Causality-Based Visual Analysis of Questionnaire Responses

Renzhong Li, Weiwei Cui, Tianqi Song, Xiao Xie, Rui Ding, Yun Wang, Haidong Zhang, Hong Zhou, and Yingcai Wu

Abstract—As the final stage of questionnaire analysis, causal reasoning is the key to turning responses into valuable insights and actionable items for decision-makers. During the questionnaire analysis, classical statistical methods (e.g., Differences-in-Differences) have been widely exploited to evaluate causality between questions. However, due to the huge search space and complex causal structure in data, causal reasoning is still extremely challenging and time-consuming, and often conducted in a trial-and-error manner. On the other hand, existing visual methods of causal reasoning face the challenge of bringing scalability and expert knowledge together and can hardly be used in the questionnaire scenario. In this work, we present a systematic solution to help analysts effectively and efficiently explore questionnaire data and derive causality. Based on the association mining algorithm, we dig question combinations with potential inner causality and help analysts interactively explore the causal sub-graph of each question combination. Furthermore, leveraging the requirements collected from the experts, we built a visualization tool and conducted a comparative study with the state-of-the-art system to show the usability and efficiency of our system.

Index Terms—Causal analysis, Questionnaire, Design study

1 Introduction

Questionnaires are a useful tool to collect quantitative information from individuals and widely used in scientific studies. Via online tools like SurveyMonkey [2] and Microsoft Forms [1], analysts can distribute questionnaires and gather responses at a low cost and high efficiency. Once responses are collected, analysts adopt various analysis methods (e.g., correlation analysis, causal reasoning) to explore data, discover causality between responses, and eventually derive insights for decision-making. However, due to the huge search space and complex causal structure in data, causal reasoning, as the final and critical step in a typical survey analysis [30], is still extremely challenging and time-consuming, and often conducted in a trial-and-error manner.

Causal reasoning in a typical questionnaire analysis is mainly based on hypothesis and regression analysis [6]. Firstly, analysts hypothesize the answers to several questions are the reasons for the answer to a target question on the ground of their prior knowledge and calculated correlation coefficient between questions. Secondly, they choose different regression methods (e.g., Regression Discontinuity Design [23], Differences-in-Differences [9]) according to the sample size of data and type of questions to test the hypothesis. Thirdly, they determine whether they have found the correct reasons via the p-value calculated by these methods. Such a pipeline is direct and easy to understand. However, the repercussions are also obvious. First of all, it is practically impossible for analysts to exhaust the huge search space involving different combinations of questions. Hence, it is desired to have a method to organize the space and, more importantly, to provide an intuitive and explainable way to explore and find promising question combinations to drill in. On the other hand, focusing on reasons for the target question may neglect the global causal structure of all questions, leading to getting lost in the mediating variables and indirect causes.

In the visualization community, causal reasoning is also an ongoing topic. Existing works on analytic systems [21, 39, 44] mainly focus on visualizing the global causal structure and encoding it with Directed Acyclic Graph (DAG). Specifically, the DAG design represents questions with nodes and causal relationships with directed edges. This encoding is successfully applied to many domains, ranging from computational fairness [21] to pollution control [16, 17] and clinical decision [32]. However, one of the primary drawbacks of DAGs is visual clutter, which limits their effectiveness when applied to large datasets [24]. While Xie et al. [44] proposed a new hierarchical layout that partially mitigates visual clutter by hiding certain edges, applying DAGs to questionnaire scenarios still poses several challenges. Firstly, for comprehensive questionnaires with dozens or hundreds of questions, The DAG can have serious crossover problems, which even the design of Xie et al. is insufficient to address. Secondly, analysts often possess domain knowledge about the questionnaire, which should be leveraged to visualize a more accurate causality. Wang and Mueller [39] summarized prior knowledge that could impact causal structure. However, the constraints they summarized may not be comprehensive enough for questionnaire analysis. For example, analysts may explicitly specify that the answer to gender cannot be the effect of any question, which is beyond the scope of their summary. The support of these needs of questionnaire experts has not been sufficiently studied in existing work. Thirdly, as the number of questions grows, the DAG structures before and after adding/removing constraints will have a great difference. This creates an unavoidable obstacle for analysts, who must frequently re-orient themselves to the changing DAG during the exploration process.

In this paper, we propose Questionnaire Explorer [1] (QE), a novel visual analytic system that helps experts explore the causal structure of single-choice questions in questionnaire data efficiently and explainably. To address the scalability issue of existing works, we introduce a causal reasoning pipeline based on the sub-graph structures in the whole DAG. First, we propose an association mining algorithm to find explainable question combinations which are more likely to have a direct causal relationship. The various question combinations serve as starting point of exploration, which are summarized and visualized in a matrix-based view. Once users have identified questions of interest, our system extracts and visualizes the most relevant sub-graph about causality for analysts to explore.Furthermore, a series of interactions are designed crossing the matrix view and graph view to ensure that the analysts are not limited to a specific sub-graph and can flexibly explore the whole dataset. To evaluate the usefulness of QE, we conduct a comparative study with the state-of-the-art system [44]. The result shows that QE can significantly improve analysis efficiency and user experience on large-scale questionnaires. The contributions are as follows:

- We propose an analysis pipeline based on sub-graphs to improve the scalability of casual graph analysis.
- We build a visual analytic system to enable users to interactively explore, comprehend, and conduct causal analysis tasks in questionnaire dataset.
- We evaluate our approach with usage scenarios and a controlled user study.

- *R. Li, T. Song, and Y. Wu are with the State Key Lab of CAD&CG, Zhejiang University. E-mail: {renzhongli, holly1027, ycwu}@zju.edu.cn.*
- *W. Cui, R. Ding, Y. Wang, and H. Zhang are with the Microsoft Research Asia. E-mail: {weiweicu, juding, wangyun, haizhang}@microsoft.com.*
- *X. Xie is with the Department of Sports Science, Zhejiang University. E-mail: xxie@zju.edu.cn. X. Xie is the corresponding author.*
- *H. Zhou is with the College of Computer Science and Software Engineering, Shenzhen University. E-mail: hzhou@szu.edu.cn.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

[1] https://github.com/evenlasting/Questionnaire-Explorer

2 Related Work

2.1 Causal Reasoning in Questionnaire Analysis

Questionnaires are one of the most useful tools for analysts to gather information. Once information is gathered, causal reasoning is an essential follow-up to convert data into insights. The workflow of causal reasoning in questionnaire analysis can be divided into two steps [6], i.e. *hypothesis generation* and *hypothesis test*.

Hypothesis generation. At the beginning of the workflow, analysts need to find a causal structure, including a target question and several dependent questions from the whole data set. This unproven causal structure is called hypothesis in the questionnaire domain. In the next step, the *hypothesis test*, analysts can verify the hypothesis with the help of statistical methods and obtain a final conclusion. The finding process of the causal structure is mainly based on the semantic meaning of each question and the correlation between the target and dependent questions. The process is direct and easy to understand. However, the disadvantages are also obvious. First, the hypothesized causal structure is often incomplete and neglects the global causal structure of all questions, leading to getting lost in the mediating variables and indirect causes. Second, the process is highly dependent on the background knowledge of analysts and is often stochastic. Finally, it is inefficient to search in the dataset of a questionnaire by enumeration. Since any number of questions may have an inner correlation, the search space is obviously too large for trial-and-error.

Some heuristic algorithms [11, 13, 28] have been proposed to deal with the inefficiency challenge. They could recommend combinations of relevant questions automatically by association mining algorithms. Hence, analysts can perform hypothesis generation on a smaller exploration space. Although reduced in size, the exploration space computed by algorithms may also contain numerous fuzzy associations and lacks an efficient method to explore relationships of corresponding questions. Therefore, the analysts need additional support or interpretations of the found question combinations to help them understand.

Our work is dedicated to solving the above challenges. Via an explainable association mining algorithm, we help analysts understand the obtained problem combinations. At the same time, our work proposes a visual analytic system to enable users to balance the exploration of question combinations and the automatically-generated global causal structure, aiming to better integrate users' expert knowledge, association mining algorithms, and causal reasoning algorithms.

Hypothesis test. Since this step is not the focus of our work, we just briefly introduce the related test methods. After generating hypotheses based on expert knowledge and the correlation coefficients, analysts need to test whether the causal structures are real. The most effective way to test the hypothesis is through randomized experiments [6]. However, randomization is usually too expensive to perform or completely impossible. Therefore, many non-experimental tests on the ground of regression are proposed to test the causality between questions (also called variables at this step). The analysts may pick different methods depending on the nature of questions and the dataset size. For example, Differences-in-Differences [9] can be used for binary variables that received an exogenous treatment, while regression discontinuity design [23] is more suitable for continuous variables with a modeled cut-off. Our visual analysis system (QE) does not focus on any specific test method but on the general scalability problem when applying any method in a questionnaire analysis of a large number of questions.

