Effect of multiple visualization elements on task performance in group-in-a-box layouts

Hiroaki Natsukawa, Yuki Ueno, Nozomi Aoyama, and Koji Koyamada

Abstract—For effective graph visualization, in addition to the optimization of computational measures such as total ink and edge crossing, user experiments to quantify task performance are often performed. However, in a sophisticated graph drawing involving multiple visual elements, the elements can affect users' performance either positively or negatively. The group-in-a-box (GIB) layout is an efficient graph drawing method designed to visualize the group structure of graphs. It consists of multiple elements such as a node-link diagram and boxes that explicitly illustrate the group structure. In this study, using GIB as an example, we measured the performance and eye movements of participants performing the task of identifying the group with the largest internal edges. We examined the effect of visualization elements on task performance while controlling for the density of internal edges and box size. The results revealed that the density of internal edges and box size in a GIB layout significantly affects the accuracy of the task, and the presence and absence of boxes caused fluctuations in accuracy. Usage of sophisticated visualizations can be beneficial if used successfully, but it can also lead to unexpected misjudges.

Index Terms—Group-in-a-box layout, eye tracking, user study

1 Introduction

As data becomes available on a large scale and in a wide variety of forms, various graph-drawing layouts, edge-bundling techniques, and sophisticated visualization methods have been proposed. Easily seen graphs have been generated by optimizing computational measures such as total ink, edge crossing, and developer-defined energy values. Visualizations with improved computational measures have a superior appearance; however it is unclear whether users can easily retrieve information from these visualizations. Consequently, user experiments and subjective evaluation have been conducted to evaluate the effectiveness of visualizations. In user experiments, visualizations are evaluated by measuring user task performance, including correct answer rate and completion time for tasks that involve reading encoded data. Visualization is in essence mediation between humans and computers. User experiments for evaluating visualization go beyond simply analyzing the difference in performance for multiple layout; they incorporate experimental approaches from the fields of visual cognition and cognitive psychology. The field of visualization currently faces several challenges, including determining which visual elements have crucial effects on task performance, and determining the just noticeable difference of a particular visual attribute.

Visualization is generally not considered a well-controlled or simple expression in a visual cognitive sense, but rather a combination of multiple visual elements such as colors, points, bars, and lines. Consequently, multiple visual elements presented on a screen can affect performance either positively or negatively in a task that involves extracting information from a visualization. Although visualization should be as simple as possible, excessive visual elements may be used unintentionally in a real-world visualization setting or visualization system.

In this study, we examine visualizations combining several visual elements that can be used in an actual visualization system, and evaluate whether the visual elements have an effect on user task performance, and what the size of that effect is. The group-in-a-box (GIB) layout is an efficient graph drawing method designed to visualize the group structure of graphs. It consists of multiple elements, including a node-link diagram and boxes that illustrate the group structure explicitly. In this study, using GIB as an example, we measured the task performance and eye movements of participants performing a task involving determining which group has the largest internal edges. Here internal edges signify the edges connecting nodes in the same group (i.e., the edges in a given box). In a task involving extracting information from a graph with a group structure, the visual elements of the graph can determine the accuracy of extraction. However, other visual elements and interaction effect of visual attributes may also have a significant effect on task performances. Therefore, we examined the effect of visualization elements on task performance while controlling for the density of internal edges and box size.

Specifically, to analyze the effect of boxes and other visual attributes in the task of extracting the internal edge features of a GIB layout, the following three conditions were used: 1) only boxes are displayed, 2) only nodes and edges are displayed, and 3) boxes, nodes, and edges are displayed. By analyzing the difference in task performance between the three conditions, we were able to quantitatively evaluate the effects of boxes and other visual attributes. Twenty-seven participants performed three tasks for two GIB layouts, TR-GIB and FD-GIB. In addition to measuring task accuracy, we also obtained eye-tracking data to determine where the participants' gaze was focused when they performed the tasks. Our results revealed that the density of internal edges and box size in a GIB layout greatly affected the accuracy of the task, and the presence or absence of boxes caused fluctuations in accuracy.

The key contributions of this study are as follows:

- User experiments to evaluate the effect of visual elements in a GIB layout on task performance
- Quantification of the effect of visual elements on task accuracy in a GIB layout

- *Hiroaki Natsukawa is with Kyoto University. E-mail: natsukawa.hiroaki.3u@kyoto-u.ac.jp.*
- *Yuki Ueno is with Kyoto University. E-mail: ueno.yuki.78x@st.kyoto-u.ac.jp.*
- *Nozomi Aoyama is with Kyoto University. E-mail: aoyama.nozomi.37x@st.kyoto-u.ac.jp.*
- *Koji Koyamada is with Kyoto University. E-mail: koyamada.koji.3w@kyoto-u.ac.jp.*

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

2 Related work

In this section, we discuss related research in which user experiments are conducted to evaluate visual design, the analysis of exploration behavior is performed using eye-tracking data, and the GIB layout is examined.

2.1 Evaluation of visual design through user experiments

In the fields of cognitive neuroscience, visual cognition and psychophysics, psychophysiological experiments are commonly conducted to determine the types of primitive visual stimuli that have an effect on human cognition; this information serves to inform human visual characteristics. Psychophysiological experiments and user experiments in the field of visualization have also been conducted [24, 22, 8, 20, 15]. Szafir performed a qualitative analysis of color difference perceptions in common visualizations, discovering that perceived color differences vary inversely with size and that colors are more discriminable on elongated bars and lines than on points. Additionally, controlled user studies have also been conducted to evaluate visual comparison [15], color assignment in multiclass scatterplots [24], color map data [22, 20], and distribution data [8]..

Research on evaluating visualizations can lead not only to improved visualization design, but also to more thorough understanding of human high-order visual processing. In this study, we evaluated the effect of complex elements in a graph visualization from a different perspective from the related studies described above. In addition to performing a quantitative evaluation, we also examined participants' exploration behavior by measuring thier eye movements during the experiments.

2.2 Analysis of exploration behavior using eye-tracking data

The analysis of user behavior while using visualization systems has been performed in several studies [11, 6, 13, 10, 14]. In these studies, an eye-tracking system is used to record participants' eye movements. Eye-tracking data makes it possible to understand the manner in which participants use the developed visualization system; additionally, it provides insights into the participants' reasoning methods and problem-solving strategies [1]. Thus, eye-tracking data can be used to improve visualization systems by evaluating the usefulness and readability of visualization technology in terms of visual cognition.

For example, Netzel et al. [14] evaluated four variants of geographic map annotation: within-image annotation, grid reference annotation, directional annotation, and miniature annotation. Participants were instructed to identify the specified label within the map as quickly and accurately as possible. As they performed the task, both eye-tracking data and completion time were recorded. The results indicated that the within-image annotation outperformed all other annotation methods. Additionally, eye-tracking data revealed that the participants used different task strategies for different geographic map annotations. Burch et al. [6] explored three types of tree diagrams: a traditional tree layout, orthogonal tree layout, and radial tree layout. Participants were instructed to identify the least common ancestor of a given set of marked leaf nodes, considered a typical hierarchical exploration task. During this task, eye-tracking data was recorded using an eye-tracker. Additionally, the accuracy and completion time of the task was recorded as well. From the eye-tracking data, it can be seen that the exploration strategies differed for each method. The participants frequently cross-checked their solutions and required more time to complete the task when using the radial layout than the other layouts.

In this study, we measured participants' eye movements during a task to reveal their exploration behaviors. This data elucidated the manner in which the participants searched the GIB graph before reaching an answer, as well as the focus of their gaze.

