Evaluating Overall Quality of Graph Visualizations Based on Aesthetics Aggregation

Weidong Huang[a,*], Mao Lin Huang[b,c], Chun-Cheng Lin[d]

[a]*School of Engineering and ICT, University of Tasmania, Australia*
[b]*School of Computer Software,Tianjin University, China*
[c]*School of Software, FEIT, University of Technology, Sydney, Australia*
[d]*Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan*

Abstract

Aesthetics are often used to measure the layout quality of graph drawings and it is commonly accepted that drawings with good layout are effective in conveying the embedded data information to end users. However, existing aesthetic criteria are useful only in judging the extents to which a drawing conforms to specific drawing rules. They have limitations in evaluating overall quality. Currently graph visualizations are mainly evaluated based on personal judgments and user studies for their overall quality. Personal judgments are not reliable while user studies can be costly to run. Therefore, there is a need for a direct measure of overall quality. In an attempt to meet this need, we propose a measurement that measures overall quality based on individual aesthetics and gives a single numerical score. We present a user study that validates this measure by demonstrating its sensibility in detecting quality changes and its capacity in predicting the performance of human graph comprehension. The implications of our proposed measure for future research are discussed.

Keywords: Graph drawing, overall quality, aesthetics, measurement, effectiveness

1. Introduction

In order to take advantage of the powerful human visual perception system, node-link diagrams are often used as a visual tool for the purposes of commu-

*Corresponding author. E-mail: tony.huang@csiro.au

nication and understanding of non-visual graph data. When used for graph data, node-link diagrams are also called graph drawings, and are sometimes simply called graphs, drawings or visualizations if no confusions are caused. However, drawing graphs into node-link diagrams does not automatically make the process of communication and understanding better as a graph can be laid out in very different ways. Empirical research has shown that layout affects how a graph is perceived [27]: a good layout facilitates the process, while a poor layout may hinder the process. Therefore, it is important to know how effective an intended layout is in conveying the underlying data to end users when drawing a graph.

Layout rules, or aesthetics, have been used as quality criteria to guide the choice between layouts. It is commonly accepted that drawings conforming to these aesthetics are of good quality and can be effective [3]. Two examples of these aesthetics are the minimum number of edge crossings and the maximum display of symmetries. In other words, a drawing with fewer crossings and more symmetries are better. However, the existing aesthetics are useful only in judging the extents to which a drawing conforms to specific drawing rules; they have limitations in evaluating overall quality. One of the causes for the limitation was the fact that most of the aesthetics conflict with each other; it is not possible to implement all of them to the fullest at the same time. Optimizing one aesthetic can be achieved only at the cost of other aesthetics, leaving the overall quality uncertain. Figure 1 gives a simple example of conflicting aesthetics. It shows two drawings of a graph. To draw the graph with maximum symmetries, more crossings are required (left). However, maximum symmetries are no longer possible when it is drawn with minimum crossings (right). The conflicting relationship affects current practices of graph visualization greatly. On the one hand, many algorithms for automatic graph drawing are designed to optimize only one or two aesthetics, and different algorithms focus on different aesthetics. This makes it difficult for an algorithm user to choose which algorithm to use when he or she has more than one algorithm candidate at hand. The reason for this is because there is no easy way available to tell whether, for example, an algorithm that is to minimize the number of crossings will produce better drawings than another algorithm that is to maximize symmetries in terms of overall quality.

On the other hand, it is generally acknowledged that the best layout is the balance of aesthetics. This is partly reflected in the fact that force-directed algorithms have been the most widely used tools for graph visualization. This particular type of algorithms simulate a graph as a physical system and assign numerical weight values to forces that represent the aesthetics considered. These forces work together and a balanced layout is reached in the end. This final layout is a