2.2 Visualization of Causality

Causality is not only important in questionnaire analysis, but also a basic topic for visual analytic researchers. The study of causality in the visualization area has its root on what is the best method to visualize causality. Empirical studies [8, 26, 45, 46] have tested the performance of different charts or animations in expressing causality. Deng et al. [17] summarized that many basic charts (e.g., line charts, bar charts) are not good choices. Bae et al. [7] recommended using node-link diagrams to visualize causality and summarized several design guidelines for effective analysis. These studies provide a solid foundation for the visualization of causality as well as the design of our system.

On this basis, researchers have developed various visual analytic systems to assist users in exploring the causality across different datasets, such as time series data [12, 25] and tabular data [44, 47]. Questionnaire data is a typical form of tabular data. Therefore, we focus on the systems with the input of tabular data in this section. Wang and Mueller [39] proposed the very first system that bridges the causality mining algorithm (e.g., PC [36], F-GES [35]) and visualizations. Users are enabled to interactively explore the automatically-generated causal structure and can also add basic causal constraints (e.g., variable A is the cause of variable B) via their priori knowledge. Furthermore, they solve the challenge of exploration and comparison of causality within different subgroups [40]. The follow-up systems [24, 44] focus on assisting users in understanding the model and performing counterfactual analysis. All of these systems are based on DAG visualizations, which are widely applied in numerous visual analytics systems [27, 31]. With the main drawback of visual clutter, DAGs hamper the scalability of the abovementioned systems. Xie et al. [44] proposed a new hierarchical layout that selectively hides edges in order to manage the visual complexity. However, their method still visualizes the entire dataset, which can result in severe cluttering particularly for questionnaire datasets with dozens or even hundreds of questions.

To alleviate the scalability issue of existing causality visualizations, this work proposes a visual analysis pipeline based on sub-graphs of the whole causal structure. We also conduct a comparative study with the state-of-the-art system of Xie et al. [44] to evaluate the usefulness of such a pipeline and receive positive feedback.

3 Informing the Design

We have collaborated with data analysts from a reputable technology company to conduct this study. The employees and customers of this company fill out a large number of questionnaires every year. However, the current method of questionnaire analysis is mainly based on enumeration and statistical tests (see Sec. 2.1). It often takes one or two weeks to obtain an analysis report of a questionnaire (e.g., [41]). Causal reasoning, as the last and one of the most essential components of questionnaire analysis, always takes up a large portion of the time spent. In this section, we introduce an interview study to collect the design needs on causal reasoning of analysts, which drives our model and visualization designs.

3.1 Participants and Process

We invited four domain analysts from a technology company. Two of them are experts in conducting surveys, who are interested in the quick generation of causal structure, as it may contribute to completing the analysis report more efficiently (P1-2). The other two are scientists who specialize in causality analysis and have more than five years of experience in modeling variables and discovering causality (P3-4).

We conducted two semi-structured interviews with experts in each domain respectively, allowing experts within the same domain to complement each other and avoiding divergent discussions among experts from different domains. Each interview is introduced by a questionnaire report that describes the basic analysis process [41]. In the first interview, we asked all experts to share their experience with questionnaire analysis, specifically focusing on causal reasoning, tools used for causality reasoning, and the challenges they faced in their daily work. We encouraged analysts to provide examples to illustrate the difficulties they encountered. Afterward, we summarized all the requirements and conducted a follow-up interview in which we asked the experts to design possible algorithmic solutions for questionnaire analysts based on the identified challenges.

3.2 Requirement Analysis

According to the interviews, we summarized seven critical needs. At the beginning of the interview, we discussed with experts how to increase the efficiency of the causal reasoning process and whether it is feasible to achieve this by providing an automatically-generated causal graph of the whole dataset. During the discussion, questionnaire analysts commented that showing the whole causal structure is quite good for their current analysis pipeline. However, they were unfamiliar with either causal graph visualization techniques or causal automated mining algorithms. Since the causal reasoning pipeline in questionnaire analysis is well studied, they wanted us to design some auxiliary tools around their current pipeline (Sec. 2.1). More specifically, the question combinations with possible causality should be the core concept of the design. Their current approaches to generating such question combinations include enumeration and association mining. For the sake of efficiency, it is critical for our system to **perform association mining algorithm automatically to dig question combinations with potential causality (N1)** instead of enumeration. However, current association mining algorithms showed weakness in supporting the semantic interpretation and the efficient exploration of the result. Therefore, **these combinations (associations) should be explainable (N2)** and **can be orderly explored via the help of quantitative indicators (N3)**.

Moreover, questionnaire analysts commented that the causal mining algorithm could be applied to a selected combination to help them identify the mediating or indirect variables within the question combination. At this point, the causality analyst (P3) came up with an idea ***"why not cut out the part of the global causal graph which contains the question combination they are interested in?"*** **(N4)** This idea balanced the experts' analytical habits [37] with the exploration of global causality and was agreed upon by all. Meanwhile, P3 also mentioned that automated algorithms for causal reasoning could not be fully trusted. Models need to be adjusted manually based on users' prior knowledge. To be specific, the system requires **interactive provision (N5)** and visualization of the causal relationship between selected questions.

The identified needs can assist analysts in discovering several reasons for a target question. Experts also pointed out that questionnaires can contain many valuable causal patterns beyond just finding the reasons for a target question. For instance, in a survey on remote work policies, analysts may discover that childcare and personal relationships significantly influence an employee's choice of workplace. There may also be other valuable causal relationships, such as the effect of age on productivity and the effect of personal relationships on job satisfaction. Therefore, analysts would like to explore the data without specific target questions, discovering new and unexpected patterns. At this stage, question combinations showing high relevance to others are likely to pique analysts' interest. Hence, it is necessary to **visualize the relationships between different question combinations (N6)**, as the stronger the relationship between a combination of questions, the more likely it is to be at the center of the causality. This type of question combination could serve as a good starting point for free exploration.

At last, after the exploration of a causal structure, analysts would take other factors into consideration, such as wondering whether having children performs the same influence on the younger and older employees' preferences. This requires the system to support **selecting a sub-group of respondents for further analysis (N7)**.

4 Association Modeling

In this section, we introduce the formal definition of association rules [5] (Sec.4.1) and propose a new computing method to recommend explainable question combinations by association aggregation (N1, N2) (Sec.4.2).

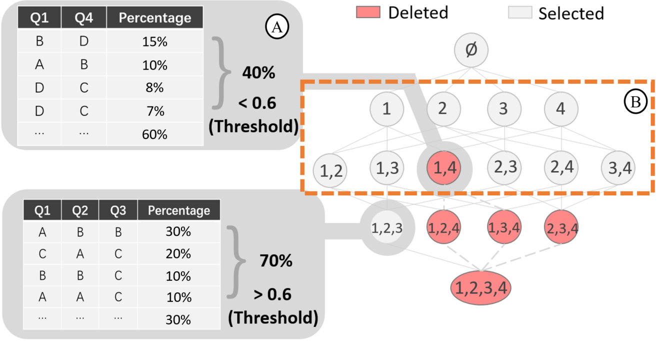
4.1 Question Association

Having their origin in Market Basket Analysis, association rules are one of the most popular tools in data mining. With marketers' attempts to study customers' shopping habits, association rules are applied to analyze what kinds of commodities will be purchased simultaneously. Formally, a set of items is defined as $I = \{I_1, I_2, \ldots, I_m\}$, and a purchase record is defined as $D = \{d_1, d_2, \ldots, d_n\}$. Every element in D is a non-empty subset of I. For an association rule $X \Rightarrow Y$ $(X,Y \subseteq I \; and \; X \cap Y = \emptyset)$, there are following requirements:

$$\begin{aligned} P(Y \cup X) &> \min_{support} \\ P(Y \mid X) &> \min_{confidence} \end{aligned} \qquad (1)$$

where $P(X \cup Y)$ represents the probability of X and Y being purchased in D at the same time. $P(Y \mid X)$ represents the probability of Y being purchased under the condition X being purchased in D. $\min_{support}$ and $\min_{confidence}$ are two thresholds, which mean how frequently the item set appears in the dataset and how often the rule is true, respectively. By treating options of questions as items and answers of respondents as purchase records, we can apply association rules to questionnaire datasets. In this way, we can solve association rules $X \Rightarrow Y$ that option set X leads to option set Y.

With the design needs from N2 and N3, our goal is to obtain the combination of relevant question sets and quantitatively measure the relationships among them. However, there are two challenges in transforming associations of options to the relationship between questions and relevant question combinations. First, associations apply to question options, but not to questions. However, traditional questionnaire causal analysis typically focuses on questions rather than options, and relationships between options are too detailed for analysts. Therefore, we need to define a method to combine the associations of options into questions. Second, the association of options has directions that are obtained by the methods. In the process of combining the association of options, We need to change the direction between options into direction between questions or eliminate the direction between options.



Q1	Q4	Percentage
B	D	15%
A	B	10%
D	C	8%
D	C	7%
...	...	60%

Q1	Q2	Q3	Percentage
A	B	B	30%
C	A	C	20%
B	B	C	10%
A	A	C	10%
...	...	...	30%

Fig. 1: (A) Distribution of option combinations in Q1 and Q4: The sum of responses in the top n (4 here) categories is less than the threshold (0.6 here). (B) A step of the Alg. 1: The algorithm searches from the $\emptyset$. The candidates in upper layer are merged with new questions to form candidates of the next layer. Then, the algorithm checks each question combination candidate, prunes unqualified candidates (**red** nodes in the figure), and repeats this step.