2.3 GIB Layout

The GIB layout is a graph-drawing method designed to visualize the group structure of graphs [21, 7, 16]. In GIB, all nodes in a group are placed within a box whose size is proportional to the number of nodes. Therefore, using GIB, it is possible to simultaneously visualize group structure, the relationship between various groups, and the size of the groups in the graph. In this study, the GIB layout was selected as the evaluation target for the following reasons. First, GIB layouts consist of various visualization elements such as lines, points, and boxes, and multiple pieces of information can thus be obtained from the layout. Second, GIB layouts are suitable for performing eye-tracking analyses based on an area of interest (AOI). Specifically, in GIB layouts, a graph diagram is divided into boxes which can be regarded as AOIs. There are several GIB layouts, including FD-GIB and TR-GIB, which were used in this study and are described below. Fig. 1 presents examples of these layouts.

FD-GIB Force-directed GIB (FD-GIB) was developed by Chaturvedi et al. [7]. This method uses a force-directed layout to arrange each box according to its attraction to the center and the repulsion between boxes. Because this layout can create overlaps, we used the PRISM method to decrease any overlaps [9]. Although this layout is suitable for depicting the topology of an entire network, it may present challenges in understanding the relationships that exist in a single group, as each box can only occupy a small area. However, the aspect ratio of each box can be made constant in this layout; therefore, users should be able to easily compare box sizes.

TR-GIB Another layout was proposed by Onoue et al. [16]: the tree-reordered GIB (TR-GIB). This layout is based on the squarified treemap GIB (ST-GIB) developed by Rodrigues et al. [21], which is coordinated by squarified treemaps proposed by Bruls et al. [5]. The ST-GIB layout does not consider the relationship between nodes when the boxes are arranged; therefore, it includes edge crossing, tending to hamper users' understanding of the depicted networks [4, 17, 18, 19]. TR-GIB is optimized so that the lengths of all edges of the ST-GIB are minimized. More specifically, the TR-GIB layout minimizes the weighted sum of the distances between groups by reordering the sibling nodes in the ST-GIB layout. Because the TR-GIB layout is optimized to minimize the distance between groups, it has fewer edge crossings than in the ST-GIB. Thus, this layout has the advantage of ST-GIB's favorable aspect ratio and effective screen use as well as the property of fewer edge crossings.

To evaluate GIB layouts, Chaturvedi et al. performed computational experiments on the following three layouts: ST-GIB, FD-GIB, and croissant-and-doughnut GIB (CD-GIB); here the latter layout improves ST-GIB by considering the link information connecting a node to another node belonging to a different group [7]. Onoue et al. demonstrated that TR-GIB is advantageous over ST-GIB in terms of computational measures [16]. In our previous study, we evaluated the four above-mentioned layouts (ST-GIB, CD-GIB, FD-GIB, TR-GIB) from the perspective of human cognition through user experiments in which eye-tracking data was collected [3]. The optimal layouts were FD-GIB and TR-GIB, both of which offered various advantages and disadvantages. FD-GIB was effective in determining the number of links and abstract information, while TR-GIB was effective in representing more concrete relationships, such as links between specific nodes. The eye-tracking data provided evidence to support these results; additionally, it provided insights to help reveal the features in the GIB layout that affected task performance.

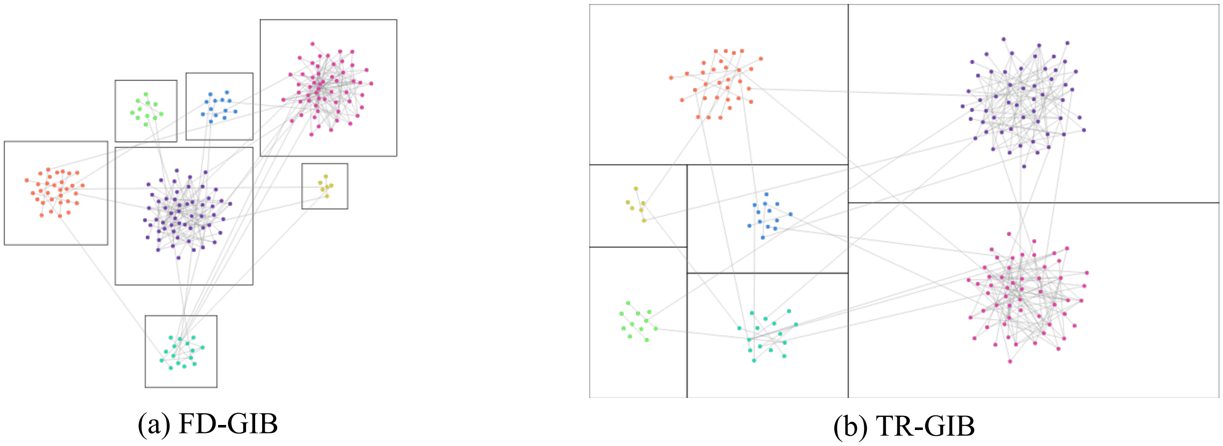


Fig. 1. Examples of group-in-a-box (GIB) layouts: (a) force-directed GIB (FD-GIB), (b) tree-reordered GIB (TR-GIB).

We conducted an additional controlled laboratory experiment to examine multiple exploration behaviors in a task [23]. In the experiment, we fixed the number of groups to 7 and 14 and investigated which visualization factors affect the task performance. The results indicated that the exploration behavior was determined by whether the correct answer was the box with the largest area, a visualization factor that considerably affects the correct answer rate.

From the results of these experiments, we were ale to identify the tradeoffs in GIB layouts and the factors that determined performance in GIB representations. However, the extent to which accuracy was affected by whether the answer was the box with the largest area remained an open question. Additionally, modeling the task of estimating internal edges in GIB was also unclear. By designing and conducting controlled laboratory experiments to address these questions, we were not only able to propose effective usage of GIB based on human cognitive abilities, but we were also able to provide a guideline for sophisticated visualization usage.

3 Experiment

In the task of identifying the box with the largest number of internal edges in a group, previous experiments [3, 23] determined that the size of the box as well as the difference in the internal edge affected the solution. In other words, in visualization systems that present multiple visual elements, other visualization elements may affect performance when extracting information from a graph. The purpose of this study is to quantify this effect and to model task performance in an effort to understand human performance characteristics using the GIB layout. Additionally, this study aims to obtain further knowledge for determining optimal visualization usage.

3.1 Task

Previous experimental results [3, 23] suggest that not only the number of edges in a group but also the size of the boxes affects the correct answer rate. To quantify these effects and model task performance, it is necessary to measure the correct answer rate for the following three cases;

1. The box size affects the accuracy.

2. The number or density of the edges in the group affects the accuracy.

3. Both the box size and density of the edges affect the accuracy.

The effects in these three cases can be modeled by designing an experiment that controls for the visualization factors that may affect task accuracy. We therefore created the following three tasks using the GIB layout.

Task 1 Determine which group has the largest area (i.e, maximum number of nodes) when only nodes and boxes are displayed.

Task 2 Determine which group has the largest number of internal edges when only nodes and edges are displayed.

Task 3 Determine which group has the largest number of internal edges when nodes, edges, and boxes are displayed.

Fig. 2 presents an example of each task.

