demonstrating the correlation map by Zhang et al. in [6]. We can now make more explicit decisions by ways of causality analysis with our new interface.

Heart disease Dataset: This is a realistic dataset on heart disease diagnosis, retrieved from the UCI Machine Learning Repository [29]. The dataset has 270 diagnosis records, each per person, with 7 categorical variables: *sex, chestPainType,* fastBloodSuger, *restECD (electrocardiographic), angina*, *thalassemia, and disease*; and 6 numeric attributes: *bloodPressure, serumChol (Cholestoral), maxHeartRate, exerST(ST depression induced by exercise), slopeExerST,* and *numVessels (colored by flourosopy).*

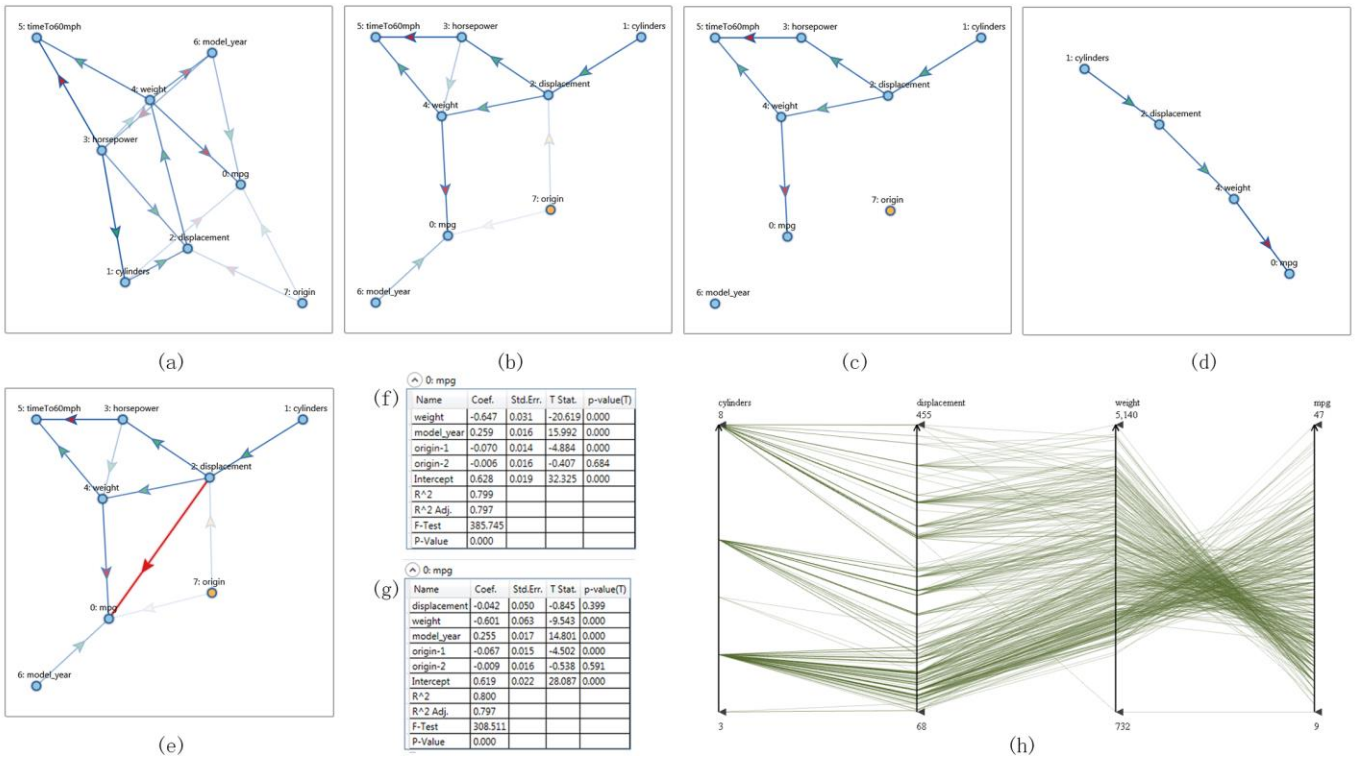


Fig. 5. Causal reasoning on auto MPG dataset with the Visual Causality Analyst. (a) The causal graph generated by randomly assigning values to *origin*'s categories, which introduces several error edges. (b) The causal graph generated by assigning globally optimized values to *origin*'s categories. (c) The graph with regression coefficient threshold of 0.4. Weak causal relations are filtered away. (d) The graph relevant to *mpg*, which is a chain of causal relationships from *cylinders* to *mpg*. (e) The causal graph in which an edge from *displacement* to *mpg* is added and highlighted. (f) A screen shot of linear regression analysis on *mpg* without *displacement*. (g) A screen shot of linear regression analysis on *mpg* with *displacement*. We see that *displacement* has a large p-value in (g), also the F-Test in (g) is decreased from that in (f), so *displacement* should not be considered as a direct cause of *mpg*. (h) The parallel coordinate view of the variables related to *mpg*, in the order consistent to the causal relationships represented in (d). A clear flow of data variable relations can be observed.