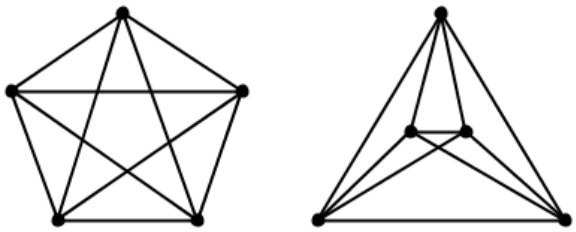

Figure 1: An example of conflicting aesthetics. The left drawing shows maximum symmetries, but with more crossings, while the right one has minimum crossings, but with less dimensions of symmetry

compromise between the forces, or the aesthetics in consideration. In addition, given several drawings of different layouts, by computing and comparing values of aesthetics, we are able to find out which drawing has achieved a better trade-off between aesthetics considered [11]. However, seeking a compromise between aesthetics only gives us a better chance of producing good drawings [21]. Different aesthetics affect human graph comprehension to different degrees. Without conducting a user study, we are unable to know, for example, whether a drawing produced based on one set of aesthetics is better than another drawing that is produced based on a different set of aesthetics. There are no empirically verified guidelines or quality measures available for us to make such type of evaluation at the design stage.

Due to the lack of appropriate measures or methods, graph visualizations are evaluated mainly based on personal judgments and user studies. However, personal judgments are subjective and are not reliable, while user studies can be costly to run and can only be done after the visualization has been completed. Therefore, there is a need for a reliable and objective measure so that we can evaluate overall quality at the early design stage of a visualization process. This measure will help visualization designers to quickly judge or compare the quality of the drawings in consideration and make decisions accordingly.

In an effort toward this need, we propose an overall quality measure of layout. This measure takes into account individual aesthetic criteria and gives a single numerical value. In this paper, we first briefly review current practices of quality evaluation of graph drawings. This is followed by an explanation of how our proposed measure is formulated and computed. Then, we present a user study for the validation of the new measure. This study has two sets of drawing stimuli. The first set of drawings are used to test its sensibility, while the other set are to demonstrate its capacity in predicting the performance of human graph comprehension.

The paper concludes with a general discussion.

The main contributions of this work include:

1. A new overall quality measure was proposed based on the normalized z scores of individual aesthetics.
2. Empirical evidence was provided that demonstrates the sensibility and the predictive capacity of the new measure.
3. We found that the proposed measure was more sensitive to quality changes than traditional performance measures.

2. Related Work

A typical graph visualization system includes two basic components: graph drawings and interaction methods that are intended to present these drawings in specific ways so that information embedded in the drawings can be processed effectively and efficiently by human users. There is a growing body of work on graph evaluation appearing in the literature [9, 17, 18]. This body of work can be divided into three major categories: system evaluation (including interfaces), interaction evaluation and graph drawing evaluation. In this section, we selectively review *evaluations of graph drawings with a focus on quality measurement*. Studies that contribute to a direct measurement of overall quality are also reviewed.

2.1. Evaluation of graph drawing quality

In graph drawing, two different types of quality measures have been used: 1) computational measures, including commonly used aesthetics and specifically developed measures that reflect algorithm goals or graph structural characteristics; and 2) empirical measures, including expert opinion, user preference and task performance.

Computationally, Didimo et al. [11] conducted an experimental study that compared two new topology-driven heuristics with three existing graph drawing algorithms. The quality of the resultant drawings was compared based on the extent to which they conformed to each of a set of readability aesthetics. The aesthetic criteria used for comparison included the number of crossings, crossing angle resolution, geodesic edge tendency and vertex angle resolution. The results indicated that drawings of the topology-driven algorithms had better trade-offs between these criteria than others. Similar approaches were also used by Di Battista et al. [4] and by Argyriou et al. [2].

In comparing stress-minimization algorithms for drawing offline dynamic graphs, Brandes and Mader [7] measured quality based on stress and stability. The former was measured as the proportion of the stress of the baseline layout and the stress of the current layout, while the latter was measured as the relative decrease of positional difference of the current layout relative to the baseline layout. Relating these measures to the baseline layout was to normalize them over the sequence graphs that were of different sizes. In this study, three variants of the stress-minimization approach: aggregation, anchoring and linking, were compared. The results indicated that linking was more preferable over the other two in terms of stress and stability. In another study, Brandes and Pich [8] compared distance-based graph drawing algorithms. In this study, the drawing quality was measured as how well the Euclidean distance between any two nodes represented their graph-theoretic distance. The results suggested that minimization of weighted stress yielded better layout than force-directed placement in terms of pairwise distances between nodes. Similar quality measurements were also developed in other studies [6, 30].