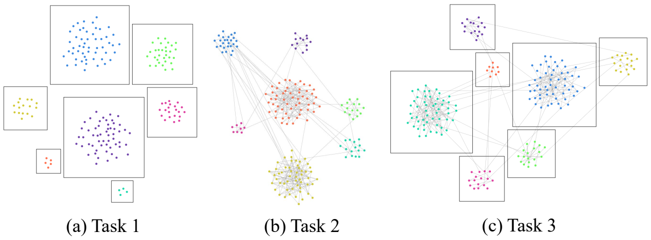


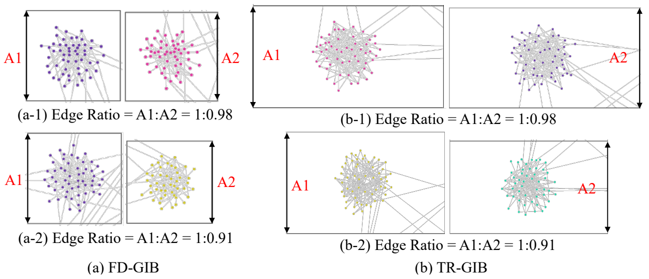
Fig. 2. Examples of each task. (a) Task 1: Determine which group has the largest area when only nodes and boxes are displayed. (b) Task 2: Determine which group has the largest number of internal edges when only nodes and edges are displayed (c) Task 3: Determine which group has the largest number of internal edges when the original GIB are displayed.

3.2 Data and layout generation

To implement the task described in Sect. 3.1, data in this experiment was generated by a method different from that used in previous research [3, 23]. Although the manner of displaying data differed between the three tasks, to unify the experimental conditions, the data was generated using the same method for all tasks. It was necessary to control for the size of the box in Task 1 and the size of the box and the number of edges for Tasks 2 and 3. To control for these conditions, each element of the GIB layout was set as follows.

3.2.1 Number of groups

Because the purpose of the experiment was not to observe changes in task performance due to differences in group size, the number of groups was fixed at seven.



(a-1) Edge Ratio = A1:A2 = 1:0.98

(b-1) Edge Ratio = A1:A2 = 1:0.98

(a-2) Edge Ratio = A1:A2 = 1:0.91

(b-2) Edge Ratio = A1:A2 = 1:0.91

(a) FD-GIB

(b) TR-GIB

Fig. 3. Example of the edge ratio of a box that is the correct answer candidate for each GIB layout

3.2.2 Number of nodes

In this task, the size of the box to be controlled was dependent on the number of nodes in the group. In a series of tasks, the size of a box is considered to affect task performance. Therefore, we created two cases; one in which the sizes of the two boxes that are the correct answer candidates were easy to compare, and one in which they were difficult to compare. However, the difference in the side lengths of the boxes was not identical in TR-GIB and FD-GIB even when the number of nodes in the two correct answer candidates was the same. The reason for that is TR-GIB has higher space utilization efficiency. Additionally, if the differences in the lengths of the sides of the two boxes were equal for TR-GIB and FD-GIB, it would be easier to recognize the differences in FD-GIB, having lower space utilization efficiency than TR-GIB. Therefore, the ratio of the sides of the two correct answer candidates were aligned in both layouts. The edge ratio (Er: Edge ratio) of the two correct answer candidates was set to $1 : 0.98$ and $1 : 0.91$. In TR-GIB, due to layout characteristics, two rectangular correct answer candidates were arranged vertically and had a common long side; thus, the side ratio was applied to the remaining short sides. In FD-GIB, the two correct answer candidates were both square; thus, the above-mentioned side ratio was applied to all sides.

The total number of nodes was 185, which is the average of the total number of nodes used in the previous study [23]. Let $N1$ and $N2$ be the number of nodes in the two candidate boxes, and let $N3$–$N7$ be the number of nodes in the remaining boxes. The number of nodes is set as follows;

$$(N1, N2) = \begin{cases} (55, 54) & (\text{TR-GIB}, Er = 1 : 0.98) \\ (55, 50) & (\text{TR-GIB}, Er = 1 : 0.91) \\ (55, 53) & (\text{FD-GIB}, Er = 1 : 0.98) \\ (55, 46) & (\text{FD-GIB}, Er = 1 : 0.91) \end{cases} \tag{1}$$

$$4 \leq N3, ..., N7 \leq N2 - 10 \tag{2}$$

Fig. 3 presents an example of the edge ratio of the correct answer candidates with $N1$ and $N2$ nodes.

To prevent $N3$–$N7$ from being the correct answer candidates, the ratio to the side of the second largest box was set to be smaller than 0.91 and was randomly generated.

3.2.3 Number of intergroup edges and internal edges

In the task of selecting a box with the largest number of internal edges, when the number of internal edges was large, participants could simply select the higher density box instead of counting and selecting the number of edges. Here, let C be the area of a circular region surrounding all nodes in a certain group (expressed in $pixel^2$), and let L be the number of edges in the group. The density D of the internal edges is then defined as follows:

$$D = L/C \tag{3}$$

Because the density of the internal edges directly affects task performance, the task difficulty level was controlled for not by adjusting the difference in the number of internal edges, but by adjusting the internal edge density difference. Regarding the two correct answer candidates, let $L1$ and $L2$ be the number of edges in the group with $N1$ and $N2$ nodes, respectively. Let $C1$ and $C2$ be the area of the circles that enclose all nodes in these groups, respectively. The density difference ΔD of the internal edge is expressed as follows:

$$\Delta D = L1/C1 - L2/C2 \tag{4}$$

With reference to data from previous research, we set the range of the difference in internal density as follows:

$$-9 \times 10^{-4} \leq \Delta D \leq 9 \times 10^{-4} \tag{5}$$

Here we consider the cases in which ΔD is positive or negative. Positive ΔD signifies that the box with the largest area has the largest number of internal edges. Similarly, negative ΔD signifies that the box with the second largest area has the largest number of edges.

In this layout, the circle enclosing all nodes in each group is determined by the number of nodes, internal edges, and intergroup edges. Thus, the circular area cannot be determined prior to generating the visualization. Of the data generated based on the fixed number of nodes and variable number of edges, this experiment used data satisfying the following conditions:

If $L1$ is the largest number of internal edges

$$\Delta D \approx \begin{cases} 1 \times 10^{-4} \\ 3 \times 10^{-4} \\ 5 \times 10^{-4} \\ 7 \times 10^{-4} \\ 9 \times 10^{-4} \end{cases} \tag{6}$$

$$L1 > L2 > L3, \cdots, L7 \geq 1 \tag{7}$$

$$L1/C1 > L2/C2 > L3/C3, \cdots, L7/C7 \tag{8}$$

$$|L'1/C1 - L'2/C2| \approx 1 \times 10^{-4} \tag{9}$$

$$20 \geq L'1, L'2 > L'3, \cdots, L'7 \geq 1 \tag{10}$$

If $L2$ is the largest number of internal edges

$$\Delta D \approx \begin{cases} -1 \times 10^{-4} \\ -3 \times 10^{-4} \\ -5 \times 10^{-4} \\ -7 \times 10^{-4} \\ -9 \times 10^{-4} \end{cases} \tag{11}$$

$$L2 > L1 > L3, \cdots, L7 \geq 1 \tag{12}$$

$$L2/C2 > L1/C1 > L3/C3, \cdots, L7/C7 \tag{13}$$

$$|L'1/C1 - L'2/C2| \approx 1 \times 10^{-4} \tag{14}$$

$$20 \geq L'1, L'2 > L'3, \cdots, L'7 \geq 1 \tag{15}$$

Here, let $C1$–$C7$ be the area of the circles surrounding all the nodes belonging to groups with $N1$–$N7$ nodes. Let the number of internal edges in each group be $L1$–$L7$, and the number of intergroup edges be $L'1$ to $L'7$. The number of edges between groups was set to a maximum of 20. This was to avoid a situation in which there were too many edges between groups, reducing the overall readability. When the number of edges between groups is large in the GIB layout, a method called edge bundling is often used to bundle the edges to improve readability. However, in this experiment, a certain degree of visibility was required, as it was not our aim to quantify the efficacy of edge bundling in GIB. Fig. 4 presents an example of the density difference of the correct answer candidates with internal edge numbers L1 and L2 in FD-GIB.