4.2 Association Aggregation

To address these challenges, we used a two-step approach to aggregate association rules. In the first step, we aggregated rules of different directions. In the second step, we aggregated options:

1. Merge association rules [5] $X \Rightarrow Y$, if the option set of $X \cup Y$ is equal. Use the results of $X \cup Y$ to represent them. For example, we will merge $(Q1: yes, Q2: no \Rightarrow Q3: yes)$ and $(Q3: yes, Q2: no \Rightarrow Q1: yes)$. And $(Q1: yes, Q2: no, Q3: yes)$ will be used to represent the above mentioned two association rules.
2. Merge the results of the previous step if their option sets correspond to the same set of questions. Use the questions to represent them. For example, we get two option sets: $(Q1: yes, Q2: no, Q3: yes)$ and $(Q1: yes, Q2: yes, Q3: no)$. We can combine them and use $(Q1, Q2, Q3)$ to represent them.

Now we obtain the combination of relevant question sets. Intuitively, the more association rules are merged in a question set, the more relevant questions in the set are. To achieve this, we first need to find shared association rules. However, if we collect them with the most commonly used algorithm, Apriori algorithm [5], the time complexity is $O(n_{people}(2^{n_{option}})^2)$, which is too large even for a small number of options. Based on our definition of aggregation, we propose an optimized approximation algorithm (Alg. 1). For each question combination, we only consider the top n_{top} set of option combinations chosen by most respondents (Fig. 1A), leading to a reduced time complexity $O(n_{people} 2^{n_{question}-1})$. The analysts can set n_{top}. The default value of n_{top} is $n_{top} = floor(\prod_i^{n_{option}} *0.3)(n_{option} \in question_i)$ where n_{option} means the number of options in question i.

Based on such an algorithm, we can process questions with any type of answer (e.g., numerical, categorical). But the questions must be close-ended and single-choice. Association rules between questions identified by our algorithm have some characteristics. First, each association rule, e.g. $(Q1, Q2, Q3)$, classifies respondents concerning the differences in their responses. The option combinations, e.g. $(Q1: yes, Q2: no, Q3: yes)$, chosen by different respondent groups reveal the relevance between multiple questions and give a good explainability of relevant question combinations (N2). Furthermore, we can evaluate the relevance of question combinations by the number of respondents who choose the top n_{top} set of option combinations (N3).

Algorithm 1: An optimized approximation algorithm for detecting aggregated association rules.

```
Input: A record of all responses $R$, a set of all questions $Q$, the
       constant support $s$, the number of combinations of
       responses considered $top_n$
Output: A set of question combinations $Q_c$ ($Q_{c_i}$ means an item
        in the set)
$Q_{c_1} \leftarrow \{q \in Q | FindTopOptionSum(q, top_n) > s\}$  ▷ For easier
  understanding, $FindTopOptionSum([Q1, Q4], 4)$ is shown in
  Fig. 1A, which means finding the total percentage of top 4
  option combinations of Q1 and Q4 ;
for $i \leftarrow 2$ to $length(Q)$ do
    $C_{c_i} \leftarrow AprioriJoin(Q_{c_{i-1}})$  ▷ AprioriJoin is the join step of
      Apriori algorithm [10]. For easier understanding,
      $AprioriJoin([Q1, Q2, Q3, Q4])$ is shown in Fig. 1B which
      means generating itemset of two questions
      ($[Q1, Q2], [Q1, Q3], ..., [Q3, Q4]$) from itemsets of one
      question ($[Q1, Q2, Q3, Q4]$) by joining each. ;
    $Q_{c_i} \leftarrow \{q_{c_i} \in C_{c_i} | FindTopOptionSum(q_{c_i}, top_n) > s\}$;
    if $Q_{c_i} = \emptyset$ then break;
return $Q_c$
```

5 System Design

Informed by the interview study, we iteratively designed the system for causal reasoning in questionnaire analysis and exploration. During a one-year iteration, many prototypes were built and tested with the participation of two domain experts. In this section, we describe how we designed the system based on their requirements and feedback.

5.1 Overview and Workflow

The user interface of QE is composed of three major views (Fig. 2): a question combination view for providing the overview of all question combinations with potential causality (N4, N6), a causal view for showing causality and linear regression among questions in a question combination (N5), and a respondent view for showing the distribution of respondent involved in a question combination (N3, N7). Meanwhile, a question list view is given to display the content of each question, which keeps users informed of the questions' details at any time.

The main workflow of this system is as follows. Users explore the question combination view and see the overview of the whole dataset. Next, they may find several question combinations that have significant patterns (e.g., ranking higher or covering more respondents). To dive into the inner causality of these questions, users will explore the causal structure of them in the causal view and inspect the distribution of respondents in the respondent view. After analysts gain a deeper understanding of the data, they may filter a certain group of respondents and repeat the exploration process again.

5.2 Question Combination View

As shown in Sec. 4, we identify numerous relevant question combinations via an association mining algorithm. Following the visual information-seeking mantra [14], we start by designing the overview of these question combinations. A combination of questions is absolutely a set. In order to help users explore the numerous combinations (sets), we adopt UpSet [29], which is a highly scalable set visualization approach and also widely used in the visualization community [48] to visualize combinations. However, UpSet does not directly meet all the design needs of analysts. According to the design needs and the workflow, this view should provide users with details of each question combination (N2, N3) and how individual questions are related to each combination (N6). Besides, the respondent groups should also be displayed to help analysts compare the number of major respondents between different question combinations (N2). Based on these needs, we have made the following designs on the ground of UpSet.

5.2.1 Question Combinations

Question combinations are the core concept in our analysis workflow. This view uses a matrix-based method from UpSet to visualize question combinations (Fig. 2B). Each row represents a question, while each column represents a combination. For each question, we label its id on the left side. And analysts can view the question description by hovering the mouse on the id. For combinations, each of them has two properties: respondent classification (N2) and question relevance (N3).

Respondent classification. For a combination, we classify respondents by their answers to the questions in the combination. In our implementation, we take the top n (n is calculated in Sec. 4.2) groups of question options in each combination, representing the answers of the majority of respondents to those specific questions. In other words, the majority group for a question combination includes all the respondents whose answers fall into the top option groups. The more concentrated the responses are, the more easily this combination can be interpreted. We adopt bar charts to reveal this information. The height of each bar is used to encode the number of the majority group, which can effectively show how concentrated or diverse the majority of answers to a question combination are. For example, if there are a great number of respondents choosing the same options for the questions, it is considered an abnormal pattern that is worth further investigation.

Question relevance. A question combination may contain two or more questions. We use black and gray colors to indicate whether a question is contained by a combination or not. And we use a black line to connect all included questions following the design of UpSet.

Beyond the inner questions, the experts want to look further at how important a question combination is, and how relevant a non-included question is to the hyperedge. Thus they can be more targeted in some questions and perform further analysis. Therefore, for each question combination, we use the width of connection lines to encode the number of occurrences of the current combinations among all solved combinations. For example, the solved question combinations are: [Q1, Q2], [Q1, Q3], [Q2,Q3], [Q1, Q2, Q3]. Then the width of connection line encoding [Q1, Q2] (which appears in both [Q1, Q2] and [Q1, Q2, Q3]) is wider than [Q1, Q2, Q3] (which only appears in [Q1, Q2, Q3]). This encoding allows users to simply find important combinations which occur most often in the large space of all combinations. For the questions that are not included in that particular combination, we will analyze their correlation with the combination. For the most relevant question, we use darker gray color to encode it in this view. And the darkness is linearly correlated to the calculated value. This question must be relevant to the corresponding question combination and can hint about the direction for further exploration.

With these visual encodings, users can more intuitively discover the pattern of questions and combinations in the dataset.

5.2.2 Order of Question Combinations

The display order of question combinations would influence analysts' comprehension since people tend to observe the items positioned at the beginning. To determine an appropriate order of question combinations, we adopt three rules to place them hierarchically. First, analysts may manually choose an interesting question by clicking its name. All question combinations that contain this question will be placed to the left. Second, question combinations with more questions are preferred and placed to the left. Finally, question combinations covering more respondents are preferred and placed to the left.

5.2.3 Cluster of Questions

Our design is based on UpSet [29]. However, the matrix of UpSet becomes sparse when the number of questions is large. Meanwhile, the designer of a questionnaire tends to design questions of different focuses, which are, unfortunately, hard to differentiate using UpSet.

To address the issue of sparsity and help analysts understand the different aspects of the questions, we adopted a clustering algorithm based on weighted sets [42] to cluster the questions automatically. We consider the number of majority respondents of a question combination as weights. This method uses a greedy algorithm to divide the questions automatically and does not require the user to specify the number of clusters in advance. The visualization of each cluster is based on UpSet as described above and is vertically laid out in the question combination view. Furthermore, we provide users with a button to toggle whether to display questions by cluster or not. Users can choose it according to their preferences and the features of the dataset.

After the exploration of question combinations, analysts will further select the combination of interest via information from different encodings. The information in a question combination can be divided into the causal structure between questions (N4) and the relationship between respondents and questions (N2). These two types of information can be checked respectively in the causal view and respondent view.