Empirically, Hachul and Junger [14] conducted a study that compared algorithms for drawing general large graphs. In this study, the authors evaluated the quality of drawings based on how well their individual layout displayed the graph structure. Fruchterman and Reingold [13] developed a force-directed algorithm. A number of drawing examples were given to demonstrate that the proposed method was able to generate the best possible layout for graphs that have regular structure patterns, such as symmetric and planar graphs. In these studies, the quality of drawings was evaluated largely based on the authors' expert opinions.

User preference and task performance have been used either independently or jointly for quality evaluation in various studies. In those studies, drawing quality was typically measured by one or more of the following metrics: rating scores, task completion time, response accuracy, and mental effort. Despite their common usage, there is one issue with these metrics. That is, they are not necessarily always consistent with each other. For example, the most preferred layout may not be the one that produces the best task performance. A layout that takes less time may be the one that induces more errors. This inconsistency makes the judgment of overall quality difficult. In response to this difficulty, Huang et al. [20] proposed a quality measure, *visualization efficiency*, that takes all these metrics into consideration and yields a single score. This measure has been used in recent studies [21].

The above-mentioned approaches are necessary and important at certain stages and in certain circumstances of a visualization process. However, a direct objec-

tive measure of overall quality is missing. Despite this, some attempts have been made towards this direction which we present is the next sub-section.

2.2. Studies towards a direct quality measure

Ware et al. [29] conducted a study in which subjects were asked to find the shortest path between two pre-specified nodes and indicate the length of the path. Task response times were recorded and the aesthetics in consideration were computed. The relationships between these aesthetics and the recorded times were investigated based on linear regression tests. This study resulted in an equation that could be used to predict the response time of the shortest-path search task based on aesthetics and task specific measures. In this equation, the aesthetics were path continuity and the number of crossings on the path, while the task specific measures were the path length and the number of branches on the path.

In her pioneering work toward normalized measurements of aesthetics, Purchase [26] developed continuous metrics for seven commonly used aesthetics a decade ago. Each metric gives a real number between 0 and 1, which helps to quantify the extent to which a drawing conforms to the corresponding aesthetic criterion. These normalized metrics not only help to evaluate drawing quality in terms of individual aesthetics on the same scale, but also have potential to evaluate overall quality by adding them together. Taking a similar approach, Taylor and Rodgers [28] quantified a set of graph drawing and graphical design based criteria within a range between 0 and 1. The criteria included uniform edge length, angular resolution and homogeneity. Then, a weighted sum of the scores of these criteria was used as an overall quality index for a hill climbing optimization system.

These studies are useful steps towards a direct measure of overall quality, further studies are needed though. Firstly, we need an overall quality measure that is generic, rather than task specific. Secondly, quantifying aesthetics within a bounded range can be difficult to achieve as not all aesthetics are bounded in nature. Thirdly, empirical evidence is needed to validate quantified aesthetics for their relevance to human graph comprehension.

3. The Proposed Measure of Overall Quality

The existing aesthetics have either been empirically validated or widely acknowledged for their association with human graph comprehension [3]. Further, when it comes to the performance of human graph comprehension, each aesthetic has a role to play, and it is the joint effect of these aesthetics that is more relevant

to the overall quality. We therefore propose to measure overall quality as a function of individual aesthetics. In particular, suppose we have n aesthetic criteria denoted as x_1, x_2 ..... x_n, then overall visual quality (y) can be measured as an aggregation of these aesthetics as below:

$$y = \sum_{i=1}^{n} x_i \tag{1}$$

Although being subjective, it is our belief that for this measure to be generically useful, aesthetics to be considered in equitation 1 should be context or application independent, be applicable to general graphs and reflect specific local features of layout. Keeping these requirements in mind, we chose the following four of the most discussed aesthetics as the elements of equation 1 for the purpose of this research:

1. Minimize the number of edge crossings (*cross#*)
2. Maximize crossing angle resolution (*crossRes*)
3. Maximize node angular resolution(*angularRes*)
4. Uniformize edge lengths ($uniEdge$)

Among these aesthetics, *cross#* is measured as the number of crossings in the drawing; a smaller value is better. *CrossRes* is measured as the minimum size of all crossing angles; a larger value is better. *AngularRes* is measured as the minimum size of angles formed by any two neighboring edges; a larger value is better. $uniEdge$ is measured as the standard deviation of all edge lengths; a smaller value is better. These aesthetics are measured on different scales. Their z scores are used instead to be able to combine them into a single measure.