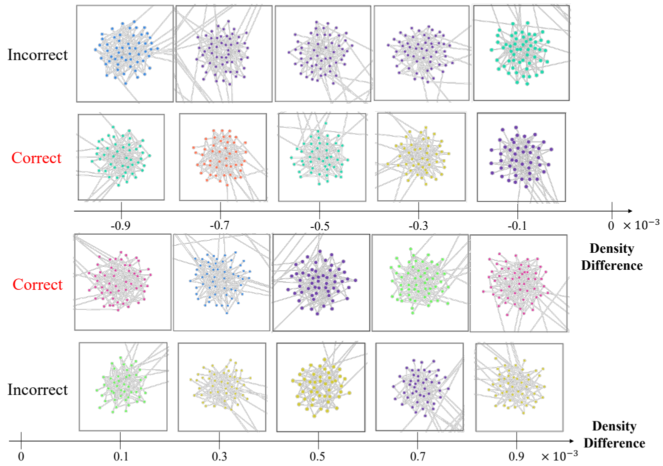


Fig. 4. Density difference of the correct answer candidates in the ratio 0.91 in FD-GIB

3.2.4 Data amounts

We constructed 20 types of data created from the combination of the visualization method (TR-GIB, FD-GIB), edge ratio ($1 : 0.98$, $1 : 0.91$), and difference in density of edges within groups ($-9\times10^{-4}\sim9\times10^{-4}$). In Task 1, one dataset was generated for each type, while in Task 2 and 3, five datasets were generated for each type. Thus, 440 datasets (=20 types × (1 dataset + 5 datasets + 5 datasets)) were generated in total for the entire task.

3.2.5 Layout

The generated data was visualized using the TR-GIB and FD-GIB methods. Further, the GIB layouts could only determine the arrangement of the boxes; therefore, it was necessary to determine the node coordinates in each box. Of the layouts available for arranging nodes in a box, we used the force layout, as it is known to reduce edge crossing and increase readability [12]. In this method, nodes are arranged according to the repulsion and the attraction between them, and the gravity level from the center of the group to which they belong. The color scheme of each group was set randomly. Although different colors produce different psychological effects and may affect task performance, this effect was eliminated in our study by randomly arranging the colors. A sample GIB layout can be output at our open-access website [2].

3.3 Study design

In this experiment, Task 1 involved the ratio of two types of sides and had a total of 20 trials (ratio of two types of sides × 10 trials). Task 2 and 3 involved the ratio of two sides and the difference in density of 10 types, resulting in 100 trials in total (ratio of two types of sides × difference of 10 types of density × 5 trials). Because Task 2 did not display a box when visualizing data, we could not verify the ratio of sides visually; however, to equate the conditions in Tasks 2 and 3, the data was generated identically. There were two layouts, resulting in a total of 440 trials (2 × (20 trials + 100 trials + 100 trials)). In Task 1, 40 trials were divided into two sets, where one set had 20 trials. In Tasks 2 and 3, 100 trials were divided into eight sets, where one set had 25 trials. The same type of GIB layout was displayed in each set. The order of the sets and trials in the set was randomized to eliminate any possible effects resulting from trial order. The participants took a short break of approximately 30 seconds following each set, and a long break of up to 5 minutes between each task and between the fourth and fifth sets during Tasks 2 and 3. The eye tracking system was re-calibrated after every break.

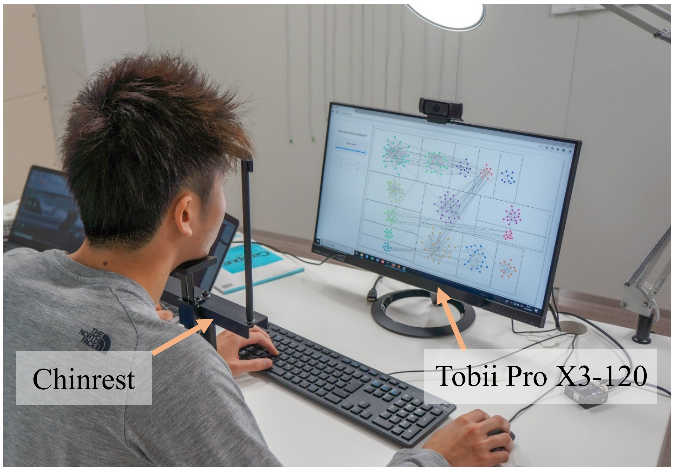


Fig. 5. Experimental environment. The eye movement was recorded using an eye tracker attached to the bottom of the monitor.

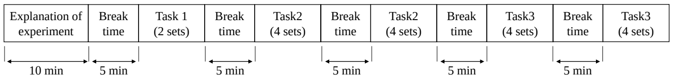


Fig. 6. Illustration of experimental paradigm

3.4 Experimental setup

The experiment was conducted in our laboratory, which was illuminated with artificial lighting. The task was displayed on a 24-inch monitor, with a resolution of $1,920 \times 1,080$ pixels, and eye movements were recorded using the Tobii Pro X3-120 eye-tracking system. Additionally, we used a chin rest to prevent participants' head from moving during the experiment and leading to noise in the eye-tracking data. Fig. 5 presents an overview of the experimental setup.

3.5 Study procedure

The experiment took 1.5–2 hours, including preparation, explanation, and rest. First, participants were provided an explanation of eye-tracking and each GIB layout. Then, they sat 65 cm from the monitor and were given instructions on the task while performing a tutorial prior to each task. They trained sufficiently to avoid the influence of habituation in the actual experiment. The experiment was performed as described in Sect. 3.3. The participants were instructed to perform each task correctly, and no time limit was set for each task so that the participants had enough time to select the correct answer. If the participants focused on answering quickly, their false answer rate would likely be high and chaotic eye movements would be obtained, which was not intended in this experiment. Participants continued to the following task after clicking on an answer and pressing the enter key. Fig. 6 illustrated the experimental paradigm.

3.6 Participants

The participants were 27 healthy adults with normal or corrected normal vision; 21 were male and 6 were female. The participants were 21–33 years old with an average age of 24 years. Three participants were familiar with GIB, four participants were familiar with visualization, while the remaining participants had no prior knowledge of GIB. The latter subjects had not been involved in visualization studies but had experience reading information from figures and tables. Informed consent was obtained in advance from all participants.

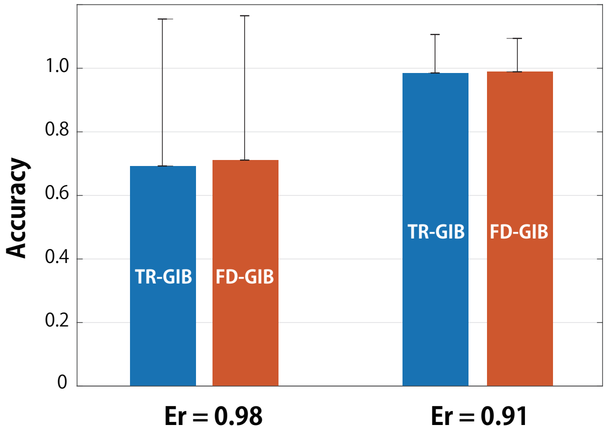


Fig. 7. Relationship between edge ratio and accuracy in each GIB layout in Task 1

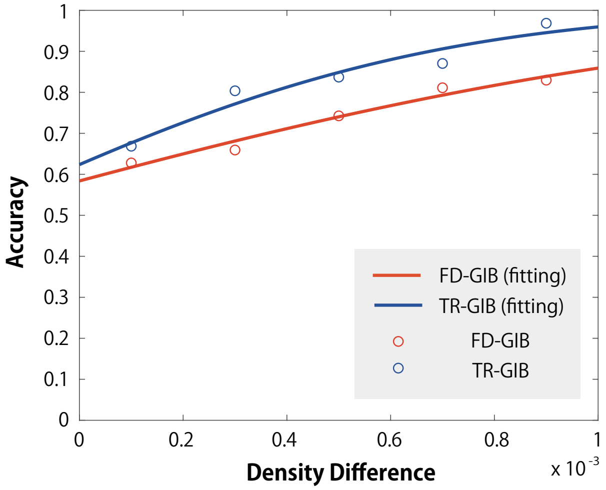


Fig. 8. Relationship between density difference and accuracy in Task 2 with FD-GIB and TR-GIB

4 Results

4.1 Relationship between box size and accuracy

In Task 1, in which only nodes and boxes were displayed, participants performed a task to answer which group had the largest area. Fig. 7 presents the results of this task. For both GIB layouts, the accuracy was high when the side ratio was small (i.e., when the difference between sides was large). No significant difference was detected between FD-GIB and TR-GIB by the Wilcoxon signed rank test.

