DOI: 10.1111/cgf.13680
Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

An Exploratory User Study of Visual Causality Analysis

Chi-Hsien Eric Yen, Aditya Parameswaran, Wai-Tat Fu

University of Illinois at Urbana-Champaign, USA

Abstract

Interactive visualization tools are being used by an increasing number of members of the general public; however, little is known about how, and how well, people use visualizations to infer causality. Adapted from the mediation causal model, we designed an analytic framework to systematically evaluate human performance, strategies, and pitfalls in a visual causal reasoning task. We recruited 24 participants and asked them to identify the mediators in a fictitious dataset using bar charts and scatter plots within our visualization interface. The results showed that the accuracy of their responses as to whether a variable is a mediator significantly decreased when a confounding variable directly influenced the variable being analyzed. Further analysis demonstrated how individual visualization exploration strategies and interfaces might influence reasoning performance. We also identified common strategies and pitfalls in their causal reasoning processes. Design implications for how future visual analytics tools can be designed to better support causal inference are discussed.

CCS Concepts

•***Human-centered computing*** → ***Empirical studies in visualization; Visualization design and evaluation methods;***

1. Introduction

With the open data movement and the rise in popularity of business intelligence software, data visualization is within the reach of a large percentage of the population. More and more public datasets can be found online from governments, health organizations, and research institutions, and can be effortlessly explored in easy-to-use tools, such as Tableau [MHS07], MS Excel, Qlik. These tools enable a broad range of people, many without data analytics experience or knowledge of statistics, to discover correlations and trends from the data and inform data-driven decisions in their daily lives [MBG*14]. Causal inference, which allows people to explain the relationships among variables and predict outcomes, is thus performed implicitly while generating actionable insights.

In the statistics and visual analytics communities, considerable effort has been devoted to the development of statistical frameworks or systems specially designed for causal inference [Pea09, SGS00, WM, WM16]; however, such tools require programming skills or statistics knowledge to use. As a result, non-expert users may choose to use visualizations to infer causality. While existing general-purpose visual analytics tools have made constructing a visualization simple, they do not effectively support causal inference and help users avoid common reasoning pitfalls. In addition, even for experts, visualizations are commonly used during exploratory data analysis to form initial hypotheses. Reducing error and biases in such early stages of analysis could save considerable downstream time. To the best of our knowledge, little is known about how and how well people perform causal inference using general-purpose visual analytics tools. Therefore, our study focuses on studying what factors might influence human causal reasoning performance and understanding how visual analytics tools can be designed to support causal inference better.

Making causal inferences based on visualizations, called *visual causality analysis* hereafter, is arguably tricky, as visualizations only show correlations, which do not always imply causation. In fact, neither visualizations nor statistical mediation analysis [FSM11] always provides real-world ground truth in terms of causality. In this paper, we therefore do not study how individuals can prove causal links through visualizations, but consider causal inference as a broad term referring to making reasonable causal hypotheses based on available data. We aim to understand how we can best assist human causal reasoning so that more accurate inferences can be made.

As a first step, we focus on understanding how individuals generate causal hypotheses during a common and important task in visual analytics: identifying *mediators* between two potentially related variables. A mediation model is a fundamental causal model stating that the independent variable influences a mediating variable, which in turn influences the outcome variable [BK86]. We chose this task because it is a type of analysis that is often conducted when assessing causality.

Neglecting to take into account possible mediators while performing causal inference could lead to incorrect findings and pose great risks to individuals who make decisions based on visualizations. Say a market analyst of a company studying customer demographics finds that male customers purchase their products more than females do; thus they decide to invest marketing resources towards male customers. However, such a relationship between gender and purchase behavior might be mediated by the height of the customer, i.e., the product is favored by people who are taller, and males are generally taller than females. In this case, a better marketing strategy would be targeting tall customers, while the previously mentioned strategy is made inaccurately due to erroneous visual causality analysis.

To design a visual analytics system that effectively assists causal inference, a critical aspect is to understand how well people make inferences based on visualizations, and what are common strategies to reach those inferences. Here, we designed an analytic framework based on the mediation model, which allows us to derive the optimal strategy, and use it to systematically evaluate human reasoning performance and identify sources of error.

In our exploratory user study, participants were given a dataset where two out of six variables are mediating a relationship between another two variables. They were instructed on how to use the visualization tool and were asked to identify which variables could explain the target relationship. Besides comparing participants' performance when the variables being analyzed have different correlation patterns, we also tested whether and how their behavior would be affected by an experimental interface feature. During the study, visualization sequences were logged and verbal data was collected using a think-aloud protocol for further analysis. With our analytic framework, this empirical study aimed to answer the following research questions:
RQ1: How well do people perform mediation analysis using a visualization tool (such as ours), and what factors influence their performance?
RQ2: What are common strategies adopted by people when using visualizations to detect mediators in a dataset? What mistakes or errors lead them to incorrect inferences?

Our primary contributions are threefold: 1) we present an analytic framework and conduct a user study to understand how well people infer causal relationships using visualizations, and how their performance is affected by various factors; and 2) we examine the most common strategies people apply for causal relationship discovery, and identify common reasoning errors in such strategies; and 3) based on these findings, we discuss design implications for the development of future visual analytics systems for causal reasoning.

2. Related Work

2.1. Causal Inference

Using mathematics to infer causality has been extensively studied for decades, as it is in our nature to try to explain why an event occurs and how variables influence each other. Mediation analysis, proposed by Baron and Kenny [BK86], is one of the fundamental mathematical frameworks used to explain a correlation between two variables using a third intervening variable. As more variables and more complex causal relationships are analyzed, more advanced mathematical frameworks are necessary, such as path analysis [Wri21], structural equation modeling, and Bayesian networks [Pea09]. These techniques are available in many statistical tools and visual analytics systems [WM, WM16].

However, using these mathematical frameworks and tools requires knowledge of statistics and programming skills. As a result, people who lack relevant training might rely on visualization to draw causal inferences. Even for experts, visualization is commonly used in exploratory data analysis to form initial hypotheses. While a wide range of visualization tools, e.g., Tableau, Excel, Google Sheet, are deliberately made easy to use and intuitive for the general public, they do not actively help users avoid causal reasoning pitfalls. With the increasing popularity of such visualization tools, there is a strong need to understand how people infer causality from visualizations and how visual analytics systems can be designed to support such tasks effectively.

Some researchers have proposed new interface features or visualization methods that could be beneficial for causal inference. For example, Armstrong and Wattenberg proposed a new visualization technique, called comet chart, designed to visualize mixed effects and help users understand Simpson's Paradox, one of the common scenarios in causal reasoning [AW14]. Guo et al. developed algorithms to automatically detect Simpson's Paradox and help users avoid reasoning with spurious correlations [GBK17]. Doris et. al. [LDH*19] designed a system, *VisPilot*, to avoid drill-down fallacies, and Zgraggen et. al. [ZZZK18] presented methods to address the multiple comparisons problem; both are common sources of error in making inferences with visualization tools. While these techniques are promising in improving human causal reasoning, there is a lack of evaluation studies that help us understand how these techniques improve causal reasoning performance.

2.2. Evaluating Human Reasoning Performance

Evaluating human causal reasoning performance has been studied extensively in psychology and cognitive science. For example, the Wason Selection Task (WST) is a logical puzzle game designed to test whether participants can deduce logically correct actions given a conditional statement [Was68]. Since then, WST has been adapted and employed in various contexts [SSBZ00, OC94, CT92], including causal inference [CLAR91], and many other causal inference experiments have been designed to deepen our understanding of human cognitive architecture [WS04, Bur05, OPH*16]. However, most of these tasks are text-based and not in an interactive visualization setting.