5.3 Causal View

We designed this view to visualize the causal structure of selected question combinations (N4). Since analysts often use regression-based methods to test the truth of causality (Sec. 2.1), we incorporated linear regressions into this view to help analysts determine the influence of each cause and the fitness of target questions. Linear regression uses dependent variables to fit an independent variable, which represents the target question in questionnaire scenario. Therefore, it is crucial to calculate the causal structure first to identify the dependent and independent variables. The system then needs to provide an intuitive visualization for analysts to explore and verify the causality and linear regression coefficients. Fortunately, both calculating and visualizing causal relationships have been well-studied for a long period of time. In our system, we directly adopted the Fast Greedy Equivalence Search (F-GES) algorithm [35] to construct causal relationships between questions and used a DAG design [33] to visualize the final structure.

5.3.1 Encoding of Nodes and Links

Refer to Causality Explorer [44], every question (node) is represented by a pie chart (Fig. 2C), where each sector encodes the proportion of an answer. This can help analysts learn the distribution of answers for each question and provide guidance for exploration. For instance, the analyst can quickly identify questions that most respondents share the same answer. The outer ring of nodes reveals how well a question is fitted by the linear regression of its cause questions (Fig. 2C). The better the fit, the greater the angle.

Links represent causal relationships and their directions are consistent from the lower node to the upper node. For instance, in Fig. 2C$_1$, there is a link from *"Q15:Are you frequently absent from school?"* to *"Q16: Students' grades"* which indicates that the habit of skipping classes likely influences the students' grades. Each link possesses two properties: uncertainty and linear regression coefficient. Uncertainty is generated by the bootstrapping F-GES algorithm, which will be introduced in Sec 5.3.2. As one of the most important variables when analyzing, we encode uncertainty using link thickness, which is considered the most effective visual channel [22] for lines. Thicker edges represent lower uncertainties. Linear regression coefficients, on the other hand, not only indicate the strength of the influence, but also show whether the influence is positive or negative. They are another important property that analysts are concerned about, so we just place the precise values of coefficients right next to the edges.

5.3.2 Graph Layout

We adopt the DAG layout proposed by Xiao et al. [44], where cell positions are determined by their casual relationships. For example, if Q1 causes Q2, the vertical position of Q1 is lower than Q2, which helps analysts quickly understand the directions of causal relationships. Our method contains four steps.

Step1: F-GES. In the first step, we use the state-of-the-art F-GES [35] algorithm to obtain the causal relationships between all questions. Specifically, F-GES adopts a greedy strategy to enumerate the best Bayesian Information Criterion (BIC) by adding and deleting causal relationships. The uncertainty of a causal relationship is the decreasing value of BIC after deleting itself:

$$\text{Uncertainty}(e) = \text{BIC}(G) - \text{BIC}\left(G_{-e}\right). \tag{2}$$

This algorithm guarantees that the uncertainty is positive.

Step2: BFS. Directly showing all casual relationships of the whole dataset will cause severe visual clutter. To deal with this problem, we only display causal relationships between questions in a question combination. However, questions of a selected combination may not be directly connected by casual relationships, but through some intermediate questions. For example, the *gender of the family member who takes care of you primarily* does not change your *grades* directly, but it may affect your character and habits, which in turn influence your *grades*. This suggests the necessity to explore questions of interest in the context of directly relevant questions. Unfortunately, finding the least number of questions to make the causal DAG connected is an NP-hard problem, known as Steiner Tree Problem [20]. Hence, we take BFS to deal with this situation and obtain a sub-tree that connects all questions in the question combination.

Step3: Topological Sorting. Our design uses arrows to encode the direction of causal relationships. However, arrows may intersect with each other, leading to severe visual clutter. To address this issue, we adopt the topological sort method [44] to divide a casual graph into layers for better readability.

Step4: User Constraints. The automatic result may not perfectly match experts' domain knowledge. Therefore, our system allows users to manually add or remove prior constraints to guide the causal graph generation in the first step of F-GES. These constraints help F-GES recompute and optimize other relevant casual relationships, and the visualization will be updated accordingly to reflect the changes. For example, experts may specify that the childcare question should not be the direct cause of work hour questions, by removing the corresponding link in the user interface. The constraint will then be incorporated into the F-GES model. After recomputing, these two questions are bridged by the work efficiency question with updated uncertainty scores. Mathematically, we support four types of prior constraint: $Q_1 \to Q_2$, $Q_1 \not\to Q_2$, $\forall Q \not\to Q_1$, and $Q_1 \not\to \forall Q$, corresponding to required causal relations, non-existing causal relations, setting independent questions and setting dependent questions.

Each constraint matches an interaction. (1) By linking two nodes (Fig. 2C), users can add a prior knowledge that Q15 is the reason for Q16. (2) By crossing out a link (Fig. 2C$_1$), users can add prior knowledge that Q9 is never the reason for Q16. (3) Users can set a problem as a dependent question by clicking on it with the right mouse button and choosing the corresponding option. (4) Users can also set a problem as an independent question in the same way.

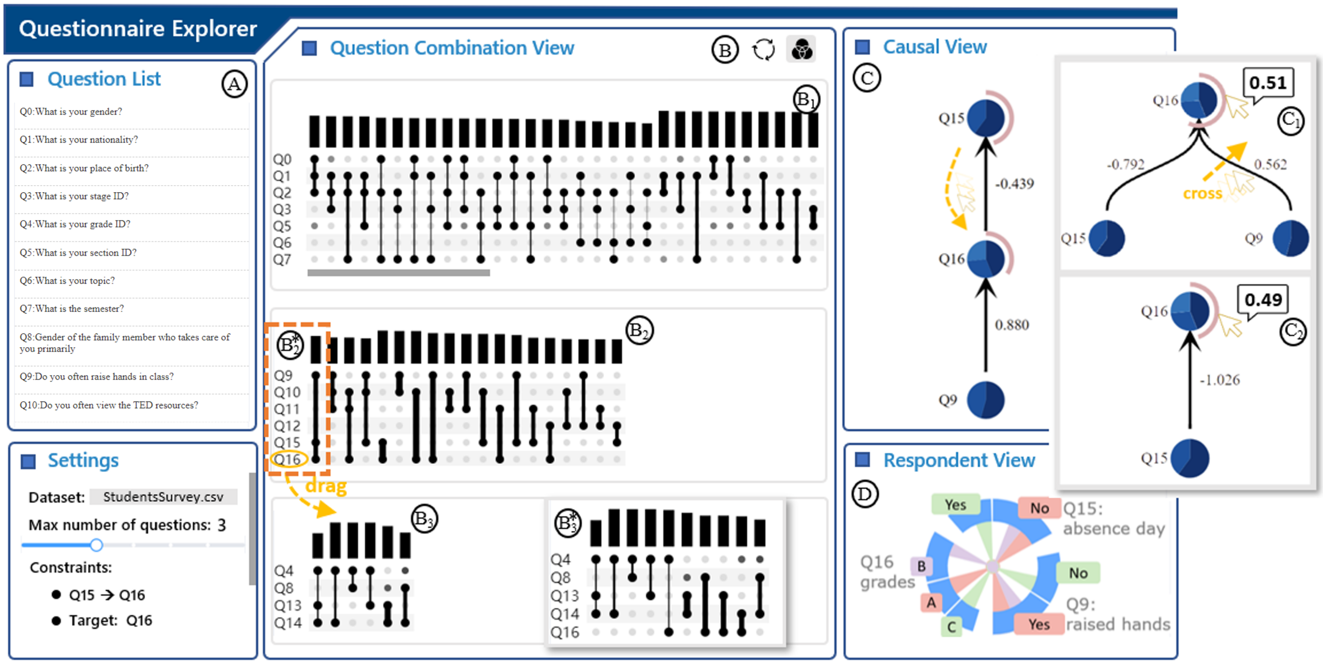
5.4 Respondent View

Although the question combination view contains bar charts showing a summary of major respondent categories, it is still unclear what options these respondents select in their answers. In response to this limitation, when users click a question combination in the question combination view, the respondent view illustrates the detailed relationships between respondent categories and question options. The main component of the view consists of a donut chart and a set of fan-shaped glyphs in the center (Fig. 2D). The outer ring of the component is evenly segmented into several sectors corresponding to the questions in a question combination. Then, for each question sector, we further divide it into smaller option sectors with angles proportional to the number of respondents who choose the corresponding options.

Inside the ring, we place fan-shaped glyphs corresponding to the top respondent categories of the question combination. For each glyph, the fans of the same color connect the option sectors, which means a major respondent category and their answers. The centered circle radius and fan angles are used to encode the number of respondents in the category of corresponding colors. Combined with their domain knowledge, experts can go through the options of the major respondent categories to determine whether the relationship between these questions is explainable or unexplainable. Explainable relationships can be used to test experts' hypotheses, while unexplainable relationships may be caused by coincidence and should be assigned a lower priority in the analysis. Also, analysts can learn about the characteristics of different categories of respondents. Some categories of respondents may attract the analysts' interest, and the analysts can set this group of people as the analysis target of QE by clicking the corresponding fans or sectors.