4.2 Relationship between edge density in group and accuracy

In Task 2, in which the nodes and edges were displayed, participants performed a task to answer which group had the largest number of internal edges. The results of this task is shown in Fig. 8.

For both layouts, the accuracy increased as the density difference increased. The results were modeled by performing fitting (i.e., probit regression analysis) on the results of Task 2 to the cumulative normal distribution function often used in psychophysics. The principle of probit regression analysis is described below. First, the cumulative normal distribution function can be expressed as follows:

$$y \approx n(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} \exp\left\{\frac{-(t-\mu)^2}{2\sigma^2}\right\} dt \tag{16}$$

Table 1. Results of goodness-of-fit test for the modeling in task 3. The null hypothesis (this model formula fits the measured value) is not rejected under the condition of $p > 0.05$.

GIB layout	FD-GIB				TR-GIB			
Edge Ratio	0.98		0.91		0.98		0.91	
Difference of Density	$\Delta D < 0$	$\Delta D > 0$	$\Delta D < 0$	$\Delta D > 0$	$\Delta D < 0$	$\Delta D > 0$	$\Delta D < 0$	$\Delta D > 0$
p-value	0.479	0.182	0.222	0.119	2.57×10^{-4}	5.99×10^{-10}	9.18×10^{-2}	1.65×10^{-3}

In this experiment, it was necessary to determine the unknown parameters μ and σ that best fit the actual data, where x is the density difference of the internal edge and y is the accuracy. The cumulative normal distribution function is a monotonically increasing function, which has an inverse function. If the inverse function is n_{std}^{-1}, the relationship between n and n_{std}^{-1} can be expressed as follows:

$$\begin{aligned} n(x) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} \exp\left\{\frac{-(t-\mu)^2}{2\sigma^2}\right\} dt && (17)\\ &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\frac{x-\mu}{\sigma}} \exp\left\{\frac{-u^2}{2}\right\} \sigma du && (18)\\ \Bigg(\quad \because \quad & u = \frac{t-\mu}{\sigma}\Bigg) \\ &= n_{std}\left(\frac{x-\mu}{\sigma}\right) && (19)\\ & && (20) \end{aligned}$$

Here, the value of the density difference is $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]$, and the accuracy is $\boldsymbol{y} = [y_1, y_2, \cdots, y_N]$. Fitting the relationship between x and y to the normal cumulative distribution function is expressed as follows:

$$y_1 \approx n(x_1) = n_{std}\left(\frac{x_1-\mu}{\sigma}\right). \tag{21}$$

$$n_{std}^{-1}(y_1) \approx \frac{x_1-\mu}{\sigma} \tag{22}$$

Here, n_{std}^{-1} is called a z-value, a value obtained from a numerical table. Performing the same procedure for all data results in the following:

$$\begin{aligned} n_{std}^{-1}(y_1) &\approx \frac{x_1-\mu}{\sigma}\\ n_{std}^{-1}(y_2) &\approx \frac{x_2-\mu}{\sigma}\\ &\vdots\\ n_{std}^{-1}(y_N) &\approx \frac{x_N-\mu}{\sigma}\\ . \end{aligned}$$

Where $n_{std}^{-1}(y)$ is w, $1/\sigma$ is a, $-\mu/\ sigma$ is b, then $w = ax+b$, and μ, σ can be obtained by simple linear regression. Modeling the data obtained in Task 2 using this method makes it possible to plot a regression curve as a solid line in Fig. 8.

4.3 Relationship between box size, density of edges in group and accuracy

In Task 3, in which all elements of the GIB layout –nodes, edges, and boxes– were displayed, participants performed a task to identify which group had the largest number of edges. Fig. 9 presents the results of Task 3. For both GIB layouts, a larger absolute value of the density difference led to a higher percentage of correct answers. The regression curve was obtained in Task 3 as a solid line in Fig. 9. From the goodness-of-fit test, the null hypothesis cannot be rejected except for a difference in density of > 0, where the ratio of the sides in TR-GIB is 0.98 and 0.91.

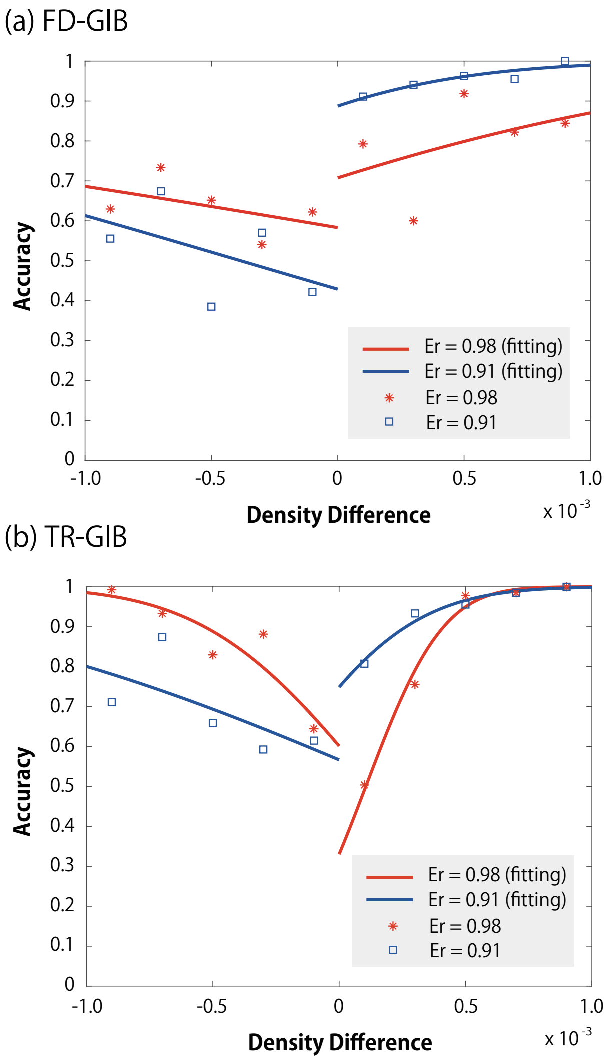


Fig. 9. Results of modeling the relationship between density difference and accuracy in (a) FD-GIB and (b) TR-GIB in task 3

5 Discussion

5.1 Relationship between box size and accuracy

In Task 3, in both GIB layouts, the smaller the ratio of the box sides in the range of density difference > 0 led to the higher accuracy. On the other hand, in the range of density difference < 0, accuracy is lower as the box side ratio is smaller. Table 2 shows the result of testing whether there is a difference in accuracy between conditions of different edge ratio in Task 3.

As shown in Fig. 9 and Table 2, when the edge ratio Er $=$ 0.91, accuracy tends to be significantly increased in the range of the density difference > 0, while accuracy tends to be decreased in the range of the density difference < 0. Therefore, not only the density difference but also the difference in box size is one of the factors in determining the accuracy. This is a natural result because, under the same density difference, the larger difference in box size (Er $= 0.91$) led to the larger difference in the number of internal edges, results in the higher accuracy on task.