Unfortunately, little has been done in measuring the correctness of human causal reasoning results in visual analytics. Most of the work measures human performance in lower-level tasks, such as reading values, finding extremums, comparing trends, and estimating correlation strength from a visualization [YHR*18]. The answers of these low-level tasks are well-defined and can be easily compared to participants' answers; however, for complex reasoning tasks such as causal inference, the answers are often open-ended and too complex to compare. As a result, instead of measuring correctness rate, researchers usually look into human analysis behavior, such as how people reach their findings, how interface designs influence analytic strategies, and what cognitive biases might occur during the analysis. The relevant work in these three aspects is discussed below in turn.

First, to understand how individuals reach conclusions in complex reasoning-based visual analytics tasks, many empirical studies have been conducted to collect and analyze user actions throughout the process [GGZL16, DJS*09, GTS10, cKFY11, DC17]. Multiple types of user action data have been used. For example, prior studies have used think-aloud audio records and screen video capture to extract meaningful actions and reactions of the participants [JSVDBG04, BJK*16, LSD*10, EFN12], or to confirm the results from quantitative analysis [GZ09]. Interaction logs and visualization sequences are also increasingly used to analyze human reasoning processes [GGZL16, DJS*09]. For instance, Guo et al. [GGZL16] used interaction logs and an insight-based evaluation method to understand what actions led to insights. Dou et al. [DJS*09] demonstrated that by examining interaction logs, coders were able to recover the reasoning strategies and methods used by participants.

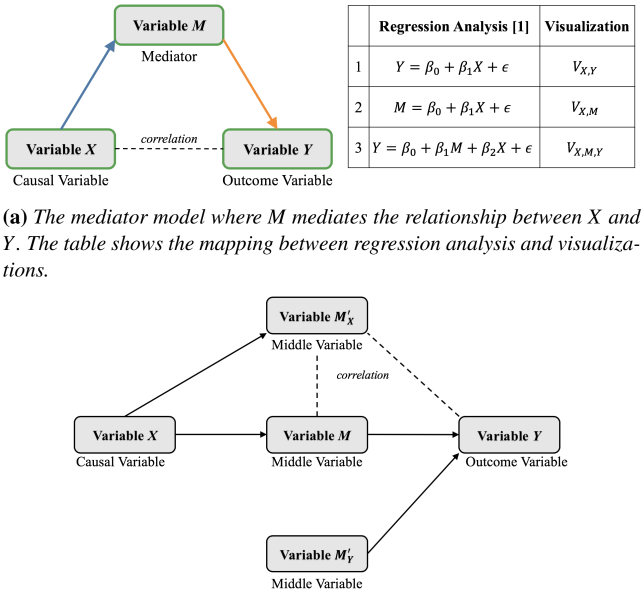
Second, interface design is shown to be one of the important factors in human reasoning performance. Jianu and Laidlaw [JL12] demonstrated that small interface changes could encourage users to consider more alternative hypotheses and collect more evidence in a causal reasoning task. Ottley et al. [OPH*16] showed that the amount and presentation of information influences human performance on conditional probability evaluation tasks. To explore the effects of interface design, we also tested an experimental interface feature in our study.

Finally, prior studies also revealed common reasoning biases or errors in visual analytics, such as confirmation bias [CS08, Nic98], priming effects [VZS18], framing effects [KLK18], and selection bias [GSC16]. These biases are additional sources of human error when it comes to causal inference. For example, when an individual generates an incorrect inference, is it because they have a flaw in their causal reasoning logic? Or they do perform logical actions but interpret the visualizations incorrectly due to biases?

In our exploratory study, we attempt to not only measure how various factors might influence human performance during visual causality analysis, but also observe their analytic strategies and common reasoning errors. If the actual causal model in a causality analysis task is too complex, it would be too difficult to systematically evaluate and compare human analysis behavior. Therefore, as a first step, we focus on the simpler yet important and common task, mediation analysis. This task allow us to develop an effective analytic framework and provide initial insights for future research.

3. Our Analytic Framework

In this section, we present our analytic framework that adopts the mediation model as the underlying causal model. The mediation model proposes that the outcome variable (Y) is influenced by an intermediate variable (mediator, M), which is in turn influenced by the causal variable (X). While the causal variable does not directly influence the outcome variable, an observer might find a correlation between the two variables when the mediator is not controlled due to the causal path through the mediator. Figure 1a displays the graphical model of an mediator.



	Regression Analysis [1]	Visualization
1	$Y = \beta_0 + \beta_1 X + \epsilon$	$V_{X,Y}$
2	$M = \beta_0 + \beta_1 X + \epsilon$	$V_{X,M}$
3	$Y = \beta_0 + \beta_1 M + \beta_2 X + \epsilon$	$V_{X,M,Y}$

(a) *The mediator model where M mediates the relationship between X and Y. The table shows the mapping between regression analysis and visualizations.*

(b) *The extended model where M'_X and M'_Y are included in our study*

Figure 1: *Graphical models of (a) a mediator, (b) our user study's dataset*

Even though the model seems simple, identifying an mediator is challenging because one needs to consider whether a correlation is spurious. Baron and Kenny have proposed a well-known regression-based analysis approach to test the this [BK86]. Here, we extend this statistical approach to a visualization domain, which allows us to effectively understand why, and when, people make mistakes when causally exploring visualizations.

3.1. Identifying Mediators Using Visualizations

The statistical approach consists of three regression equations: 1) regressing Y on X, the coefficient of X must be significant; 2) regressing M on X, the coefficient of X must be significant; and 3) regressing Y on both X and M, the coefficient of M must be significant. Following the same procedure, one can use visualizations to find evidence that supports or refutes the hypothesis that M is the mediator. Specifically, the following three visualizations must be inspected. (We denote a visualization that uses a given set of variables as $V_{variables}$; for example, a visualization that plots variable X and M is denoted as $V_{X,M}$. A table of symbols used in this paper is provided in the supplementary materials.)

1) X-Y Visualization ($V_{X,Y}$)**:** The first aspect to confirm from the visualization is that a correlation exists between X and Y.

2) X-M Visualization ($V_{X,M}$) The X-M visualization should be used to check if there is a correlation between X and M. Without one, M cannot explain the X-Y relationship, even though M might affect Y.

3) X-M-Y Visualization ($V_{X,M,Y}$) Even if one finds correlations between X and M, between X and Y, and between M and Y, these pairwise correlations do not provide sufficient evidence of mediation. Rather, $V_{X,M,Y}$ is required to establish whether, when X is

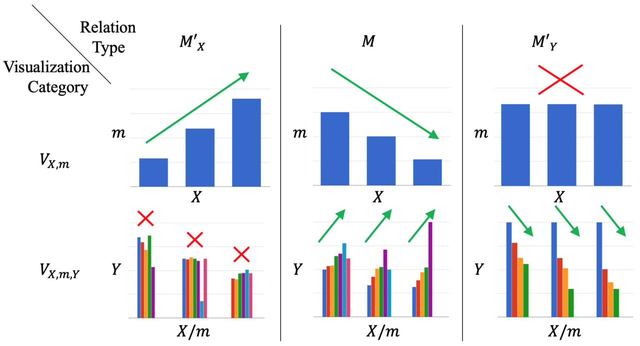


Figure 2: *This figure compiles the key visualizations that are mandatory to correctly identify M, M'_X, or M'_Y in our user study. The green arrows and red crosses represent the existence or nonexistence of key correlations respectively. Note that the arrows and crosses were not shown in our actual user study.*

controlled, the correlation between M and Y is still present. This is analogous to calculating the coefficient of M when regressing Y on X and M in the statistical procedure.