5.1 Causality Analysis: Auto MPG Dataset

We firstly present the basic concepts of our Visual Causality Analyst interface with the auto MPG dataset. Fig. 5a gives an initial casual graph generated by randomly assigning values to levels of *origin*. We see here that *horsepower* is mistakenly drawn as the positive cause of *cylinders* and *displacement*. It is common knowledge that these two edges should at least be reversed. The reason for such errors is typical. The feature extraction found an undirected edge between *horsepower* and *origin* with *origin*'s random level values, then cancelled it in the d-separating set search and directed the causal relation as *horsepowe* → *displacement* and *origin* → *displacement*. This error then spread in later processes and affected the direction between *horsepower*-*cylinders*.

A better causal graph is shown in Fig. 5b, which is generated by using globally optimized level values of *origin*. Now we can see that all causal relations between *horsepower*, *cylinders* and *displacement* are correct. We also found that the categorical variable *origin* plays a weak (low edge opacity) cause of *displacement* and *mpg*. As *origin* will turn into dummy variables, it is represented by a yellow vertex. Also the arrows on edges leaving from *origin* are colored yellow as each denotes multiple coefficients. Fig. 5c shows an enhanced causal graph after setting the regression coefficient threshold to 0.4. All weak causal relations are filtered away. We now observe that *origin* and *model_year* are independent of all other variables, and the direct relation between *horsepower* and *weight* has also been eliminated.

Our original purpose for this dataset was to predict car *mpg* and find the direct and indirect causes for it. Fig. 5c suggests that *timeTo60mph* and *horsepower* are not related to *mpg* since there is no causal edge pointing to it. Thus we may unselect them and only lay out those variables that have strong direct or indirect causal relations with *mpg*. Having done this we obtain the causal graph of Fig. 5d, which is a chain of causal relations with four variables. Fig. 5h shows the parallel coordinates plot of these variables in the order of the causal chain. We can clearly observe a flow of associations from *cylinders* to *mpg*.

This chain is consistent with the mechanics of cars, at least when it comes to cars captured in this dataset. Adding cylinders to such a car increases its displacement, but not the other way around since we can also increase displacement by adding volume to the current set of cylinders. More displacement (and the power it affords) requires a heavier car for stability. But moving the extra weight around requires more gasoline, decreasing mpg.

5.1.1 Interactive exploration of causal relationships

One may want to further explore the potential causal relationships that are not suggested by the graph in Fig. 5b. This can also be easily achieved with the interactive tools provided by our framework.

For example, the causal graph did not draw direct edges between *displacement* and *mpg*. However, we might wish to test the hypothesis if this causal relation actually exists. To do this, we can simply select the pair of variables as cause and effect, respectively, in the control panel and assign the edge. The resulting causal graph is shown in Fig. 5e, with the added edge highlighted. Colors and opacities of other edges on the graph may change accordingly if the causal relations they represent are affected by such operation.

To determine whether this causal relation holds, we need to refer to statistical analyses. Two screen shots of the linear regression analyses results before and after adding the edge are shown in Fig. 5f and g. Since the p-value of *displacement* from the student t-test is too large ($p = 0.339$) in Fig. 5g, together with the dropping of F-value, the direct causal relation between *displacement* and *mpg* should not be considered as existing. Hence there is no direct relationship between *displacement* and *mpg*. Raising the displacement of a car