On the other hand, since the correct box is the second largest box in the range of density difference < 0, the density information and the size difference information contradict each other in this range, indicating that accuracy got lower than that of the density difference > 0.

Table 2. The result of testing whether there is a significant difference in accuracy by the edge ratio for each density difference of each GIB layout in Task 3 by the Wilcoxon sign rank test. The highlighted part in red is the data showing significant difference ($p < 0.05$).

Difference of Density	-9×10^{-4}	-7×10^{-4}	-5×10^{-4}	-3×10^{-4}	-1×10^{-4}	1×10^{-4}	3×10^{-4}	5×10^{-4}	7×10^{-4}	9×10^{-4}
FD-GIB	0.0709	0.265	2.05×10^{-4}	0.497	6.90×10^{-4}	9.05×10^{-3}	1.57×10^{-5}	0.146	1.10×10^{-3}	6.10×10^{-5}
TR-GIB	5.57×10^{-5}	0.194	9.46×10^{-3}	4.88×10^{-4}	0.547	5.50×10^{-5}	1.24×10^{-4}	0.508	1.00	1.00

Table 3. The result of Wilcoxon's signed-rank test for the difference in accuracy between Task 2 and 3 under each condition. The red highlighted area shows where the accuracy in Task 3 is significantly larger than that in Task 2 ($p < 0.05$), and the blue highlighted area is where the accuracy in Task 3 is significantly smaller.

GIB layout		**FDGIB**		**TRGIB**	
Edge Ratio		**0.98**	**0.91**	**0.98**	**0.91**
Difference of Density	-9×10^{-4}	8.94×10^{-4}	9.21×10^{-6}	0.0625	5.92×10^{-4}
	-7×10^{-4}	0.0881	7.70×10^{-3}	0.0189	0.245
	-5×10^{-4}	0.0182	1.20×10^{-5}	0.961	1.17×10^{-3}
	-3×10^{-4}	3.89×10^{-3}	0.0756	0.0121	3.67×10^{-4}
	-1×10^{-4}	0.942	2.71×10^{-4}	0.521	0.167
	1×10^{-4}	6.50×10^{-4}	1.12×10^{-5}	2.83×10^{-4}	8.17×10^{-3}
	3×10^{-4}	0.126	1.17×10^{-5}	0.285	1.95×10^{-4}
	5×10^{-4}	2.54×10^{-5}	1.73×10^{-5}	1.01×10^{-5}	1.28×10^{-4}
	7×10^{-4}	0.524	8.15×10^{-5}	1.75×10^{-5}	1.67×10^{-5}
	9×10^{-4}	0.645	6.36×10^{-6}	0.0313	0.0313

5.2 Effects of box on accuracy

In Task 2 and Task 3, there was a difference in the display of the box. Therefore, comparing the results of Task 2 and Task 3 makes it possible to confirm the effect of the presence or absence of a box. We hypothesized that the effect of the box is as follows.

- When the edge ratio in Task 3 is 0.98, that is, there is almost no difference in edge ratio, the effect of a box explicitly representing the group size would be small. On the other hand, when the edge ratio is 0.91, that is, the difference between the edge ratio is large, the effect of a box would be large.

For each GIB layout, Fig. 10 shows the result of comparing the accuracy of Task 2 and Task 3. In addition, the results of testing the statistical difference between accuracy in Task 2 and 3 using the Wilcoxon signed-rank test are shown in 3.

When the edge ratio of the FD-GIB was 0.98, a significant difference is confirmed in the density difference 1×10^{-4} and 5×10^{-4} in the density difference > 0, showing that the accuracy of Task 3 was higher than that of Task 2. On the other hand, in the density difference -9×10^{-4}, -5×10^{-4}, -3×10^{-4}, significant differences were confirmed, showing that the accuracy of Task 3 was lower than that of Task 2.

When the edge ratio of FD-GIB was 0.91, significant difference was found except for the density difference -3×10^{-4}. At density differences showing significance, the accuracy of Task 3 was higher than Task 2.

In addition, when the edge ratio of TR-GIB was 0.98, the significant density difference between Task 2 and 3 were confirmed at -7×10^{-4}, -3×10^{-4}, 1×10^{-4}, 5×10^{-4}, 7×10^{-4}, and 9×10^{-4}. The accuracy of Task 3 was higher than Task 2 in the density difference 5×10^{-4}, 7×10^{-4}, 9×10^{-4}.

In the density difference -7×10^{-4}, -3×10^{-4}, the accuracy in Task 3 is higher than that in Task 2, and the density difference 1×10^{-4} had a lower accuracy than Task 2. When the side ratio of TR-GIB was 0.91, significant differences were found except for the density difference -7×10^{-4} and -1×10^{-4}. At density differences showing significance, the accuracy was higher than task 2.

From this result, when the edge ratio was 0.91, the difference between Task 2 and Task 3 was generally large, indicating

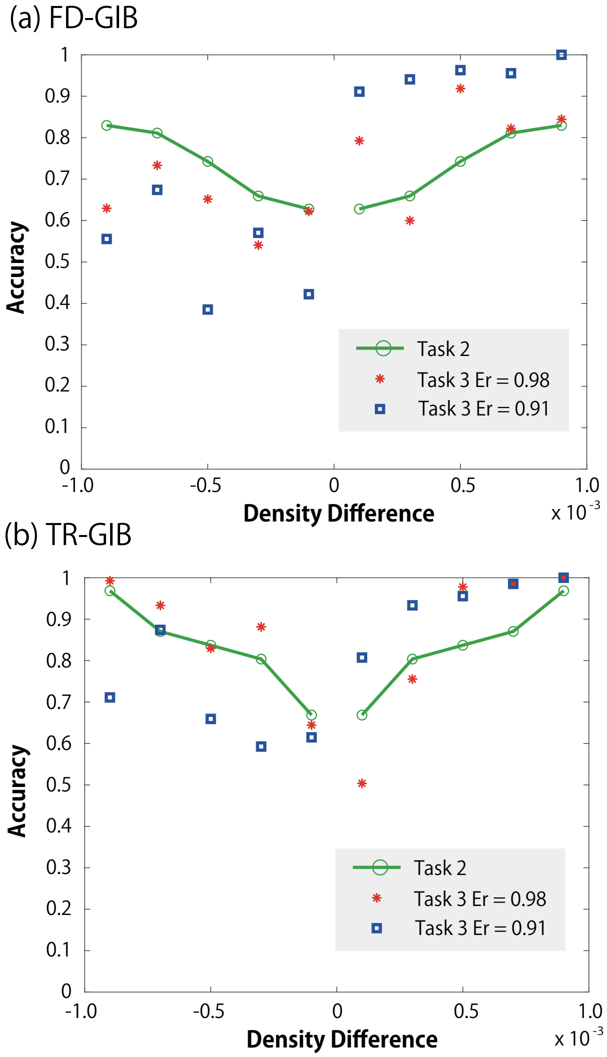


Fig. 10. Relationship between density difference and accuracy in (a) FD-GIB, (b) TR-GIB in Task 2 and 3

that an effect of box would increase the accuracy. The presence of the box that explicitly indicates the group size could allow information on the group size to be used to estimate the difference between the internal edges.

The same effect was partially seen when the edge ratio is 0.98, however, as shown in the blue area of Table 3, the effect of lowering the accuracy was also confirmed by the presence of a box. The effect of lowering the accuracy was observed in which the accuracy was extremely low for both FD-GIB and TR-GIB. Displaying the box explicitly could be an adverse effect in these cases, leads to misjudging the number of internal edges. Thus, the presence or absence of box can cause a fluctuation in accuracy either positively or negatively.