The last step is critical, because it is possible that the M-Y correlation does not exist despite the three pairwise correlations all being significant. In such cases, one would still find a correlation between M and Y, which is called a *spurious correlation*, while X is termed as a *confounding factor* in statistics.

3.2. Extended Model in Our User Study

In our framework, we extended the causal model by adding two other types of variables: one affected by X but not affecting Y (denoted M'_X) and the other not affected by X but affecting Y (denoted M'_Y) (Figure 1b). These two types of variables, M'_X and M'_Y, are not true mediators and were intentionally included as distracting variables to test whether users could correctly identify true mediators. In the rest of the paper, we use lowercase symbol m, or the term "middle variable", to denote any variable that is of relation type M, M'_X, or M'_Y, whenever the actual type of the variable is not determined yet.

Depending on the relation type of the middle variable m, $V_{X,m}$ and $V_{X,m,Y}$ would show different patterns in terms of the existence of correlations. Figure 2 illustrates these visual patterns using a dataset from our study. Each column shows $V_{X,m}$ and $V_{X,m,Y}$ visualizations for M'_X, M, and M'_Y, respectively. As shown in the figure, $V_{X,m}$ shows strong correlation for both M'_X and M, but not for M'_Y. On the other hand, $V_{X,m,Y}$, which shows the trend between Y and m within each group of X (grouped by X and color-coded by m), helps differentiate M'_X from M and M'_Y due to the lack of correlation between Y and m within each group.

With the visualization procedure mentioned above in mind, we can systematically evaluate how people find the true mediators and what mistakes could be made during this process. For example, if a person does not plot V_{X,M'_Y}, they might falsely identify M'_Y as a mediator because that variable does affect Y, which is shown in $V_{X,M'_Y,Y}$. On the other hand, if a person does not plot $V_{X,M'_X,Y}$, they might misidentify M'_X as a mediator due to the spurious correlation between M'_X and Y.

In our study, we tested how human reasoning performance could be affected by the variables' relation types, specifically, by comparing the frequency with which people correctly identify M, M'_X, and M'_Y as mediators. As spurious correlations exist between M'_X and Y, which might mislead people to falsely infer causation, we expected that the performance in identifying the mediator status of M'_Y and M would be better than that of M'_X. Furthermore, different exploration strategies adopted by individuals may or may not include these key visualizations, and therefore affect causal inference performance. This again motivated us to investigate the visualization sequences and study what the common exploration strategies are.

4. User Study

4.1. Synthetic Dataset

The proposed analytic framework necessitates the use of a dataset that has been manipulated to contain certain relationships: specifically, those shown in Figure 1b. Therefore, we adopted the generative data model approach, which has been widely used in evaluating visualization techniques [SNEA*16], to synthesize our dataset.

Due to the wide media exposure and familiarity of the average person to crowdfunding markets such as GoFundMe and Kickstarter, we selected a crowdfunding context for our synthetic dataset. Crowdfunding is a popular fundraising approach whereby anyone can appeal for money from crowds via the Internet. The success of any given crowdfunding campaign depends on many variables, such as the fundraising goal and campaign category, all of which can be manipulated in our synthetic dataset. The target relationship we chose was between the success of a crowdfunding campaign (*Success*) and the month when it was launched (*Month*). This relationship is designed as a negative trend that the projects' success rates decreased every month. Thus, the analytic task is to identify the most plausible reasons for that trend, for example, the success rate decreases because the campaigns launched later have higher fundraising goals, while high fundraising goal campaigns have a smaller chance to succeed. In terms of Figure 1b, variable X is *Success*, Y is *Month*, and the other variables are one of M, M'_X, and M'_Y.

There are two major benefits of contextualizing the dataset in the crowdfunding context as described above. First, as mentioned before, the context is familiar to most people and analytics of this type are likely to occur in real world, thus the test has a good ecological validity. Second, it rules out the reverse causal effects [Bol98], which states that a mediator might actually be caused by the outcome variable. Since the final success outcome of a campaign cannot possibly influence campaign characteristics, reverse causal effects are avoided in this context. Note that we also ruled out the possible direct causal relationships from X to Y by using numbers (1, 2, 3) to represent months instead of using names (e.g., January, February, etc.), thus eliminating possible explanations introduced by month names (for example, charities usually receive less donation in summer months in the real world).

For each participant, we generated 9000 entities in the dataset;

Variable Name	Possible Value	Description
Success	True or False	Whether the project succeeded or not
Category	Art or Food	The category the project belongs to
Country	US or CA	The country the project was launched in
Month	1st, 2nd, or 3rd	In which month the project was launched
Goal	Positive Number	The fundraising goal in USD
Duration	Positive Number	How many days the project lasted
Description Length	Positive Number	How many words in the project essay
Media Shares	Positive Number	How many social media shares of the project

Figure 3: *Description of the variables in our user study's dataset*

each representing a crowdfunding project characterized by the eight variables listed and defined in Figure 3. The six variables other than *Success* and *Month* were randomly assigned to the three causal relation types, i.e., M, M'_X, and M'_Y, in such a way that each type had exactly two variables assigned to it. We randomized the relation type of the variables for each participant to reduce the effect of confirmation bias. For example, since it is common to believe a higher fundraising goal is more difficult to achieve, people might tend to select *Goal* as a mediator more often. By randomly assigning relation types to the variables for each participant, we can ensure the overall performance difference between different variables are due to the relation type but not confirmation bias.

For each of these six variables, we designed two relationships: 1) from X to m and 2) from m to Y, that would be reasonable in reality. We then embedded the two relationships into the dataset according to which relation type each variable was assigned to. A variable M would have both the relationships, while M'_X or M'_Y would only have relationship 1 or 2, respectively. The value of each variable for each entity was generated according to the order of the causal relationships. *Month* was always generated first, using a uniform distribution from 1 to 3. Then, the two M'_X's and the two M's were generated based on the value of *Month* and their predefined causal relationships to it, while the two M'_Y's were generated from a predefined distribution. Finally, *Success* was generated based on the values of the two M'_X's and two M's and their relationships to *Success*.

After the dataset was generated, we ran a series of regressions on it to ensure that the key relationships had been embedded as we intended (all p-values are < 0.001), and that the visual differences in visualizations were perceivable as shown in Figure 2. Specifically, for all pairwise relationships with Month, we ensure high absolute correlation coefficients (mean=0.788, std=0.068) for numerical attributes, and high absolute logistic regression coefficients (mean=0.838, std=0.057) for binary attributes. Details are provided in the supplementary materials.

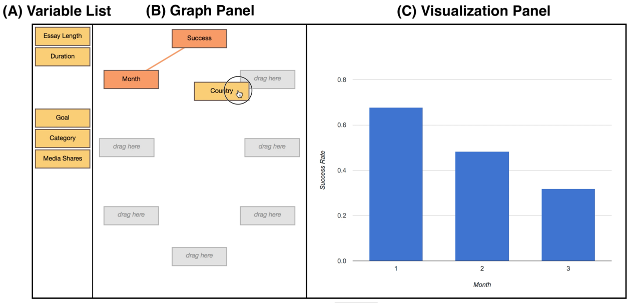


Figure 4: *Interface of the visual analytics system in our studies. The graph panel (B) is an experimental feature that is only available in the treatment group.*

4.2. Visual Analytics System Interface

Figure 4 shows the interface of our visual analytics system, consisting of three components: (A) Variable List, (B) Graph Panel, and (C) Visualization Panel. The first component shows all of the dataset variables, which the user can drag and drop into the grey placeholders in the graph panel. Lines between each pair of variables are drawn automatically. Users can then click on the variable in the graph to include or exclude it from the visualization. They can also click on a line to quickly plot a visualization of the two connected variables. The visualizations are automatically generated based on the selected variables; the generation algorithm can be found in the supplementary materials. In Figure 4, the participant has already dragged in and clicked on *Month* and *Success* (highlighted in orange) to plot the bar chart shown on the right; and is in the process of dragging in a third variable, *Country*, to start a new visualization.