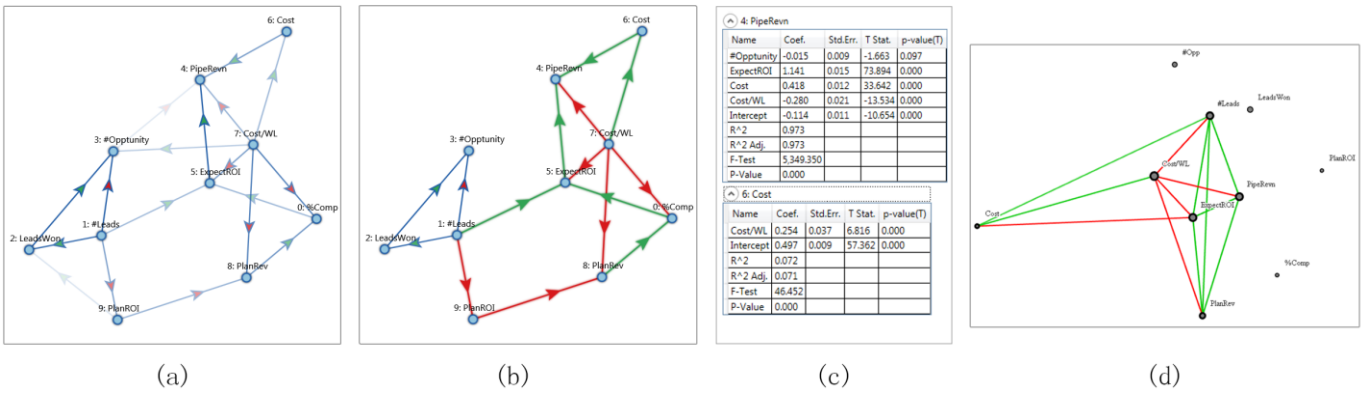


Fig. 6. Strategizing with the Visual Causality Analyst for the sales campaign dataset. (a) The causal graph generated from the dataset showing how the sales system works. (b) All the routes pointing to *PipeRevn* from some variable, indicating possible strategies to increase pipeline revenue. Here *#Leads* and *Cost/WL* are two variables that all routes start from. (c) Screen shots of linear regression analyses on *PipeRevn* and *Cost*. Here the purpose is to investigate the effect of *Cost/WL* on *PipeRevn*. As the scale of the direct effect of *Cost/WL* on *PipeRevn* is larger than the indirect effect via *Cost*, the total effect of rising *Cost/WL* will be the reduction of *PipeRevn*. (d) A correlation map view browsing only variables in a similar strategizing scenario. However, variables correlated to each other may not necessarily imply any causal relationship.

usually will not directly lead to the decrease of its mpg. But when displacement increases, usually the car will be heavier (as mentioned before), which causes the mpg to reduce, since weight negatively changes mpg. In this case, weight serves as a *mediator variable* completing the chain of displacement and mpg.

Many more conclusions can be drawn and many more explorations can be done from this single causality visualization. Therefore we believe our Visual Causality Analyst is powerful and effective for casual reasoning explorations, and the graphs in Fig. 5 may potentially be helpful for consumers to select cars, as well as for car manufacturers to balance their offering of models.

5.2 Strategizing: Sales Campaign Dataset

In this example, we use the Sales Campaign dataset to show how business executives may analyze sales behaviors and strategize with our Visual Causality Analyst software.

To give some background, a sales pipeline typically starts with a *lead* generator responsible for developing prospective customers called *leads* with whom salespersons may actually close deals. *Leads* may become *won leads* when they give positive feedback and then *opportunities* when they offer further interests. For each *won leads*, an increased sales pitch at cost *per won lead (cost/WL)* will be invested. The goal of the entire sales effort is to increase the expected return on investment *(ExpectROI)*, and ultimately maximize pipeline revenue *(PipeRevn)*. In [6] a correlation map was used on the same dataset to showcase its features. In the following, we will demonstrate that by upgrading to the Visual Causality Analyst, the decision making process becomes even more explicit.

Suppose a team of sales data analysts in the company are busy analyzing the sales strategy for the following year, basing on last years' data from their sales teams. Their first step is to build the causal graph of all data attributes to get an overview on how their sales system actually functions. This process is straightforward – import the data with an Excel spreadsheet and lay out the initial causal graph shown in Fig. 6a.

5.2.1 Strategy development

After drawing the sales system's causal graph, the analysts proceed in developing effective business strategies using our interface.

To achieve the goal of increasing the pipeline revenue, the analysts first filter out weak relations in the graph and select a series of related causal relations, highlighted in Fig. 6b. These relations form several routes starting from some variables and finally pointing to *PipeRevn*. Clearly on the graph, the variable *#Leads* is the starting point of multiple routes to the final goal. In all of these routes, *#Leads* plays a positive factor for *PipeRevn*, which means increasing the former will lead to an increase of the latter in the end. So the first strategy might be to generate more leads, i.e. reach out to more people to look for potential customers, simply and clearly.