5.3 Comparison of accuracy by two GIB layouts

We examined whether there was a significant difference in accuracy between FD-GIB and TR-GIB. Fig. 11 presents the result of comparing the accuracy of each GIB layout in Task 3.

4 shows the result of testing the significant difference of accuracy of two GIB layouts by the Wilcoxon's signed-rank test in each density difference.

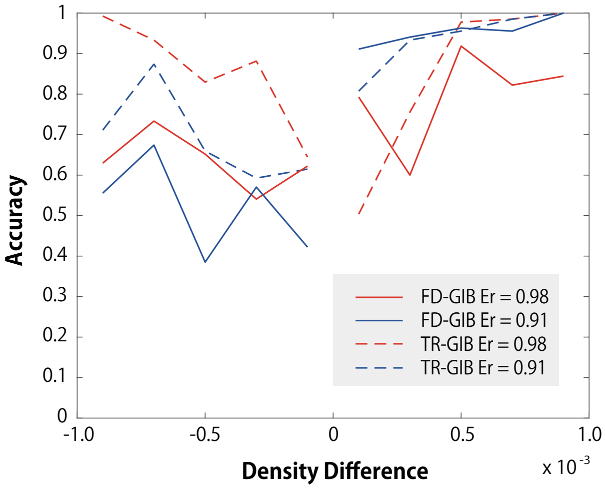


Fig. 11. Accuracy comparison of FD-GIB and TR-GIB in task 3

Table 4. The result of testing whether or not there is a significant difference in accuracy of two GIB layouts in Task 3 by Wilcoxon's signed-rank test. The red highlighted area shows significant difference ($p < 0.05$).

Difference of Density		-9×10^{-4}	-7×10^{-4}	-5×10^{-4}	-3×10^{-4}	-1×10^{-4}	1×10^{-4}	3×10^{-4}	5×10^{-4}	7×10^{-4}	9×10^{-4}
Edge Ratio	**0.98**	9.56×10^{-6}	8.22×10^{-4}	2.99×10^{-4}	1.37×10^{-5}	0.639	1.12×10^{-4}	9.62×10^{-3}	0.0352	8.21×10^{-4}	6.10×10^{-5}
	0.91	0.0128	2.51×10^{-3}	1.75×10^{-4}	0.740	3.80×10^{-3}	0.0357	1.00	1.00	0.289	1.00

In the range of density difference < 0, TR-GIB had a significantly higher accuracy than FD-GIB at either edge ratio except for two cases (Er=0.98, ΔD=-1×10^{-4} and Er=0.91, ΔD=-3×10^{-4}). On the other hand, in the range of density difference > 0, when the edge ratio was 0.98 and the density difference was 1×10^{-4}, FD-GIB showed higher accuracy than TR-GIB. However, other than that, TR-GIB showed higher accuracy. Also, when the edge ratio was 0.91, a significant difference is confirmed only when the density difference was 1×10^{-4}, and the accuracy in FD-GIB is high.

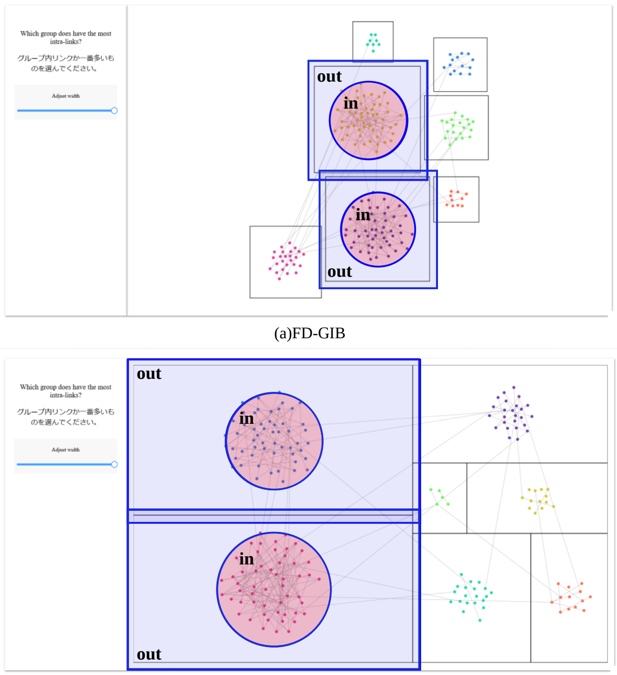
TR-GIB tended to have a higher accuracy rate as a whole, but this is because the paper utilization efficiency was higher in TR-GIB and the area of the box is larger than in FD-GIB, leads to improved visibility. Moreover, in the previous study [3], the correct answer rate of FD-GIB was relatively high, and this study [23] showed that the significant difference in accuracy between the GIB layouts could not be confirmed. The reason for inconsistency is why the difference in density, the area of the box, and the number of boxes used in these experiments are not the same as the current experiment.

5.4 Analysis of eye-tracking data

In this experiment, eye-tracking data was acquired from the subject performing the task. We conducted an AOI-based analysis on eye-tracking data to investigate in detail from the viewpoint of human exploration behavior why the accuracy would change depending on the presence or absence of a box.

It is expected that the size of the graph in the box could be estimated efficiently by displaying the box. If such information is actively used by participants, we hypothesized that it may change the fixation duration inside the graph and fixation duration at the periphery of the box outside the graph depending on the presence or absence of the box. Therefore, in this analysis, only eye-tracking data related to two correct answer candidates was analyzed. AOIs were determined as Fig. 12.

For each box we define an AOI called "in" and "out". "In" is a circle area that encloses all the nodes belonging to the group, and "out" is the area slightly larger area (13.5 pixels margin) than the box excluding "in" area. Although the range of " out " is slightly larger than that of the box, when the side lengths of the box are compared in Task 1, the gaze positions


Which group does have the most intra-links?

グループ内リンクが一番多いものを選んでください。

Adjust width

(a)FD-GIB

Which group does have the most intra-links?

グループ内リンクが一番多いものを選んでください。

Adjust width

(b)TR-GIB

Fig. 12. Definition of AOI. (a) FD-GIB, (b) TR-GIB. The area of the circle surrounding all the nodes belonging to a certain group is defined as ”in” and the area slightly larger than the box excluding ”in” is defined as ”out”.

are distributed at the periphery of the box. To count those peripheral distribution of gaze as a fixation duration of ”out”, we enlarge the AOI of ”out”.

In this experiment, eye-tracking data were analyzed to find out which AOIs the subjects focused. When subjects compare box sizes, the fixation duration for the ”out” region could increases, and when the internal edges are compared, an increase in fixation duration for the ”in” region could increase.

Of the 27 subjects, six subjects were excluded from this analysis because their datasets were noisy. In each trial of each task, we calculated the ratio ”in” against the time when the subject saw ”in” and ”out” for the time from 100 ms after the stimulus onset to the end of the stimulus. The average value of the ratio of AOIs was shown in Fig. 13.

Compared to Tasks 2 and 3 in both GIB layouts, the percentage of fixation duration to see ”in” in Task 1 was lower, indicating that the subjects pay attention to the periphery of a box when comparing the box sizes. The ratio of fixation duration looking at ”in” in Tasks 2 and 3 was high, we tended to focus on the node link diagram in the box itself. No significant difference between the fixation duration of task 2 and task 3 was confirmed in either GIB layout.

Since Task 3 displays a box, we expected to see a lower percentage of fixation duration at “ in ” compared to Task 2. However, from the results of tasks 2 and 3, it can be seen that the subject is focusing on the ”in” region regardless of the presence or absence of the box. The subjects are not focusing on the periphery of the box, but has a strategy of focusing on the node-link diagram. However, the box information obtained by the peripheral vision could influenced the accuracy unintentionally.