Note that we did not place statistical data, such as error bars or p-values, on the visualizations because a general user is usually not trained in statistics and does not typically use them in such tasks. While the lack of statistics inhibits the ability to statistically verify causal links, statistical testing is not the focus in our exploratory study. Instead, we manipulate the data so that the visual differences are large enough for anyone to perceive a correlation (refer to the supplementary materials for details). Using these visualizations, an expert who is familiar with mediation analysis would be able to identify the most plausible mediators from the dataset using the method described in Sec. 3. In our study, we aim to observe how people search and integrate those findings for causal inference.

The graph panel, Figure 4 (B), is an experimental feature that allows us to test whether interface design influences people's causal-reasoning strategies and performance. Using a graph to represent relationships among variables is intuitive and has been used in many causal reasoning tools [WM16, JL12]. We chose to test this graph panel design as it might benefit the reasoning task in two ways. First, it explicitly draws connections between the variables, which may help externalize their mental models of how the variables are related. Second, it allows them to place the variables on the graph based on their own strategies, which may reduce the cognitive load for memorizing findings. For example, they can drag the variables that they believe are not relevant out of the graph, or keep the most important variables on the top. However, human causal analytic strategies are not always flawless. When using incorrect strategies, the graph panel may exacerbate the potential biases and errors induced by these incorrect strategies.

To test the effect of the feature, the graph panel was hidden so that only (A) and (C) were shown in the control condition, where participants could directly include or exclude variables from the visualization by clicking on the variable list. Because the graph panel does not provide any statistics, the information and visualization space provided by the interface with and without the graph panel are essentially the same, with the only difference being how visualizations are plotted. This reflects the current study's primary interest in the extent to which such variations in interface design can affect human exploration strategies.

4.3. Participants

We recruited participants initially via flyers placed on bulletin boards in multiple areas of a university campus, followed by snowball sampling through social networks, yielding an initial pool of 25 participants. Even though our inclusion criteria included basic visualization-reading ability for bar charts and scatter plots, no professional background in statistics or data analysis was required, as our focus was on how non-experts would use visualizations for causal reasoning. The data of one participant who exhibited difficulties in understanding the task and the visualizations was removed before our analysis.

The remaining participants (13 males, 11 females) were between 20 to 32 years old (median = 23), holding diverse degrees (12 bachelor's, 9 master's, and three doctoral) and programs (10 engineering, 5 business, 5 liberal arts and science, and 4 from other departments). None of them were professional data analyst. The participants were randomly assigned to one of two conditions: 1) the control group, in which the visual analytics system's graph panel was disabled; or 2) the graph group, in which the graph panel was enabled, as shown in Figure 4. We used a between-subjects experimental design to eliminate carry-over effects.

4.4. Procedure

At the beginning of the study, each participant was given an training session on how to use our visual analytics system, using a fictitious admission dataset containing four variables: *Admission, Gender, GPA, and Applied Department*. In this dataset, *Applied Department* is the mediator (M) between *Gender* (X) and *Admission* (Y), while *GPA* is randomly generated and independent of all other variables (not included in Figure 1b).

Two practice tasks were then given to help the participants familiarize themselves with the interface and the structure of the task. The first was to answer the question, *"What is the relation between GPA and Gender?"* This task was used to test whether the user understood the interface and was able to interpret the visualizations. The second practice task was *"Given the data, please determine whether the admission result is affected by gender or department the applicant applied to."* While the correct answer was that only the department affected the admission results, we did not inform them whether they answered correctly or not, avoiding influencing their reasoning strategy. Throughout all of the tasks, the experimenter never hinted whether a participant was approaching the problem in a right or wrong way, but only answered interface-related questions.

We adopted the think-aloud protocol, which encourages participants to say whatever comes into their mind as they are performing a task. This method has been shown to be highly effective in research on human cognitive processes [BJK*16]. The experimenter thus regularly encouraged the participants to say what they wanted to do, why they wanted to do it, and what they were learning from the visualizations. The collected verbal data allowed us to understand their reasoning processes and identify pitfalls, if any.

After the participants had completed their training sessions, they were introduced to the fictitious crowdfunding dataset for the main experimental task, albeit without any clue that it was fictitious, to encourage them use their real-world knowledge to guide their reasoning processes. A printed description of the dataset and its variables was given to each participant. The actual task was presented as follows: *"Based on the data, which variable(s) is the most plausible reason that explains why the success rate of the projects launched in the third month is much lower?"* Rather than revealing that there were two correct answers, however, we merely stated that the number of plausible reasons could be more than one, or none. For the sake of realism, there was no time limit for the task; the participants were told they could end it whenever they felt satisfied with their answers. In practice this took each participant between 5 and 35 minutes.

When the task was completed, the participant needed to confirm to the experimenter which variables they believe explain the relationship between *Month* and *Success*. Therefore, their final answers to the mediation status of the six variables were binary. After the participants confirmed their final answers, a survey was administered, covering their experience of using the visual analytics tool. The entire session lasted from 30 to 60 minutes, and the participants were paid $10 per hour.

5. Results

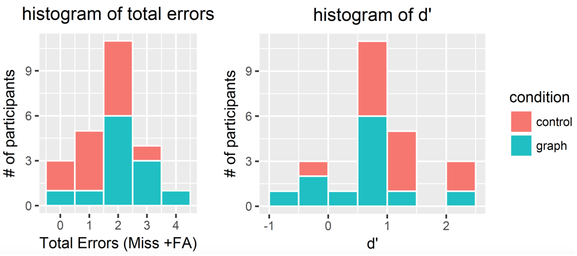
In this section, we first report on the participants' reasoning performance and analyze how it was affected by various factors including 1) the variables' relation types, 2) strategies adopted by individuals, and 3) the interface design (**RQ1**). Second, we describe the general visualization exploration procedures and strategies adopted. Then, we extract and present the common reasoning pitfalls based on audio recording and screen capture (**RQ2**).

5.1. Reasoning Performance

We compared the ground truth against each participant's final answer regarding which variables could explain the relationship between X and Y. Depending on the relation type of m and whether they identify it as a mediator, they might make a hit, a miss, a correct rejection, or a false alarm. The responses are summarized in Figure 5a. Based on signal detection theory [MC04], d' is computed ($d' = Z(hit) - Z(false\ alarm)$) to measure the performance of each participant. The histogram of d' and number of total errors are shown in Figure 5b. On average, a participant made 1.8 errors (median=2), which equates to a correctness rate of 70.1%. Among all errors, 51% are M'_X False Alarms, 35% Misses, and 14% M'_Y False Alarms. Note that the error rate should not be interpreted as the fact that humans are not able to perform mediation analysis well, since visualizations are not enough to verify mediators in practice. However, it serves as a performance metric here that allows us to study how the reasoning performance might be influenced by various factors, as discussed below.