6 Evaluation

In this section, we demonstrate the effectiveness and usefulness of our system in a two-part evaluation. We first report two usage scenarios of different datasets. Then, we compared our system with a scalable causal analysis system [44] in the questionnaire scenario with a controlled user study. These two experiments clearly demonstrate the strengths and weaknesses of our system.



Questionnaire Explorer

Question List

Q0:What is your gender?

Q1:What is your nationality?

Q2:What is your place of birth?

Q3:What is your stage ID?

Q4:What is your grade ID?

Q5:What is your section ID?

Q6:What is your topic?

Q7:What is the semester?

Q8:Gender of the family member who takes care of you primarily

Q9:Do you often raise hands in class?

Q10:Do you often view the TED resources?

Settings

Dataset: StudentsSurvey.csv

Max number of questions: 3

Constraints:

- Q15 → Q16
- Target: Q16

Fig. 2: The interface of QE: (A) The question list view displays all questions. (B) The question combination view provides an overview of the whole dataset. (C) The causal view presents the causality in a relevant question combination. (D) The respondent view visualizes the clusters of respondents divided by a set of relevant questions for users to deep dive. (other makers with subscripts) They show different patterns indicated by the exploration of students' grades in Sec 6.1.1.

6.1 Usage Scenarios

The following two cases use two real-world datasets to show different usage scenarios of our system.

6.1.1 Causal Reasoning

In the first case, we demonstrate how our system can help analysts find the questions that have strong influences on a target question. The dataset of this case is an open-source questionnaire dataset from Tianchi [3], containing students' information as well as their grades. In this case, *the students' grades* (Q16 in Tab. 1) is the target question. Bob, the teacher, needs to explore causal relationships between the target question and other questions. There are 18 questions (all single-choice questions) and 481 respondents in this dataset.

After loading the dataset, Bob found that the questions were automatically divided into three clusters (Fig. 2B). To obtain an overview of the dataset, He looked at the questions in each cluster. Combining his background knowledge, He determined that the three clusters were respectively related to *parent situations* (Fig. 2B$_3$, such as Q8), *student performances* (Fig. 2B$_2$, such as Q9 and Q11), and *student profiles* (Fig. 2B$_1$, such as Q1 and Q2).

After getting familiar with the questionnaire, Bob had to identify some questions related to the target question *grades*. He found that the target question was in the cluster of *student performances* (Fig. 2B$_2$), which led him to hypothesize that *grades* were more relevant to the questions related to *student performances* .Therefore, he first explored this cluster. There are a lot of combinations in this cluster. The question combinations were ranked based on their importance (see Sec. 5.2.2). Bob found the most important combination (Fig. 2B$_2^*$) contains three questions (*grades*, *absence days*, and *raised hands*). To further understand how these questions are causally-related, he tapped on this combination to explore them in the respondent and causal views.

In the respondent view (Fig. 2D), sectors of the same color represented a group of people who gave the same answers to *grades*, *absence days*, and *raised hands*. For example, red sectors represented the students who were absent from school less (the answer *absence days* is No), raised their hands more (the answer of *raised hands* is Yes), and got better grades (the answer of *grades* is A). In contrast, the green sectors represented students who had bad habits and struggled with achieving good grades. The relationship between performances and grades was exactly in line with Bob's background knowledge.

In the causal view (Fig. 2C), the *grades* (Q16) pie chart is connected to the *absence days* (Q15) pie chart, which meant our underlying model implied that *grades* was the cause of *absence days*. However, it did not match Bob's domain knowledge, because *grades* should only be the effect of *absence days*. To correct the wrong causal relationship, Bob manually added a new causal constraint from *absence days* (Q15) to *grades* (Q16) by a link.After recalculating, the causal graph was automatically refined by leveraging the user input (Fig. 2C$_1$). Via hovering the mouse on the outer pink ring, Bob found that *raised hands* and *absence days* fit the *grades* well linearly, with a score of 0.51. However, the red and purple parts in Fig. 2D showed that the students who raised their hands (Q9, *raised hands*) would not prefer to skip school (Q15, *absence days*), which meant these two questions are highly correlated and may have interchangeable effects on *grades*. If so, these two questions could fit *grades* well, regardless of whether together or individually [19]. Bob eliminated one causality by crossing to check the interchangeable effect. The fit did not change much (Fig. 2C$_2$), indicating that *raised hands* and *absence days* were highly interchangeable and dependent on each other. So Bob only needed to choose one of them to study *grades*. He found that the combination of *grades* and *absence days* ranked higher than the combination of *grades* and *raised hands*. Then, he chose *absence days* because its relationship with the target question is stronger.

After checking other question combinations, Bob found an interesting pattern that *caregiver gender* (Q8) is sometimes calculated to be the direct reason for student performances, and the indirect reason for *grades*. Because *caregiver gender* is in the cluster of *parent situations*, Bob decided to continue exploring the influence of *parent situations* on the *grades*. Hence, Bob dragged the question of *grades* (Q16) to the cluster of *parent situations* (Fig. 2B$_2^*$).

The updated cluster (Fig. 2B$_3^*$) has a small number of question combinations, which means other questions in this cluster may have little to do with Q16. Considering the indirect causality of *caregiver gender* and *grades*, he tapped the question combination of these two questions, which also ranked as the first in the view, to see the explanation in the respondent view and the causal graph in the causal view. In the respondent view, Bob found there are three dominant student groups, corresponding to purple, red, and green sectors, respectively. And the patterns suggested that female caregivers (purple sectors) might lead to better grades, compared with male caregivers (red and green sectors). At the same time, the link in the causal view showed that the *caregiver's gender* affected *grades* by influencing some study habits like *raised hands* (Q9) and *announcements* (Q11). After the exploration, Bob felt *caregiver gender* could be another important reason for *grades*.

He wanted to check whether there were other important reasons. He merged all clusters together by clicking the toggle and explored some highly ranked question combinations. Based on the causal sub-graph of these question combinations, Bob believed that the other questions were unrelated to the target question or highly dependent on *absence days* and *caregiver gender*. At the same time, Bob found that *caregiver gender* and *absence days* could fit *grades* well with a score of 0.56. So Bob believed that he had successfully found two high-quality causes of *grades* and gained a deep understanding of this dataset.

6.1.2 Free Explorations

The second dataset is collected from a survey conducted in a company, regarding the working experiences during the COVID-19 pandemic period. The questionnaire dataset consists of 72 questions (66 single-choice questions) and 475 respondents.

In the first step, Bob, the analyst, used our system to explore questions related to the target question: *"which workplace do you prefer the most?"* (Q60, denoted by *workplace*), and successfully found three questions to fit *workplace* as reasons using a similar way as in Sec. 6.1.1. He obtained a satisfying linear regression fit score of 0.23. The three reason questions included *"I feel good to be closer to family"*, *"I feel good to spend less money on commute, food, etc."*, and *"compared with working in office, how has your productivity changed"*. In this section, we demonstrate how Bob used our system to do further open-ended exploration and discover some unexpected causality.

First insight. After finishing the question relationship mining task around the target question, Bob went back to see the overview in the question combination view. He merged all clusters by clicking the clustering toggle to explore the relationship of all questions. He observed that the color of *"frequency of email usage"* (Q32) was dark in columns of combinations involved *"frequency of group-wise meetings"* (Q37), *"I feel more flexible to join meetings"* (Q41), and *"I have difficulties with colloquial meetings"* (Q47)(Fig. 3B). According to Bob's experience in the former question relationship mining, he knew that these questions were in the same cluster and were about meetings. This encoding indicated that *emails* (Q32) are relevant to questions in the cluster about meetings. Thus, Bob formed a hypothesis that this was because people with more meetings may use email to book meetings or discuss online more frequently. He then would check this hypothesis by QE. He clustered all questions by clicking the clustering toggle and moved the *emails* question (Q32) to the *meeting* cluster, and then chose the top-ranked combination that involved Q32 in this cluster. The question combination included *emails* (Q32), *online documents* (Q31), and *group-wise meetings* (Q37). Bob explored the distribution of respondents (Fig. 3C) in the respondent view, which would bring some ideas about question *emails* (Q32). The sectors in the respondent view indicated the largest respondent groups (Fig. 3C_1) showed the answer with the highest number of choices was *increased* and *significantly increased*. It meant most of the employees used online tools(e.g., online documents, emails) more frequently during the epidemic. At the same time, the green and red sectors in the respondent view meant that employees who used emails more had more meetings and did use other online tools more. Therefore, Bob accepted his first insight due to the explanation in the respondent view of this hyperedge.

Second insight. Bob now wanted to see what patterns the people who sent more emails had. From Bob's domain knowledge, He guessed these people might be managers. Because during the Covid-19, managers needed to send more emails and hold more meetings to organize employees who worked from home. He chose this part of people who selected "increased" in Q32 by clicking the corresponding arc and looking at the question combination view to get an overview. In the question combination view, Bob found that most of the questions had a light gray color. But the row of Q62 (*job*) was much darker. At the same time, the vertical connection lines of question combinations involving Q62 were thicker than others. This meant that Q62 (*job*) had a strong relevance with other questions and was included in a number of question combinations. As Bob remembered, he did not see a similar pattern before selecting respondents who sent more emails.