Another variable related to the goal of the sales data analysts is *Cost/WL* which is the sales pitch invested into each won lead. However, the effect of increasing this variable can have both a

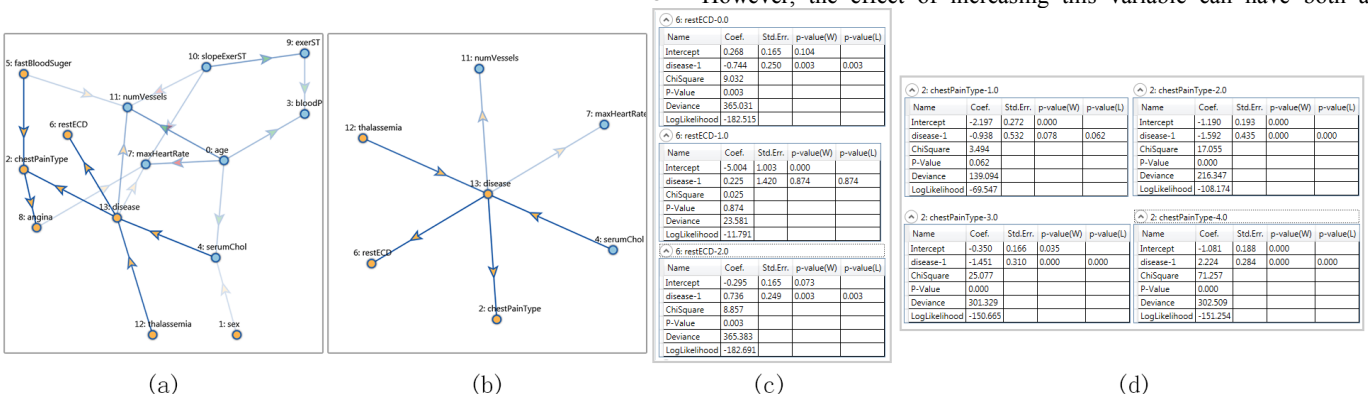


Fig. 7. Analyzing heart disease dataset with the Visual Causality Analyst. (a) An original causal graph generated by the new framework. As there are multiple categorical variables, many nodes and arrows are colored yellow. (b) The causal graph targeting *disease*, in which only variables and relations relevant to *disease* are selected and shown. (c) Screen shots of logistic regression analyses on each of the three *restECD*'s categories. Only electrocardiographic type *restECD-2.0* is found as a sign of heart disease, due to the positive regression coefficient and small p-value of *disease* in its logistic model. (d) Screen shots of logistic regression analyses on each of the four *chestPainType*'s categories. The last type *chestPainType-4.0* should be considered positively relevant to heart disease, while other types are either irrelevant or not a sign of heart disease.

positive and a negative effect on *PipeRevn*. The positive effect is through the variable *Cost*, which is the salesman's total investment. The negative effect can be both direct and indirect through multiple routes. To resolve this conflict, we can refer to the coefficients analysis view of the software. Here we present two screen shots in Fig. 6c. They show that the direct effect of *Cost/WL* in the linear regression model of *PipeRevn* is larger than that of the indirect relation (*Cost/WL* → *Cost)* × *(Cost* → *PipeRevn)*. In addition, the coefficient of *ExpectROI* on *PipeRevn* is more than twice that of *Cost,* and *Cost/WL* negatively impacts *ExpectROI*. Thus, to increase *PipeRevn* we are advised to not increase *Cost/WL* – in fact, we might rather decrease it.

These two strategies essentially imply that, to increase the revenue, each of the company's salespersons should put more effort in exploring new customers. Further, the model indicates that once a potential customer has already shown interest, there is no need to invest extra promotions. It may even have some negative impact on closing the deal.

The strategic guidance our Visual Causality Analyst can provide is explicit and assuring, partly due to its visualization of the casual relationships and partly due to its interactive response rates. When users see the causal graph, they can visually think and form hypotheses that a certain action might potentially lead to a certain outcome. Further, via the regression analysis the size of the effect can also be measured and communicated. This is a significant improvement over the correlation map proposed in prior work (Fig. 6d), by browsing which, users may learn how two variables are correlated in past data (e.g. *Cost/WL* and *ExpectROI*, *#Leads* and *ExpectROI* etc.), but variables strongly correlated to each other may not necessarily imply any causal relationship. And thus, adjusting a variable just based on the correlation map alone will not necessarily lead to the expected change in another variable in the real world.