	Control (N=12)		Graph (N=12)		
	Identified as Mediator		Identified as Mediator		Correctness Rate
	Yes	No	Yes	No	
M	18	6	15	9	68.8%
M'_X	9	15	13	11	54.2%
M'_Y	2	22	4	20	87.5%

(a) *The overall correctness rate and responses for each type of the variables being analyzed. Refer to Figure 1b for the definitions of M, M'_X, and M'_Y*



(b) *Histogram of the number of total errors (left) and d' (right) each participant had, colored by condition.*

Figure 5: *The summary of participant performance in terms of (a) overall correctness rate, and (b) number of total errors and d'.*

5.1.1. Effects of Relation Type on Performance.

The correctness rate was 68.6% for M, 54.2% for M'_X, and 87.5% for M'_Y (Figure 5a). In other words, people made many more mistakes when reasoning about M'_X or M as compared to reasoning about M'_Y. We fitted a mixed-effect logistic regression model [BBC*09] to individual responses. The dependent variable is *Correct* (True or False) and the fixed effect terms are *RelationType* (M, M'_X, or M'_Y) and *Condition* (control or graph), which captured the main effect of the interface. Random intercepts for each participant, *UserID*, are also included to capture individual differences. Likelihood ratio tests show a significant main effect for *RelationType* ($\chi^2(2) = 13.9, p < 0.001$). Using M'_Y as the reference group, the effect sizes are $\beta = -1.17, p = 0.03$ for M and $\beta = -1.82, p < 0.001$ for M'_X, which indicates that the participants were less likely to draw correct inferences about M and M'_X, compared to M'_Y. However, their reasoning performance with M and M'_X is not significantly different: when M'_X is set as the reference group in the model, the effect of M is not significant ($\beta = 0.64, p = 0.14$). Please refer to the supplementary materials for detailed statistics.

A higher correctness rate of M'_Y than that of M'_X is expected, as the spurious correlation between M'_X and Y could mislead people to falsely identify M'_X as a mediator. However, contrary to our expectation, the correctness rate of M is also lower than that of M'_Y. These results suggest that not only do people often see a causal relationship when none exists, but also often overlook such causal relationships when they do exist.

5.1.2. Effects of Individual Strategy on Performance

We next tested whether users' different strategies affected their performance. Based on how much time they spent on each visualization category, we identify the **major categories** each participant preferred to use during the task (details will be described in section 5.2). As $V_{X,m,Y}$ was required to identify the spurious relationship, we hypothesized that having $V_{X,m,Y}$ as one of the major categories would lower a person's M'_X False Alarms. On the other hand, as $V_{X,m}$ was required to reject M'_Y being a mediator, we hypothesized that having $V_{X,m}$ as one of the major categories would lower M'_Y False Alarms. To test these hypotheses, we first counted how many false alarms of M'_X and M'_Y each participant made respectively. Since the counts of false alarms do not follow a normal distribution (Shapiro-Wilk Normality test $p < 0.001$), we used Mann-Whitney's U tests to evaluate the differences. Using the number of M'_X False Alarms as the dependent variable (DV) and whether $V_{X,m,Y}$ was one of the major categories (True or False) as the independent variable (IV), the test showed that there is no significant difference ($U = 52, p = 0.34$). However, using the number of M'_Y False Alarms as the DV and whether $V_{X,m}$ was one of the major categories as the IV showed a significant difference ($U = 80.5, p < 0.05$). In other words, participants who plotted $V_{X,m}$ during the task made significantly fewer M'_Y False Alarms, but using $V_{X,m,Y}$ did not help them reduce M'_X False Alarms significantly.

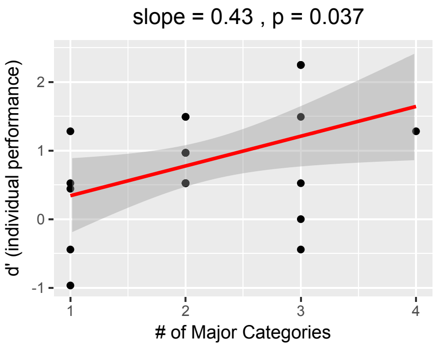


Figure 6: *Participants having more major categories performed better in our task.*

In addition, we found that participants who have a higher number of major categories performed significantly better. As the task performance d' is normally distributed (Shapiro-Wilk Normality test $p = 0.279$), we ran a linear regression on the d' of the participants against the number of major categories they had, which showed a significant effect ($\beta = 0.43, p = 0.037$, Whole model: Adjusted $R^2 = 0.18, F(1,22) = 4.946, p = 0.037$). In other words, the more different categories of visualizations the participants used extensively, the fewer errors they made (see Figure 6). The reason might be that, first, if a participant only focused on one visualization category, that is not enough to identify mediators correctly. Also, when a participant used more categories of visualizations, it might help them collect more evidence, reason more thoroughly, and have more chances to detect underlying data relationships.

5.1.3. Effects of Interface on Performance

We did not find a significant effect of the experimental condition (i.e., control vs. graph) on task performance d', using a linear regression to fit d' on *Condition* ($F(1,22) = 2.57, p = 0.12$). In

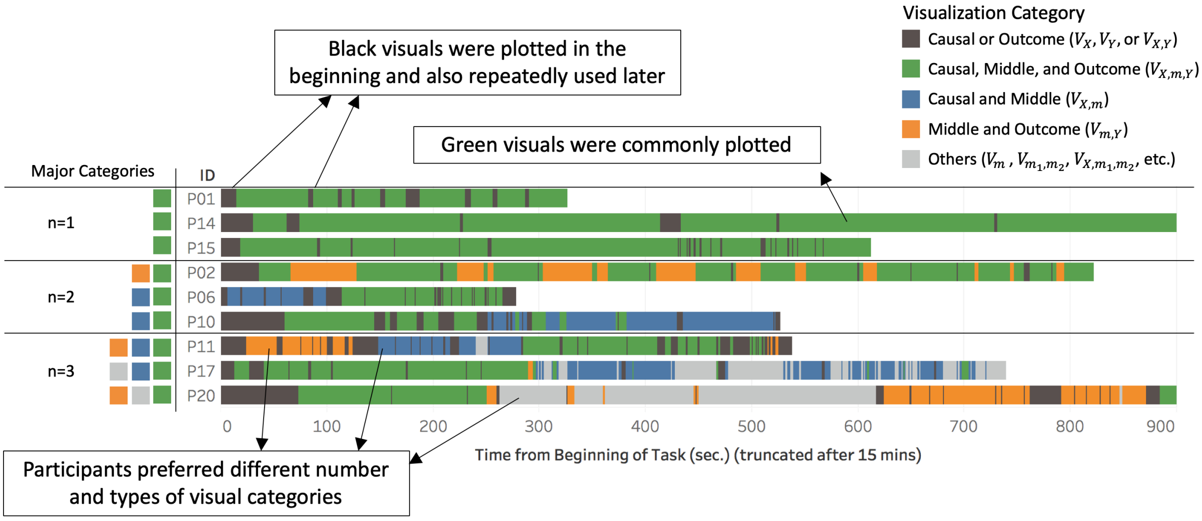


Figure 7: *Nine representative participants' visualization sequences during the task and their major categories. Three participants who had* 1, 2, *or* 3 *major categories were selected to show the similarities and differences in exploration strategies.*

fact, the average of d' for the graph condition (mean=0.59) is lower than that for control (mean=1.11) (Figure 5b). In addition, we found that participants in the graph condition tend to have lower number of major categories than in the control condition (mean=1.9 for graph, 2.4 for control). While the differences are not statistically significant, combined with the results that people having fewer major categories made more errors, this might explain why participants in the graph condition made marginally more errors than those in the control condition. The reason might be that when users already had a relation graph drawn on the interface, it was presumably easier for them to focus quickly on fewer types of relations. However, in the control condition, people had to construct such relation graphs in their minds, which may require more bottom-up processes and motivate them to plot more categories of visualizations. As the effects are not significant in our exploratory study, additional research is required to investigate these hypotheses on interface effects further.

Summary

We found that participants performed significantly better on M'_Y than on M and M'_X. Individual exploration strategies also affected their performance: using $V_{X,m}$ as one of their major categories reduced M'_Y False Alarms, and having more major categories helped them perform better in general. However, adopting $V_{X,m,Y}$ as one of the major categories did not help them reject M'_X as a mediator. We did not find significant effects of the interface feature.