Bob thought about this phenomenon and gave a hypothesis: for some jobs, nobody sent more emails. Therefore, after selection, respondents' answers were more centered on the left jobs. If respondents of Q62 (*job*) were concentrated in the corresponding options, the calculated association (question combinations) would be more and stronger in our algorithm (see Sec. 4). In order to test this hypothesis, He clicked Q62 and checked the respondent view (Fig. 3A) to see people's choice of *jobs* for all people and people who sent more emails. There were indeed some disappeared jobs shown in Fig. 3A_1, including *Design*, *HR*, *Finance*, and *District Community Process Manager*. Meanwhile, the angle of the arcs representing *Managers* in the right ring was much smaller than in the left ring. This meant that *Managers* did not think they used more emails while other jobs (e.g., engineer) did. Although this was not in line with Bob's domain knowledge, Bob could try to give an explanation: the managers were used to sending so many emails before COVID-19, so they did not think that their use of emails increased. And developers were just the opposite. Through this

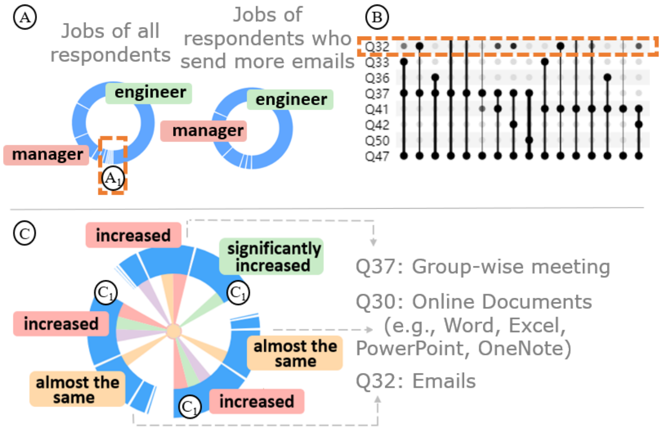


Fig. 3: (A) The distribution of answers to the question *job* (Q62) is different between all respondents and those who send more emails. (B) The *emails* question (Q32) is important in many other question combinations (C). The respondent view shows a lot of participants use emails, online documents, and group-wise meetings more often.

Table 1: Questions used in Sec. 6.1.1

Id	Question	Acronym
Q1	What is your nationality	nationality
Q2	What is your place of birth	place of birth
Q8	Gender of the family member who takes care of you primarily	caregiver gender
Q9	Do you like to raise hands in class	raised hands
Q10	Do you often view the TED resources	TED resources
Q11	Do you often view the announcements	announcements
Q15	Are you frequently absent from school	absence days
Q16	Students' grades	grades

Table 2: Questions used in the Sec. 6.1.2

Id	Question	Acronym
Q15	I have sufficient office setup	office setup
Q31	Frequency of online documents usage	online documents
Q32	Frequency of email usage	emails
Q37	Frequency of group-wise meetings	group-wise meetings
Q41	I feel more flexible to join meetings	flexible meetings
Q44	I feel easier to access team members.	members relationship
Q47	I have difficulties with colloquial meetings	colloquial meetings
Q60	Which workplace do you prefer the most?	workplace
Q62	What is the role of your job?	job

exploration, Bob gained additional insights. Online tools were crucial and frequently used during the epidemic. Moreover, changes in work style had a greater impact on developers than on managers.

Third insight. In the above exploration, Bob discovered that *job* was an important and interesting question as people's decisions may be influenced by their job roles. He wanted to explore its influence on the target question (*workplace*) further by finding reasons for the target problem for people with different jobs. He completed the question relationship mining task by following a similar process to that in Section 6.1.1 for *developers* and *managers*.

As a result, Bob achieved much higher fits for both managers and developers. For developers, Bob replaced the reason *"I feel good to be closer to family."* with *"I have better work environment."* and achieved a better fit score of 0.26. For managers, Bob replaced *"I feel good to spend less money on commute, food, etc."* and *"compared with working in office, how has your productivity changed?"* with *"Please rate your overall satisfaction with working from home during the COVID-19 period. "* and *"I feel good to have freedom for physical actions (standing, sitting, etc.)."* and got a great fit of 0.31. With the open-ended exploration, Bob gained a deeper understanding of this dataset and was able to provide better recommendations to the company.

6.2 User Study

To test the efficiency (G1) and usability (G2) of our system, we conducted a comparative study. The traditional questionnaire analysis process (usually performed using tools like Python, SPSS, Excel, etc.) involves enumerating causal structures and conducting interspersed tests, which can take several working days to complete. In contrast, QE aids users in understanding data through visual designs and provides heuristic high-quality causal structures, which is also confirmed by our tester experts. Therefore, comparing QE with traditional analysis tools is not meaningful enough. To ensure a more fair comparison, We selected Causality Explorer [44] (S2), a state-of-the-art research work for causal analysis of large tabular datasets, which could be a competitive tool for exploring the causal structure of all questions in a questionnaire. At the same time, QE (S1) adopts an identical DAG layout with Causality Explorer. The main difference between them is the analysis pipeline which is exactly the focus of our comparison: QE is based on causal sub-graphs of question combinations, while Causality Explorer shows the whole causal graph to users. Apart from the main difference, Causality Explorer does not support adding causal constraints interactively. Therefore, We have added the same interaction to the causal graph view in Causality Explorer as in QE. Then the analysts can perform the same interactions to add constraints and select part of the respondents in both QE and Causality Explorer.

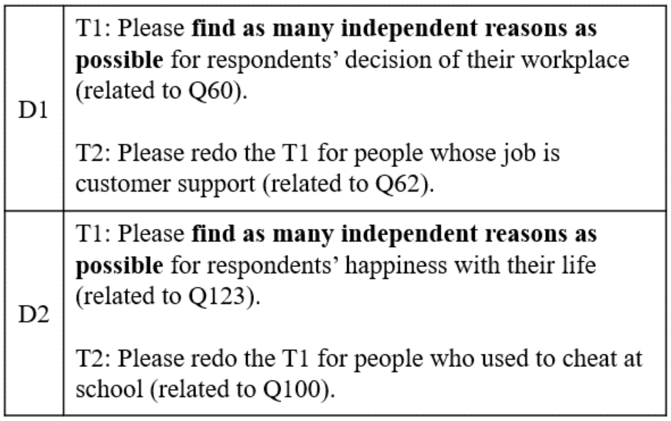
We conduct a with-in-subject design to compare these two systems based on two datasets. The first dataset (D1) is the same dataset in Sec. 6.1.2. And the second dataset (D2) is collected from Kaggle [4] which contains 150 questions to explore the preferences, interests, habits, opinions, and fears of 1010 young people. To balance the effect of different systems and datasets [43], we derived four conditions, including [D1S1, D2S2] (A participant first uses QE to complete tasks on Dataset 1, and then uses Causality Explorer to complete tasks on Dataset 2), [D1S2, D2S1], [D2S2, D1S1], and [D2S1, D1S2].

6.2.1 Participants and Study Setup

We recruited 12 participants [38], 6 males, and 6 females, aged 24-30, from local companies. All of them have Master's degrees and more than four years of experience in questionnaire analysis. In their daily work, Python, R, Excel, and SPSS are used as the analysis tools. None of them have experience in using interactive visualization systems (e.g., Causality Explorer) to perform causal reasoning tasks. QE and Causality Explorer are both deployed on the cloud. Participants took part in the study remotely via video conference on their own computers.

6.2.2 Procedure

Participants were first asked to watch two short videos to get familiar with QE and Causality Explorer. The videos contained the interface and interactions of both systems. Then, they could use the two systems to explore a training dataset (the same dataset in Sec. 6.1.1). They were encouraged to ask us questions on the systems until they were familiar enough with both QE and Causality Explorer.



D1	T1: Please **find as many independent reasons as possible** for respondents' decision of their workplace (related to Q60). T2: Please redo the T1 for people whose job is customer support (related to Q62).
D2	T1: Please **find as many independent reasons as possible** for respondents' happiness with their life (related to Q123). T2: Please redo the T1 for people who used to cheat at school (related to Q100).

Fig. 4: There are two step-by-step tasks on each dataset: (T1) Find reasons for a target question. (T2) Redo T1 for a subgroup of respondents.

After the warming-up training, each participant was asked to perform causal analysis tasks (Fig. 4) under the above-mentioned four conditions ([D1S1, D2S2], [D1S2, D2S1], [D2S2, D1S1], and [D2S1, D1S2]). In each task, the participants were allowed to analyze with a time limit of 20 minutes or stop when they found all causes they wanted. Lastly, we asked participants to complete a questionnaire about the usability of both QE and Causality Explorer, followed by a short semi-structured interview. During the interview, we discussed with them their preferences for the two systems. The whole study lasted about two hours. And each participant was paid 100 CNY after the study.

6.2.3 Results

In the following, we report results of the study to demonstrate the efficiency (G1) and usability (G2) of QE.