5.3 Analyzing Categories: Heart Disease Dataset

In this final example we will demonstrate how the Visual Causality Analyst can also be used to visually analyze the causal relationships in medical data that include mixed types of variables.

Suppose an expert on cardiology has been keeping a collection of medical records on his past patients and wishes to identify the most effective methods for diagnosing heart diseases. The expert opens our software and imports his data, then generates the initial causal graph shown in Fig. 7a. Since there are multiple variables that are categorical, we observe many nodes and arrows on edges that are colored yellow.

Any edges on the graph directly pointing to and from *disease* indicate either diagnostics (the outgoing edges) or causes (the incoming edges) of heart disease. These edges and the causal relationships they represent are of greatest interest to the expert. From Fig. 7a he learns that *restECD*, *numVessels*, *maxHeartRate*, *serumChol*, *chestPainType*, and *thalassemia* are all variables directly linked to *disease*. Thus he unselects all other variables and re-lays out the graph, which yields Fig. 7b.

In Fig. 7b, the categorical variable *restECD* has three levels where each represents a type of electrocardiographic test result. To test which type of electrocardiographic result is caused by heart disease, we need to consult the logistic regression analysis. Fig. 7c is a screen shot of the analysis result targeting each of *restECD*'s level. In the first model, *disease* has a negative coefficient and a small p-value, which means *restECD-0.0* is not a sign of heart disease, or even a sign of a healthy heart. In the second model, although *disease* has a positive coefficient, its p-value is too large. The values of other statistical metrics, such as low Chi-Squared value, high model p-value, low deviance, etc. all indicate that *restECD-1.0* is likely irrelevant to heart disease diagnosis. The last logistic model, *restECD-2.0, disease* shows both a positive coefficient and a small p-value, and therefore this test seems to be a valuable means to diagnose an impending heart disease. The expert is satisfied having succeeded in finding an effective means for heart disease diagnostics from his treasure trove of data,

A similar process can be conducted on the variable *chestPainType*. The logistic regression analysis targeting each of its four categories is shown in Fig. 7d. Here we observe that only *chestPainType-4.0* has both a positive coefficient and a zero p-value. Other statistical features of this model, e.g. high Chi-Square value and high deviance also indicate that *chestPainType-4.0* should be considered a sign of heart disease. Other types of chest pains are either irrelevant (*chestPainType-1.0*) or not a sign of disease (*chestPainType-2.0* and *3.0*).

There are many more hypotheses that the expert might discover, test and prove or disprove given his data and using our software. We cannot list them all here. But the case study presented shows that the Visual Causality Analyst is well applicable for health sciences data, as well as all other scientific dataset with mixed types of variables.

6 Conclusion and Future Work

We have presented the Visual Causal Analyst – the first interactive framework for visual causal reasoning and visual causal discovery for high-dimensional data. An added novelty of our framework is that it supports both numerical and categorical variables, which is important for real-world applications. Our interface can serve both as a causality exploration environment and as a platform to visually demonstrate, explain, and justify causal relations that exist in the data with statistical proof provided by linear regression analyses and logistic regression analyses. Our framework is general and applicable to a wide set of real cases, as demonstrated by our case studies.

A present limitation of our framework is that causal relations may exist and vary in different data clusters. Therefore a prior visualization and possibly clustering of the data might be advised. Interactive clustering algorithms, such as ClusterSculptor [31], would allow users to first isolate an independent data cluster and then deduce causalities only on this partial data.

Future work will also focus on visualizing the test statistics, such as R-squared and F-value directly in the visualization. Since the comparison of models (that is, configurations with certain causal edges missing or added) is a frequent task, we might add an information visualization widget that would allow users to compare the values of the test statistics for these alternative models and convey the statistical relevance of the different values.

Another frontier is the ability to perform visual causal reasoning with time series data. This is of great interest to scientists, policy makers, economists, etc. Although we can deal with time series data by simply treating time as a data variable, a dedicated visual analytic approach will be better, possibly using Granger causality.

Finally, we should note that causality can be affected by outliers, non-linear relationships, heteroskedasticity, and multicollinearity. To achieve more statistical robustness, techniques for outlier detection and removal, non-linear causality need to be added to our system.