5.2. Visualization Exploration Strategies

To design visual analytics tools that assist causal reasoning effectively, we need to know what common strategies and reasoning pitfalls a person may have during visual exploration. To do so, we first analyzed the visualization sequence for each participant to discover common exploration strategies. Figure 7 shows the overall visualization sequences of nine representative participants in our study (grouped by the number of major categories they had; 3 participants were selected for each group). Each colored block represents a visualization that they were reading during that time, with the color coded based on the category of the visualization. At the start of the task, most participants (19 out of 24) created a visualization using either variable X, or Y, or both, i.e., V_X, V_Y, or $V_{X,Y}$ (black blocks). From our records, among all possible visualizations, $V_{X,Y}$ was used the most frequently by the participants. While we already told them X and Y had a negative correlation before the task, many participants mentioned that they wanted to see whether Y was really decreasing with respect to X, and if so, how large this drop was. Such a strategy helped them confirm what they had learned and become more confident that they were on the right track. Many users re-plotted this type of visualization repeatedly throughout the task, as we will discuss further below.

Next, the participants generally plotted visualizations that included a third middle variable m, such as $V_{X,m}$ (blue), $V_{m,Y}$ (orange), or $V_{X,m,Y}$ (green). They used these visualizations to infer the relationship between variables X, m, and Y. After they reached a conclusion about whether variable m was a mediator, they unselected it and started reasoning about another variable using similar visualization categories. Each participant tended to use just a few categories, though which ones they chose varied widely by individual.

For example, as shown in Figure 7, participants 1, 14, and 15 only used $V_{X,m,Y}$, the visualizations that includes causal, middle, and outcome variables, as their main visualization category. On the other hand, besides using $V_{X,m,Y}$, participant 2 also often plotted the outcome variable with one middle variable, $V_{m,Y}$. As for participant 11, she first plotted the outcome variable with each middle variable ($V_{m,Y}$), then the causal variable with middle variables ($V_{X,m}$), and finally $V_{X,m,Y}$ over the course of the task.

The results show that while participants followed a similar procedure in general, people have individual preferences regarding which categories of visualizations to plot. To determine the major

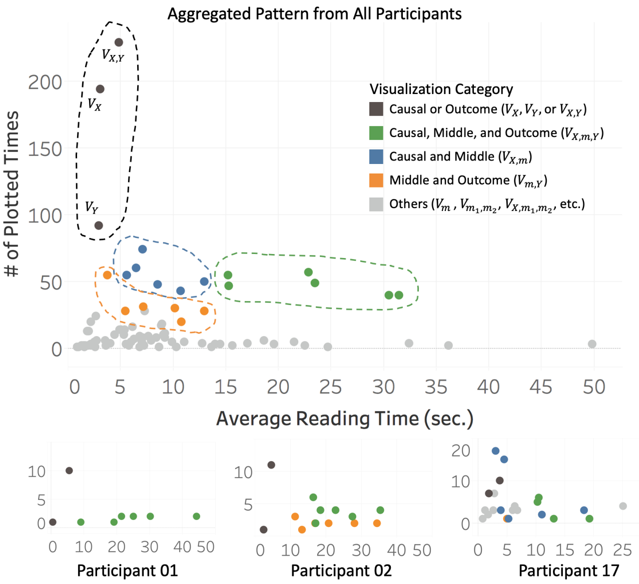


Figure 8: *Scatter plots of how many times a visualization is plotted and the average time users spent reading it per plotting. The top scatter plot shows the aggregated pattern; in the bottom, three participants' behavior pattern are plotted for comparison.*

categories each participant preferred to plot with middle variables, we first exclude the time they spent on causal or outcome visualizations (V_X, V_Y, or $V_{X,Y}$). Then, we calculated the percentages of time each user spent on each visualization category. We labeled the categories from highest percentage to lowest percentage as **major categories** until the sum of the labeled time exceeds 75%, or when the next percentage of the category is lower than average. Following this method, we identified the major categories for each participant. In Figure 7, the major categories of each participant were shown on the left side. We acknowledge that this simple rule may not be generalizable to future studies; however, the results fit to the data in our study by visual inspection as shown in the figure.

Certain categories were more frequently used by our participants as a major category: $V_{X,m,Y}$ was used by 20 participants, $V_{X,m}$ was used by 17 participants, with much fewer participants using $V_{m,Y}$ (n=6) and other visualization categories. As discussed in Section 3.1, $V_{X,Y}$, $V_{X,m}$ and $V_{X,m,Y}$ are the three key visualizations to verify a mediator. Interestingly, although our participants did not have extensive data analytics experience, their overall strategy aligned well with the discussed procedure in Section 3.1. However, at the individual level, most of our participants' strategies were not desirable. Fifteen (63%) participants did not use both $V_{X,m}$ and $V_{X,m,Y}$ as major categories. And, six participants (25%) did not use any Y-related visualizations (i.e., $V_{X,m,Y}$, $V_{m,Y}$, and $V_{m1,m2,Y}$); in other words, they only reasoned about m's relation to X, without seeking evidence that m could affect Y. This implies that the participants generally paid more attention to finding relationships between X and m, even though understanding those between m and Y was equally important for the successful completion of the task.

Figure 8 further illustrates how users behaved differently when using different visualizations. In the scatter plot, the Y-axis shows how many times a specific visualization was plotted across all participants, while the X-axis shows the average time (in seconds) a user spent on a visualization per plotting; the points are colored according to the visualization category. Note that each point represents one specific visualization, so multiple visualizations may belong to the same category. The scatter plot on the top shows aggregated patterns across all participants. The top-left area includes $V_{X,Y}$, V_X, and V_Y, which were plotted the most times. However, users usually spent little time examining these visualizations (less than 5 seconds per plot). The second most frequently plotted categories were $V_{X,m}$ (blue) and $V_{X,m,Y}$ (green), which users spent more time interpreting compared to V_X, V_Y, or $V_{X,Y}$. Within this pair, $V_{X,m,Y}$ required much more time to interpret (from 15 to 32 seconds in this group) than $V_{X,m}$ (from 5 to 13 seconds in this group); this is not unexpected, given that the former contains more variables and thus requires more effort to interpret. For the group of $V_{m,Y}$ visualizations (orange), users spent roughly the same amount of time as they did on $V_{X,m}$, but created fewer plots compared to $V_{X,m}$. These results again imply that users generally focused more on the relationship of X to m than on that of m to Y.

Summary

Overall, the participants showed similar exploration strategies, plotting causal or outcome variables the most frequently ($V_{X,Y}$, V_X, and V_Y), both at the beginning of the task and repeatedly over its later phases. The repeated plots often served as breaks during the overall task, lasting less than 5 seconds, and seem to have functioned to separate the reasoning processes that were being applied to different variable m's. Then, each participant would pick a new variable m and plot it using certain visualization categories. Our results show that participants adopted different strategies in terms of how many, and which categories to use. While $V_{X,m,Y}$ and $V_{X,m}$ were used frequently in general, at the individual level only 38% of participants used both as major categories. Therefore, future research may focus on designing interfaces that assist users to adopt better strategies based on their visualization sequences.

5.3. Reasoning Pitfalls

To better understand why participants made errors, we used their verbal data and screen capture videos to identify common pitfalls. Apart from the reason that they did not plot the necessary visualizations for analysis, we are also interested in understanding why, when those visualizations were indeed plotted and examined, some participants still made incorrect inferences, i.e., missed M and falsely identified M'_X as a mediator.