G1: Efficiency in causal reasoning. In our study, all participants were experienced in questionnaire analysis. Furthermore, after conducting our experiments, we ensured that all identified reasons satisfied two criteria: (1) no significant collinearity (tested by variance inflation factor [15]); (2) no obviously incorrect reasons (e.g., age as the reason for grades). Since the causal reasoning was highly dependent on the analysts' prior knowledge, we assumed that all participants were able to find correct causes which matched their prior knowledge. Therefore, we test G1 by analyzing the average time to find a cause instead of checking the correctness of the causes.

In our study, there are three factors that affect the average time to find a cause, including different datasets, systems, and tasks. Our aim is to check whether the time to find causes through QE is less than Causality Explorer and has no relationship with the choices of dataset and task. Considering the task to find causes for all respondents and the task to find causes for a part of respondents always occurred in pairs, we performed a paired t-test (two-tailed) on these two types of tasks and found no significant difference for them ($p = 0.37$). This means we can ignore the effect of different tasks when analyzing the difference between systems and datasets. Then, we check the interaction relationship of datasets and systems via Two-Way ANOVA. There is no significant interaction ($p = 0.43$) between datasets and systems. Therefore, we do not need to consider the impact of datasets when comparing QE and Causality Explorer. We directly analyze the average time to find a cause for each analyst using different systems. The time performance is shown by box plots in Fig.5A. Then we performed a t-test (one-tailed) to check whether analysts could find causes more efficiently on QE than on Causality Explorer. The results yielded $p < .001$, validating that the efficiency of causal reasoning was significantly improved on QE compared to Causality Explorer.

According to our observation and the post-study interview, we identified two reasons why QE is more efficient than Causality Explorer. First, the information in Causality Explorer is homogeneous and scarce. Analysts can only find new causal relationships based on the semantic meaning and their prior knowledge after checking the causes that were directly or indirectly connected to the target question. This process is inefficient and may lead to missing causal relationships. In contrast, QE provides visual encodings in the question combination view and

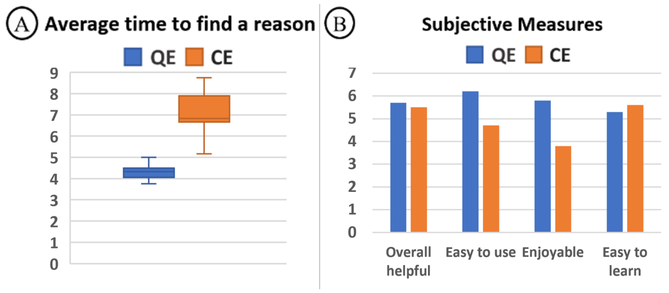


Fig. 5: Questionnaire Explorer (QE) v.s. Causality Explorer (CE). (A) Average time (in minutes) to find a reason. (B) Participants' rantings from the post-study survey (1 = "strongly disagree" and 7 = "strongly agree").

respondent view, which help analysts identify possible causes and understand the relationships between questions. These hints could help the analyst explore and validate the missing causal relationships. Second, the causal structures in Causality Explorer before and after adding constraints will have a great difference. Users of Causality Explorer had to recognize the huge DAG, including a large number of nodes and edges. However, much of the information was unhelpful in finding the causes of the target questions. In contrast, QE eliminates such an issue by showing the sub-graphs of the whole causal graph.

G2: Usability. In the post-study interview, we asked users which system they preferred to use for performing the causal reasoning task in their daily analysis. 91.7% (11/12) of users preferred QE. At the same time, users rated QE positively (Fig. 5B) in terms of *easy of use* ($M: 6.2, SD: 0.8$) and *enjoyable to use* ($M: 5.8, SD: 0.6$). In comparison, the Causality Explorer was rated by: *easy to use* ($M: 4.7, SD: 2.1$) and *enjoyable to use* ($M: 3.8, SD: 1.5$). There is a significant difference in these two questions (one-tailed t-test, *easy to use*: $p < .01$ and *enjoyable to use*: $p < .01$) between QE and Causality Explorer. This indicates that QE is more useful in the causal reasoning of scalable questionnaire data. The other ratings, including *Overall helpful* and *Easy to learn*, were not statistically significant.

According to the interview, there were two reasons that made users prefer QE. First, the workflow of QE, which is based on question combinations, is closer to the regular pipeline in users' daily work. Users reported that QE not only gives the causal structure of the problem combinations they are interested in but also lists mediating variables in the structure, which is a great aid to their daily analysis. Second, the causal graph view of Causality Explorer, including dozens of nodes and hundreds of edges, looked chaotic. These visual clutters gave rise to users' resentment and also made the system low-usability. All participants (12/12) noted this point in the interview. At the same time, QE can help them focus on useful information while alleviating clutters.

7 Discussion and Future Work

In this section, we will discuss the implications, lessons learned, and the limitations and future work of QE.

7.1 Implications

This research is the first step toward the model-assisted visual analysis of questionnaire data. Analysts can efficiently discover explainable relationships between questions and dig further into causality.

techniques. QE proposes an innovative visual analytic approach to model and present question combinations in questionnaires and designs a workflow based on question combinations to address the scalability issue in existing causality visualization.

applicability. The workflow that enables uses balance the exploration of sets (question combinations) and the corresponding causal sub-graph can be applied to many other scalable datasets (e.g., urban data based on numerous sensors). At the same time, the visual encoding and the interactions of QE are helpful for the visual community to dive into questionnaire analysis scenarios. As demonstrated in the usage scenarios, experts can easily find interesting patterns and conduct in-depth causal reasoning with the help of QE.

7.2 Lessons Learned

The first lesson is about our backend model. In a discussion with survey experts (P1-2), they mentioned that they prefer model-driven analysis over data-driven analysis. This is because model-driven analysis is easier to explain and can be used to persuade decision-makers. Surprisingly, F-GES [35], as a data-driven algorithm, has also gained their favor. Based on our understanding, as users become more involved in data-driven algorithms, the credibility of algorithms among users will increase. Ultimately, analysts will confidently use data-driven algorithms with added constraints to report to decision-makers. The second lesson is related to the design of analysis systems for tasks that involve text-intensive data, such as questionnaire analysis. It is important to provide users with hints about the details they focus on. These hints include the content of questions, the nature of scales, and basic quality metrics. They could be provided through the use of tooltip, which can help remind experts of important information and reduce the burden of memory. And for some of the more important contents, such as the question description in the respondent view, we list them directly. This will make the system more cluttered. But analysts' thoughts won't be interrupted by frequent checking the question description in the tooltips. The trade-off between aesthetics and text information needs to be more thoughtfully designed in such visualization systems.

7.3 Limitations and Future Work

To fully support the design requirements, some compromises are made in terms of scalability. First, the fan-shaped glyphs in the causal view and the respondent view are designed to encode a question consisting of multiple options and show the relationships between the questions and the options. However, the glyph design does not scale well with questions [18]. In our system, users can set the maximum number of problems for association analysis to mitigate this limitation. Second, the system employs UpSet in the question combination view. Although this method is able to display all features of the questions, it is burdensome for users to scroll horizontally and vertically to explore all question combinations and clusters. To alleviate this issue, we further sort question combinations by importance and provide visual cues to help users identify the starting point of the analysis. We also plan to explore other methods [34] to further enhance the scalability of QE.

Currently, QE focuses on the casual analysis of questionnaire data and assumes that the data are ready for statistical analysis. However, in real practice, the quality of the data and the design of the questionnaire are also critical as poor-quality questionnaire data and erroneous questionnaire design can significantly impact the analysis results. The main influencing factors for the quality of questionnaire data are the data cleaning process and statistical tests (e.g., reliability analysis, validity analysis). In the future, we will incorporate certain statistical test methods into QE to reject low-quality data input or automatically perform data cleaning for such data. Second, during the initial stage of questionnaire design, researchers often conduct small-scale surveys. At this point, they can use QE to analyze pre-experimental data. Based on the analysis results, researchers can eliminate questions with low relevance and collinearity, as well as balance the number of questions within each cluster. However, QE currently cannot detect important questions that may have been overlooked in the questionnaire, such as the influence of age on the workplace through an overlooked intermediate variable like marital status. In the future, we will explore methods to assist analysts in identifying these overlooked questions.

8 Conclusion

This study presents Questionnaire Explorer, a novel visual analytic system that enables target-based causal reasoning, association explanation, and open-ended exploration. We propose an association-based computing method to extract relevant question combinations with potential causality. To interactively explore these combinations, we designed a matrix-based system for visualization. At the same time, we innovatively enable users to explore causal sub-graphs of each question combination to alleviate the scalability issue of current causality visualizations. The proposed system has been evaluated through a comparative user study based on real-world data. The results demonstrate the efficiency and usability of our system. In the future, we plan to incorporate natural language processing methods for semantic-based analysis and expand the system's support for more data types.

Acknowledgments

The work was supported by National Key R&D Program of China (2022YFE0137800) and NSFC (62072400, 62202424), and the Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government (ZJU).