Acknowledgements

This research was partially supported by NSF grant IIS 1117132 and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the "ICT Consilience Creative Program" supervised by the IITP (Institute for Information & communications Technology Promotion)". Partial support was also provided by the US Department of Energy (DOE) Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences. Some of this research was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's OBER at Pacific Northwest National Laboratory (PNNL). PNNL is operated by the US DOE by Battelle Memorial Institute under contract No. DE-AC06-76RL0.

REFERENCES

[1] "Google Correlate," [Online]. Available: http://www.google.com/trends/correlate/. [Accessed 20 March 2014].

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

[3] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669-688, 1995.

[4] P. Spirtes, C. N. Glymour and R. Scheines, Causation, Prediction, and Search, Berlin: Springer Verlag, 1993.

[5] P. Spirtes, C. Glymour and R. Scheines, "Causality from probability," *Philosophical Studies*, vol. 64, no. 1, pp. 1-36, 1991.

[6] Z. Zhang, K. T. McDonnell, E. Zadok and K. Mueller, "Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 2, pp. 289-303, 2015.

[7] J. Pearl, "An Introduction to Causal Inference," *The International Journal of Biostatistics*, vol. 6, no. 2, pp. 1557-4679, 2010.

[8] P. Royston, D. G. Altman and W. Sauerbrei, "Dichotomizing continuous predictors in multiple regression: a bad idea," *Statistics in medicine*, vol. 25, no. 1, pp. 127-141, 2006.

[9] H. Wainera, M. Gessarolib and M. Verdib, "Finding what is not there through the unfortunate binning of results: The Mendel effect," *Chance*, vol. 19, no. 1, pp. 49-52, 2006.

[10] C. Rehn, "A Definition of Data Consistency Using Event Lattices," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications*, 2004.

[11] L. Viennot, "Parallel N-free Order Recognition," *Theoretical Computer Science*, vol. 175, no. 2, pp. 393-406, 1997.

[12] N. Elmqvist and P. Tsigas, "Animated visualization of causal relations through growing 2D geometry," *Information Visualization*, vol. 3, no. 3, pp. 154-172, 2004.

[13] N. Elmqvist and P. Tsigas, "Causality visualization using animated growing polygons," in *IEEE Symposium on Information Visualization*, 2003.

[14] N. R. Kadaba, P. P. Irani and J. Leboe, "Visualizing Causal Semantics using Animations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1254-1260, 2007.

[15] K. Wongsuphasawat and D. Gotz, "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659-2668, 2012.

[16] J.-D. Zapata-Rivera, E. Neufeld and J. E. Greer, "Visualization of Bayesian belief networks," in *Proceedings of IEEE Visualization'99, Late Breaking Hot Topics*, 1999.

[17] N. Friedman and D. Koller, "Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery," *Machine Learning*, vol. 20, no. 1-2, pp. 95-125, 2003.

[18] S. Nadkarni and P. P. Shenoy, "A Bayesian network approach to making inferences in causal maps," *European Journal of Operational Research*, vol. 128, no. 3, pp. 479-498, 2001.

[19] J. Peters, J. M. Mooij, D. Janzing and B. Schölkopf, "Causal Discovery with Continuous Additive Noise Models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2009-2053, 2014.

[20] Z. Wang and L. Chan, "An Efficient Causal Discovery Algorithm for Linear Models," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[21] K. Baba, R. Shibata and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 647-664, 2004.

[22] J.-P. Pellet and A. Elisseeff, "Using Markov Blankets for Causal Structure Learning," *Journal of Machine Learning Research*, vol. 9, pp. 1295-1342, 2008.

[23] B. G. Tabachnick and L. S. Fidell, Using multivariate statistics, New York: Harper & Row, 2001.

[24] S. Ma and J. Hellerstein, "Ordering categorical data to improve visualization," in *IEEE symposium on information visualization*, 1999.

[25] J. Cohen, P. Cohen, S. G. West and L. S. Aiken, Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.), Routledge, 2002.

[26] P. Spirtes, "Introduction to Causal Inference," *Journal of Machine Learning Research*, vol. 11, pp. 1643-1662, 2010.

[27] S. L. Morgan and C. Winship, Counterfactuals and causal inference, Cambridge: Cambridge University Press, 2007.

[28] T. S. Verma and J. Pear, "Equivalence and synthesis of causal models," in *The Sixth Annual Conference on Uncertainty in Artificial Intelligence*, Mountain View, 1990.

[29] M. Lichman, *{UCI} Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, 2013.

[30] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129-1164, 1991.

[31] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk and D. Imre, "ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data," in *Visual Analytics Science and Technology*, 2007.