5.3.1. Miss

Among the 15 misses made by the participants collectively, 9 occurred when they did plot $V_{X,m,Y}$ for the variable. One common misinterpretation arose because, when a user plotted $V_{X,m,Y}$, the pattern of variable m versus Y was similar for all values of X, which gave them an impression that the variable m had no overall impact. Such false impression occurred especially when the participant moved from one visualization to another, as an example shown in Figure 9. During the transition from $V_{m,Y}$ to $V_{X,m,Y}$, the participant was apparently hoping to see a substantial visual change in the

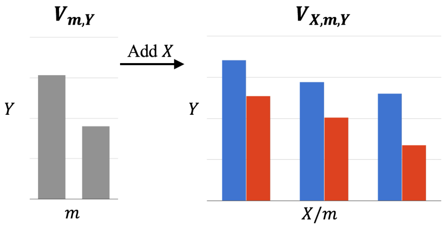


Figure 9: *An example of when a participant missed recognizing a mediator due to false impression.*

chart, while the pattern was similar except being separated to three groups. Such an impression made the participants conclude that the m was not important after seeing the plot.

Another mode of misinterpreting $V_{X,m,Y}$ was for the user to note that when the value of variable m remained unchanged, the relationship between Y and X still existed. For example, as shown in Figure 9, Y decreased with respect to X even when m is fixed (comparing the bars with the same color); this led some participants to reason that there must be another factor that caused the drop. While it was true that there was another mediator, this did not automatically imply that the variable currently being scrutinized could not itself be one of several mediators. Nevertheless, some people rejected the variable too early based on this observation.

5.3.2. M'_X False Alarm

Overall, M'_X was misidentified as a mediator 22 times by 16 different participants. In two-thirds of these cases of M'_X False Alarm (15 out of 22) the $V_{X,m,Y}$ was indeed plotted by the participants, suggesting that the main pitfall was not failing to plot it, but rather failing to detect that the plotted relationship is spurious.

We observed two common reasons that participants failed to reject M'_X. First, when the visualization showed a strong trend between X and Y for each subgroup of m, this salient pattern captured their attention and fit into their overall hypotheses, automatically providing them with a sense of confirmation. Thus, without further reasoning of the relationship between m and Y, they jumped to the conclusion that m was the important factor, and did not check whether such m-Y relation still exists when X is controlled. Second, the noise in a visualization such as outliers often creates some random patterns. Some participants would consider those patterns as evidence of the relationship between m and Y, which actually did not exist.

6. Discussions and Limitations

Our results demonstrate that the correctness of human causal inference is influenced by the relation type of the variable being analyzed and by individual exploration strategies. We also reported the common strategies and reasoning pitfalls identified in our study. While visualizations are not the best mechanism for humans to statistically verify mediators, they are powerful tools for forming initial hypotheses and are widely used by both experts and nonexperts. Therefore, there is a strong need to provide more support for people to better interpret and understand their data through visualizations. Based on our results, we categorize the common sources of error in visual causality analysis below and provide design recommendations to address these issues.

6.1. Human Error in visual causality analysis

6.1.1. Missing critical visualizations

Our analytic framework explained some of the errors made by the participants; for example, failing to plot the key visualizations would lead to different types of errors due to insufficient evidence. Without seeing the key visualizations, there is no way to differentiate the mediators and non-mediators. This type of error results in 6 out of 15 misses and 7 out of 22 M'_X False Alarms in our study. In addition, only 38% of the participants used both of the key visualizations $V_{X,m}$ and $V_{X,m,Y}$ for reasoning throughout the task. Confirmation bias may also contribute to these errors in a way that once supporting evidence is collected, people tend to not seek additional evidence that may refute their hypothesis.

6.1.2. Erroneous visualization interpretation

Our results show that, in fact, most of the incorrect answers were made when key visualizations were plotted. Therefore, in addition to suggesting key visualizations, visual analytics systems should also assist interpretation effectively. A common reason for erroneous interpretation was that the false impression from the visual patterns might make users jump to false conclusions quickly, as described in Sec. 5.3.1 and 5.3.2. Also, human perception of correlations does not always align with statistics [SBLC17]. A random pattern might be perceived as a correlation, and vice versa.

6.1.3. Inability to incorporate all visual evidence

Even when the key visualizations are charted and interpreted correctly, people might still fail to incorporate all of the visual evidence and infer the reasonable causal model, especially when certain visual patterns are conflicting. For example, Simpson's Paradox is one of the well-known examples that is considered difficult to comprehend by human analysts [AW14]. Some participants expressed their frustration when such visual patterns occurred in our study. Further studies are necessary to understand the reasoning process and how to help users resolve the confusion.

Since this is an exploratory study, we do not believe this is an exhaustive list of human errors that might occur during visual causality analysis. However, it sheds light on what type of errors might occur in each reasoning step and helps motivate the design implications as follows.

6.2. Design Implications

First, to help avoid the errors discussed in Sec. 6.1.1, a system could suggest under-explored visualizations using action logs. As shown in Figure 8, the reasoning goals, i.e., which variable relationship(s) a person wants to explain, are directly related to 1) how often a person plots a specific visualization, and 2) how much time is then spent reading it. Therefore, it is eminently possible to use log data alone to detect what target relationship a user is attempting to deduce. As the target relationship might continuously change during exploratory data analysis, auto-detecting the target relationship could save additional input effort from the analysts and avoid distraction.

With the information of user reasoning goal and context, more effective guidance can be provided. For example, a system could conduct proper statistical tests to detect possible mediators and encourage users to explore those variables. Moreover, by analyzing the behavior pattern as shown in the scatter plots (Figure 8), the system could prompt users to plot important visualization categories such as $V_{X,m}$, or $V_{X,m,Y}$, if they have not been visualized.

Second, to alleviate the errors in Sec. 6.1.2 and 6.1.3, a visual analytics system could provide high-level, statistics-backed text summaries of data patterns to assist interpretation. Moreover, if a reasoning context has been fed into the system, it could provide more direct inferences. For example, on the right-hand side of Figure 9, a context-aware help message could be "m is a potential explanation of the relationship between X and Y". More importantly, the system could also remind users about what could *not* be inferred from the data: for example, "The decreasing Y for both blue and red bars suggests that there should be reasons other than m for the decline of Y, but this does not rule out m as part of the reason." Such high-level reasoning guidance would undoubtedly help reduce human errors induced by false impressions or reasoning pitfalls.

The integration of visualization and statistical analysis is an important research area, as both are necessary to infer causality in practice. For example, visualization allows users to quickly check if there are any anomalies in the data and to ensure the correctness of statistical mediation analysis. Further studies are needed to understand how visual analytics systems should be designed in this context.

6.3. Limitations

Several limitations of this study should be noted. First, while our analytic framework can be used to evaluate other systems when users are performing similar causal reasoning as in our study, its causal model was relatively simple compared to most real datasets, which may contain dozens or hundreds of variables with highly complex interrelationships. In addition, visualizations may become ineffective when too many variables are involved. Therefore, the scope of our analytic framework may not easily extend to large and complex datasets. It may not be suitable for some causal analysis applications as well, when the underlying causal network is not well defined in a task; for example, reasoning the causes of an accident with videos or text.

Second, human reasoning processes are influenced by the dataset and the design of visual analytics systems, which can limit the study's generalizability. For instance, results might vary based on the interface and the visualization techniques made available to the study participants. The visualizations used in our study are basic infographics, but there could be other visualization techniques that could further reduce human errors. Also, people may behave differently as the sizes of their datasets increases, such as relying more on initial hypotheses, thus changing their exploration strategy.

Third, in our study, a number of participants (10 out of 24) are from engineering departments and the average age is relatively low (mean=23), which might impact the generalizability of our findings. While they are appropriate for our study as we focus on nonprofessional data analysts, people with different backgrounds or age groups might behave differently from our results.