When extracting information from a visualization diagram composed of multiple visualization elements such as GIB layout, even if the user is paying attention to the most important element for the task, other visualization elements are also came into view unintentionally during the exploration. The results of this study suggest that these unintended visualization factors may influence the correct answer rate. Therefore, the designer of visualization must design the visualization by carefully considering the characteristics of human visual cognition. To avoid the influence of unintended visualization elements on the GIB layout, one option is to change the way to visualize data interactively according to users’ needs. Taking example of this task, an interactive visualization that directly encodes the number of internal edges could be effective for decrease in misjudges.

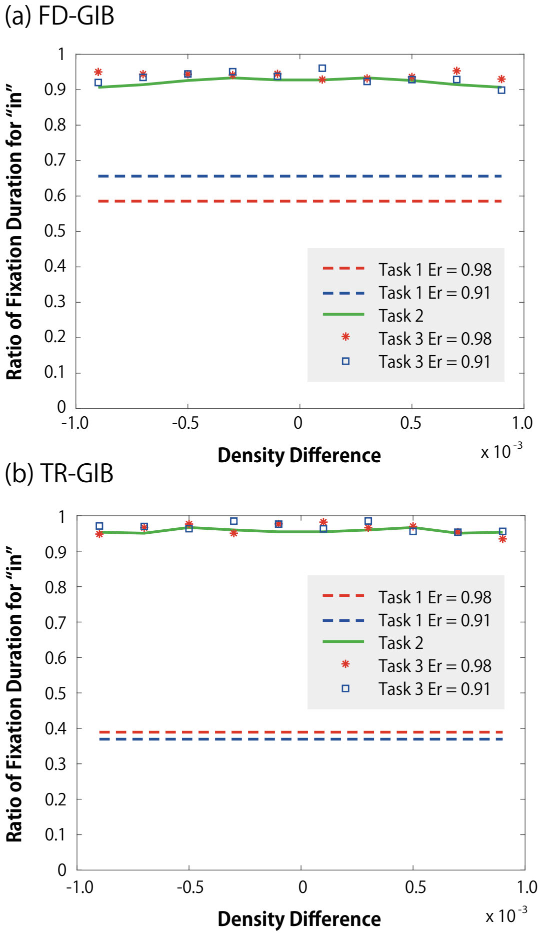


Fig. 13. The ratio of fixation durations of “ in ” to that of “ in ” and “ out ” in (a) FD-GIB layout and (b) TR-GIB layout in each task.

6 Conclusion

We measured the performance and eye movements of participants performing the task of identifying the group with the largest internal edges to investigate the effect of visualization elements on task performance. Twenty-seven subjects performed the task of determining which group has the largest number of internal edges when different combination of visualization elements are displayed. Eye-tracking results revealed that subjects performed tasks focusing on the graphs in the group even when the box was displayed. Nevertheless, the presence and absence of boxes caused fluctuations in accuracy. The GIB layout targeted in this study is an example of a complex visualization. By designing tasks for visualizations other than the GIB layout, it is possible to quantify human behaviors and performances in various cases, leads to obtain useful knowledge for designing visualization techniques.

Acknowledgments

This work was supported by JST CREST Grant Number JP-MJCR1511, Japan. This work was also partially supported by The Keihanshin Consortium for Fostering the Next Generation of Global Leaders in Research (K-CONNEX), established by Human Resource Development Program for Science and Technology, MEXT.

References

[1] G. Andrienko, N. Andrienko, M. Burch, and D. Weiskopf. Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2889–2898, 2012.

[2] N. Aoyama. A website of gib. `http://gib.jgs-hd.com/`. Accessed: 2019-04-01.

[3] N. Aoyama, Y. Onoue, Y. Ueno, H. Natsukawa, and K. Koyamada. User evaluation of group-in-a-box variants. In *Proceedings of IEEE Pacific Visualization 2019 (PacificVis 2019)*, pp. 1–10. IEEE, 2019.

[4] R. A. Becker, S. G. Eick, and A. R. Wilks. Visualizing network data. *IEEE Transactions on visualization and computer graphics*, 1(1):16–28, 1995.

[5] M. Bruls, K. Huizing, and J. J. Van Wijk. Squarified treemaps. In *Data visualization 2000*, pp. 33–42. Springer, 2000.

[6] M. Burch, N. Konevtsova, J. Heinrich, M. Hoeferlin, and D. Weiskopf. Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2440–2448, 2011.

[7] S. Chaturvedi, C. Dunne, Z. Ashktorab, R. Zachariah, and B. Shneiderman. Group-in-a-box meta-layouts for topological clusters and attribute-based groups: Space-efficient visualizations of network communities and their ties. In *Computer Graphics Forum*, vol. 33, pp. 52–68. Wiley Online Library, 2014.

[8] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE transactions on visualization and computer graphics*, 25(1):830–839, 2019.

[9] E. R. Gansner and Y. Hu. Efficient node overlap removal using a proximity stress model. In *International Symposium on Graph Drawing*, pp. 206–217. Springer, 2008.

[10] W. Huang, P. Eades, and S.-H. Hong. Beyond time and error: a cognitive approach to the evaluation of graph drawings. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 3. ACM, 2008.

[11] S.-H. Kim, Z. Dong, H. Xian, B. Upatising, and J. S. Yi. Does an eye tracker tell the truth about visualizations?: findings while investigating visualizations for decision making. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2421–2430, 2012.

[12] S. G. Kobourov. 12 force-directed drawing algorithms. 2004.

[13] R. Netzel, M. Burch, and D. Weiskopf. Comparative eye tracking study on node-link visualizations of trajectories. *IEEE transactions on visualization and computer graphics*, 20(12):2221–2230, 2014.

[14] R. Netzel, M. Hlawatsch, M. Burch, S. Balakrishnan, H. Schmauder, and D. Weiskopf. An evaluation of visual search support in maps. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):421–430, Jan 2017. doi: 10.1109/TVCG.2016.2598898

[15] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri. Face to face: Evaluating visual comparison. *IEEE transactions on visualization and computer graphics*, 25(1):861–871, 2019.

[16] Y. Onoue and K. Koyamada. Optimal tree reordering for group-in-a-box graph layouts. In *SIGGRAPH Asia 2017 Symposium on Visualization*, p. 13. ACM, 2017.

[17] H. Purchase. Which aesthetic has the greatest effect on human understanding? In *International Symposium on Graph Drawing*, pp. 248–261. Springer, 1997.

[18] H. C. Purchase. Performance of layout algorithms: Comprehension, not computation. *Journal of Visual Languages & Computing*, 9(6):647–657, 1998.

[19] H. C. Purchase, D. Carrington, and J.-A. Allder. Empirical evaluation of aesthetics-based graph layout. *Empirical Software Engineering*, 7(3):233–255, 2002.

[20] K. Reda, P. Nalawade, and K. Ansah-Koi. Graphical perception of continuous quantitative maps: the effects of spatial frequency and colormap design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 272. ACM, 2018.

[21] E. M. Rodrigues, N. Milic-Frayling, M. Smith, B. Shneiderman, and D. Hansen. Group-in-a-box layout for multi-faceted analysis of communities. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 354–361. IEEE, 2011.

[22] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):810–819, 2019.

[23] Y. Ueno, H. Natsukawa, N. Aoyama, and K. Koyamada. Exploration behavior of group-in-a-box layouts. *Visual Informatics*, in press, 2019. doi: 10.1016/j.visinf.2019.03.005

[24] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C.-W. Fu, O. Deussen, and B. Chen. Optimizing color assignment for perception of class separability in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 25(1):820–829, 2019.