References

[1] Microsoft Forms, 2020. Retrieved Dec 1st, 2020 from https://forms.office.com/. 1

[2] SurveyMonkey, 2020. Retrieved Dec 1st, 2020 from https://www.surveymonkey.com/. 1

[3] Tianchi, 2021. Retrieved Aug 21st, 2021 from https://tianchi.aliyun.com//. 6

[4] Kaggle, 2022. Retrieved Aug 21st, 2022 from https://www.kaggle.com/. 8

[5] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pp. 207–216. ACM, 1993. doi: 10.1145/170035.170072 3

[6] J. Antonakis, S. Bendahan, P. Jacquart, and R. Lalive. On making causal claims: A review and recommendations. *The leadership quarterly*, 21(6):1086–1120, 2010. doi: 10.1016/j.leaqua.2010.10.010 1, 2

[7] J. Bae, T. Helldin, and M. Riveiro. Understanding Indirect Causal Relationships in Node-Link Graphs. *Computer Graphics Forum*, 36(3):411–421, 2017. doi: 10.1111/cgf.13198 2

[8] J. Bae, E. Ventocilla, M. Riveiro, T. Helldin, and G. Falkman. Evaluating multi-attributes on cause and effect relationship visualization. In *International Conference on Information Visualization Theory and Applications*, pp. 64–74, 2017. doi: 10.5220/0006102300640074 2

[9] M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004. doi: 10.1162/003355304772839588 1, 2

[10] C. Borgelt and R. Kruse. Induction of Association Rules: Apriori Implementation. In *Proceedings of Compstat*, pp. 395–400. Springer, 2002. doi: 10.1007/978-3-642-57489-4_59 4

[11] S. H. Burton, R. G. Morris, C. G. Giraud-Carrier, J. H. West, and R. Thackeray. Mining useful association rules from questionnaire data. *Intelligent Data Analysis*, 18(3):479–494, 2014. doi: 10.3233/IDA-140652 2

[12] A. Cao, X. Xie, J. Lan, H. Lu, X. Hou, J. Wang, H. Zhang, D. Liu, and Y. Wu. Mig-viewer: Visual analytics of soccer player migration. *Visual Informatics*, 5(3):102–113, 2021. doi: 10.1016/j.visinf.2021.09.002 2

[13] Y.-L. Chen and C.-H. Weng. Mining fuzzy association rules from questionnaire data. *Knowledge-Based Systems*, 22(1):46–56, 2009. doi: 10.1016/j.knosys.2008.06.003 2

[14] B. Craft and P. Cairns. Beyond guidelines: what can we learn from the visual information seeking mantra? In *Proceedings of International Conference on Information Visualisation*, pp. 110–118. IEEE, 2005. doi: 10.1109/IV.2005.28 4

[15] T. A. Craney and J. G. Surles. Model-dependent variance inflation factor cutoff values. *Quality engineering*, 14(3):391–403, 2002. doi: 10.1081/QEN-120001878 8

[16] Z. Deng, D. Weng, Y. Liang, J. Bao, Y. Zheng, T. Schreck, M. Xu, and Y. Wu. Visual cascade analytics of large-scale spatiotemporal data. *IEEE Transactions on Visualization and Computer Graphics*, 28(6):2486–2499, 2022. doi: 10.1109/TVCG.2021.3071387 1

[17] Z. Deng, D. Weng, X. Xie, J. Bao, Y. Zheng, M. Xu, W. Chen, and Y. Wu. Compass: Towards Better Causal Analysis of Urban Time Series. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1051–1061, 2022. doi: 10.1109/TVCG.2021.3114875 1, 2

[18] S. G. Eick and A. F. Karr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11(1):22–43, 2002. doi: 10.1198/106186002317375604 9

[19] D. E. Farrar and R. R. Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pp. 92–107, 1967. doi: 10.2307/1937887 6

[20] M. R. Garey and D. S. Johnson. The Rectilinear Steiner Tree Problem is NP-complete. *SIAM Journal on Applied Mathematics*, 32(4):826–834, 1977. doi: 10.1137/0132071 5

[21] B. Ghai and K. Mueller. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):473–482, 2023. doi: 10.1109/TVCG.2022.3209484 1

[22] H. Guo, J. Huang, and D. H. Laidlaw. Representing Uncertainty in Graph Edges: An Evaluation of Paired Visual Variables. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1173–1186, 2015. doi: 10.1109/TVCG.2015.2424872 5

[23] J. Hahn, P. Todd, and W. Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001. doi: 10.1111/1468-0262.00183 1, 2

[24] M. N. Hoque and K. Mueller. Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4728–4740, 2022. doi: 10.1109/TVCG.2021.3102051 1, 2

[25] Z. Jin, S. Guo, N. Chen, D. Weiskopf, D. Gotz, and N. Cao. Visual causality analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1343–1352, 2021. doi: 10.1109/TVCG.2020.3030465 2

[26] A. Kale, Y. Wu, and J. Hullman. Causal Support: Modeling Causal Inferences with Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1150–1160, 2021. doi: 10.1109/TVCG.2021.3114824 2

[27] J. Klaus, M. Blacher, A. Goral, P. Lucas, and J. Giesen. A visual analytics workflow for probabilistic modeling. *Visual Informatics*, 7(2):72–84, 2023. doi: 10.1016/j.visinf.2023.05.001 2

[28] J. A. Krosnick. Questionnaire design. In *The Palgrave Handbook of Survey Research*, pp. 439–455. 2018. doi: 10.1007/978-3-319-54395-6_53 2

[29] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248 4

[30] C. Lonsdale, C. M. Sabiston, I. M. Taylor, and N. Ntoumanis. Measuring student motivation for physical education: Examining the psychometric properties of the Perceived Locus of Causality Questionnaire and the Situational Motivation Scale. *Psychology of Sport and Exercise*, 12(3):284–292, 2011. doi: 10.1016/j.psychsport.2010.11.003 1

[31] J. Méndez, C. Alrabbaa, P. Koopmann, R. Langner, F. Baader, and R. Dachselt. Evonne: A visual tool for explaining reasoning with owl ontologies and supporting interactive debugging. *Computer Graphics Forum*, 2023. to appear. doi: 10.1111/cgf.14730 2

[32] J. Müller-Sielaff, S. B. Beladi, S. W. Vrede, M. Meuschke, P. J. F. Lucas, J. M. A. Pijnenborg, and S. Oeltze-Jafra. Visual assistance in development and validation of bayesian networks for clinical decision support. *IEEE Transactions on Visualization and Computer Graphics*, 29(8):3602–3616, 2023. doi: 10.1109/TVCG.2022.3166071 1

[33] J. Pearl et al. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. 5

[34] A. Pister, P. Buono, J.-D. Fekete, C. Plaisant, and P. Valdivia. Integrating prior knowledge in mixed-initiative social network clustering. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1775–1785, 2021. doi: 10.1109/TVCG.2020.3030347 9

[35] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129, 2017. doi: 10.1007/s41060-016-0032-z 2, 5, 9

[36] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991. doi: 10.1177/089443939100900106 2

[37] F. van Ham and A. Perer. "Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960, 2009. doi: 10.1109/TVCG.2009.108 3

[38] R. A. Virzi. Refining the test phase of usability evaluation: How many subjects is enough? *Human factors*, 34(4):457–468, 1992. doi: 10.1177/001872089203400407 8

[39] J. Wang and K. Mueller. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):230–239, 2016. doi: 10.1109/TVCG.2015.2467931 1, 2

[40] J. Wang and K. Mueller. Visual Causality Analysis Made Practical. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 151–161, 2017. doi: 10.1109/VAST.2017.8585647 2

[41] Y. Wang, Y. Liu, W. Cui, J. Tang, H. Zhang, D. Walston, and D. Zhang. Practices around Working from Home and Early Indicators on Returning to Work after the COVID-19 Pandemic: Data from Microsoft China. 2020. 2

[42] X. Wei, J. Hu, and Y. Luo. Automatic Software Module Partition Based on Hypergraph Model. *Computer Engineering*, 42(1):71–76, 2016. doi: 10.3969/j.issn.1000-3428.2016.01.014 4

[43] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu. A semantic-based method for visualizing large image collections. *IEEE Transactions on Visualization*

and Computer Graphics, 25(7):2362–2377, 2018. doi: 10.1109/TVCG.2018.2835485 8

[44] X. Xie, F. Du, and Y. Wu. A Visual Analytics Approach for Exploratory Causal Analysis: Exploration, Validation, and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1448–1458, 2021. doi: 10.1109/TVCG.2020.3028957 1, 2, 5, 8

[45] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of Causality in Visualized Data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2020. doi: 10.1109/TVCG.2019.2934399 2

[46] C.-H. E. Yen, A. Parameswaran, and W.-T. Fu. An exploratory user study of visual causality analysis. *Computer Graphics Forum*, 38(3):173–184, 2019. doi: 10.1111/cgf.13680 2

[47] H. Yeon, H. Son, and Y. Jang. Visual performance improvement analytics of predictive model for unbalanced panel data. *Journal of Visualization*, 24:583–596, 2021. doi: 10.1007/s12650-020-00716-0 2

[48] Z. Zhu, Y. Shen, S. Zhu, G. Zhang, R. Liang, and G. Sun. Towards better pattern enhancement in temporal evolving set visualization. *Journal of Visualization*, 26(3):611–629, 2023. doi: 10.1007/s12650-022-00896-x 4