7. Conclusion

While visualizations are not enough for humans to perform thorough mediation analysis, they are useful for initial hypothesis formation and are widely used by non-experts. In this study, we aim to understand how visual analytics systems can be designed to assist visual causality analysis better. We conducted an exploratory user study to provide empirical evidence regarding an individual's visual causality analysis performance, and how such performance was impacted by the causal relationships of the variables being analyzed; by their exploration strategies; and by the interface. We recruited 24 participants and found that they were more likely to make incorrect inferences when the variable being analyzed is directly influenced by a confounding variable. We also showed that the participants adopted varied visualization exploration strategies, which in turn affected their performance. An experimental interface feature was tested and the results suggest that interface design may affect exploration strategies, while no significant effect on performance was found. We then identified the common exploration strategies and reasoning pitfalls and discussed design implications on how future visual analytics systems can assist causal inference more effectively.

8. Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions.

References

[AW14] ARMSTRONG Z., WATTENBERG M.: Visualizing statistical mix effects and simpson's paradox. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2132–2141. 2, 10

[BBC*09] BOLKER B. M., BROOKS M. E., CLARK C. J., GEANGE S. W., POULSEN J. R., STEVENS M. H. H., WHITE J.-S. S.: Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution 24*, 3 (2009), 127–135. 7

[BJK*16] BLASCHECK T., JOHN M., KURZHALS K., KOCH S., ERTL T.: Va 2: a visual analytics approach for evaluating visual analytics applications. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 61–70. 3, 6

[BK86] BARON R. M., KENNY D. A.: The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology 51*, 6 (1986), 1173. 1, 2, 3

[Bol98] BOLGER N.: Data analysis in social psychology. *Handbook of social psychology 1* (1998), 233–65. 4

[Bur05] BURNETT R. C.: Close does count: Evidence of a proximity effect in inference from causal knowledge. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2005), vol. 27. 2

[cKFY11] CHUL KWON B., FISHER B., YI J. S.: Visual analytic roadblocks for novice investigators. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 3–11. 3

[CLAR91] Cummins D. D., Lubart T., Alksnis O., Rist R.: Conditional reasoning and causation. *Memory & cognition 19*, 3 (1991), 274–282. 2

[CM84] Cleveland W. S., McGill R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association 79*, 387 (1984), 531–554. 2

[CS08] Cook M. B., Smallman H. S.: Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes. *Human Factors 50*, 5 (2008), 745–754. 3

[CT92] Cosmides L., Tooby J.: Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture 163* (1992), 163–228. 2

[DC17] Dabek F., Caban J. J.: A grammar-based approach for modeling user interactions and generating suggestions during the data exploration process. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 41–50. 3

[DJS*09] Dou W., Jeong D. H., Stukes F., Ribarsky W., Lipford H. R., Chang R.: Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications 29*, 3 (2009). 3

[EFN12] Endert A., Fiaux P., North C.: Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2879–2888. 3

[FSM11] Fiedler K., Schott M., Meiser T.: What mediation analysis can (not) do. *Journal of Experimental Social Psychology 47*, 6 (2011), 1231–1236. 1

[GBK17] Guo Y., Binnig C., Kraska T.: What you see is not what you get!: Detecting simpson's paradoxes during data exploration. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (2017), ACM, p. 2. 2

[GGZL16] Guo H., Gomez S. R., Ziemkiewicz C., Laidlaw D. H.: A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 51–60. 3

[GSC16] Gotz D., Sun S., Cao N.: Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (2016), ACM, pp. 85–95. 3

[GTS10] Grammel L., Tory M., Storey M.-A.: How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics 16*, 6 (2010), 943–952. 3

[GZ09] Gotz D., Zhou M. X.: Characterizing users' visual analytic activity for insight provenance. *Information Visualization 8*, 1 (2009), 42–55. 3

[JL12] Jianu R., Laidlaw D.: An evaluation of how small user interface changes can improve scientists' analytic strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, pp. 2953–2962. 3, 5

[JSVDBG04] Jaspers M. W., Steen T., Van Den Bos C., Geenen M.: The think aloud method: a guide to user interface design. *International journal of medical informatics 73*, 11-12 (2004), 781–795. 3

[KLK18] Kong H.-K., Liu Z., Karahalios K.: Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 438. 3

[LDH*19] Lee D. J.-L., Dev H., Hu H., Elmeleegy H., Parameswaran A.: Avoiding drill-down fallacies with vispilot: assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019), ACM, pp. 186–196. 2

[LH14] Liu Z., Heer J.: The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics 20*, 12 (2014), 2122–2131. 2

[LSD*10] Lipford H. R., Stukes F., Dou W., Hawkins M. E., Chang R.: Helping users recall their reasoning process. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on* (2010), IEEE, pp. 187–194. 3

[MBG*14] Morton K., Balazinska M., Grossman D., Kosara R., Mackinlay J.: Public data and visualizations: How are many eyes and tableau public used for collaborative analytics? *ACM SIGMOD Record 43*, 2 (2014), 17–22. 1

[MC04] Macmillan N. A., Creelman C. D.: *Detection theory: A user's guide*. Psychology press, 2004. 6

[MHS07] Mackinlay J., Hanrahan P., Stolte C.: Show me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics 13*, 6 (2007). 1

[Nic98] Nickerson R. S.: Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology 2*, 2 (1998), 175. 3

[OC94] Oaksford M., Chater N.: A rational analysis of the selection task as optimal data selection. *Psychological Review 101*, 4 (1994), 608. 2

[OPH*16] Ottley A., Peck E. M., Harrison L. T., Afergan D., Ziemkiewicz C., Taylor H. A., Han P. K., Chang R.: Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 529–538. 2, 3

[Pea09] Pearl J.: *Causality*. Cambridge university press, 2009. 1, 2

[SBLC17] Sher V., Bemis K. G., Liccardi I., Chen M.: An empirical study on the reliability of perceiving correlation indices using scatterplots. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 61–72. 10

[SGS00] Spirtes P., Glymour C. N., Scheines R.: *Causation, prediction, and search*. MIT press, 2000. 1

[SNEA*16] Schulz C., Nocaj A., El-Assady M., Frey S., Hlawatsch M., Hund M., Karch G., Netzel R., Schätzle C., Butt M., et al.: Generative data models for validation and evaluation of visualization techniques. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (2016), ACM, pp. 112–124. 4

[SSBZ00] Staller A., Sloman S. A., Ben-Zeev T.: Perspective effects in nondeontic versions of the wason selection task. *Memory & Cognition 28*, 3 (2000), 396–405.

[VZS18] Valdez A. C., Ziefle M., Sedlmair M.: Priming and anchoring effects in visualization. *IEEE Transactions on Visualization & Computer Graphics*, 1 (2018), 584–594. 3

[Was68] Wason P. C.: Reasoning about a rule. *Quarterly journal of experimental psychology 20*, 3 (1968), 273–281. 2

[WM] Wang J., Mueller K.: Visual causality analysis made practical. 1, 2

[WM16] Wang J., Mueller K.: The visual causality analyst: An interactive interface for causal reasoning. *IEEE transactions on visualization and computer graphics 22*, 1 (2016), 230–239. 1, 2, 5

[Wri21] Wright S.: Correlation and causation. *Journal of agricultural research 20*, 7 (1921), 557–585. 2

[WS04] Walsh C. R., Sloman S. A.: Revising causal beliefs. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2004), vol. 26. 2

[YHR*18] Yang F., Harrison L., Rensink R. A., Franconeri S., Chang R.: Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics* (2018). 2

[ZZZK18] Zgraggen E., Zhao Z., Zeleznik R., Kraska T.: Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 479. 2