To give an example, suppose that we have three drawings of the same graph that we would like to compare, and they have 2, 7, and 6 crossings, respectively. That is, the scores (x) of *cross#* are 2, 7 and 6. The mean of these three scores ($Mean$) is 5 and their standard deviation ($StDev$) is 2.65. Then, the z score of *cross#* for each of the three drawings can be computed as below:

$$z_{cross\#} = \frac{x - Mean}{StDev} \tag{2}$$

and it is -1.14, 0.76 and 0.38, respectively. The other aesthetics can be standardized into z scores in the same way.

In aggregating z scores, the scales must be made toward the same direction. That is, higher values are always better. As such, equation 1 can be refined and the overall quality score (O) can be computed as below:

$$O = -z_{cross\#} + z_{crossRes} + z_{angularRes} - z_{uniEdge} \quad (3)$$

It is important to note that firstly, this proposed measure is useful only when different visualization conditions are involved. A single absolute value of overall quality does not make much sense since we do not have a baseline score. Secondly, the computed quality score (O) can be zero or negative. When this is the case, it does not necessarily mean that the drawing is bad. If it is negative, it means only that negative aesthetics outweigh positive ones in equation 3. If it is zero, it indicates that negative and positive aesthetics are balanced.

4. Experiment

For a measure to be useful during the design stage of a visualization process, it should meet the following requirements:

1. Be objective (give the same value when used by different assessors);
2. Be reliable (give the same value when used at different times);
3. Be easy to measure (take only a few steps to compute);
4. Be comparable (be able to give continuous numerical values, rather than categorical);
5. Be sensitive to changes (be able to tell the difference when there is a change in quality);
6. Be predictive of human graph comprehension performance (we visualize graphs for people to understand the underlying data. An effective quality measure should positively correlate with the performance of human graph comprehension).

It is apparent that the proposed measure meets the first four requirements. In this section, we describe an experiment that was designed to validate this measure by testing its sensibility and predictability.

4.1. Design

Sensibility was tested by examining whether the quality measured by the proposed measure was consistent with the actual quality. More specifically, we generated a set of random graphs with similar structures. For each graph, a number of drawings were produced with different quality levels. These drawings were generated using a specific graph drawing algorithm so that the relative quality levels between these drawings were known beforehand. The overall quality of these

drawings was also computed using the proposed measure. The resulting scores were then tested to see whether they were consistent with the pre-known quality.

Predictability was tested by examining whether there existed a positive correlation between the measured quality and the actual user task performance. More specifically, we used a random graph with reasonable size and complexity. A number of drawings of this graph were randomly generated. However, this time we did not know their overall quality beforehand. Instead, we ran a user study in which subjects were recruited to perform typical graph reading tasks on these drawings. We recorded task response time, accuracy, cognitive load and visualization efficiency. We also computed overall quality of each drawing using the proposed measure. We then ran regression tests to see whether the task performance data were significantly correlated with the measured overall quality.

Further, as mentioned in Section 2.1, task performance measures are often used to evaluate the quality of drawings. We would like to compare how performance measures and the proposed measure performed in differentiating drawings in terms of overall quality. We therefore included the drawings of the sensibility test in our user study, and subjects were asked to perform tasks on these drawings as well.

As a result, the experiment included two blocks of the drawings: one block for sensibility and the other for predictability. The experiment employed a within-subject design. That is, each subject performed the task in each of the conditions. All subjects performed the task online with a custom-built testing system. To reduce the learning effect, the following two precautions were taken:

1. Subjects were given sufficient training and practice for them to get familiar with the task, the testing system and the procedure before the experiment began.
2. The drawings of the two blocks were mixed together and presented in a random order.

4.2. Stimuli

For sensibility, we generated 20 different graphs. To ensure that the obtained graphs had internal structures as similar as possible, these graphs were all generated using the Erdos-Renyi model of random graphs [12] with each having 30 nodes and 40 edges. These graphs then were drawn using a force-directed algorithm. A force-directed algorithm applies forces on the nodes and edges of a random initial layout, and moves them accordingly [5]. This process is repeated until an equilibrium state is reached. It is known that each time the process is

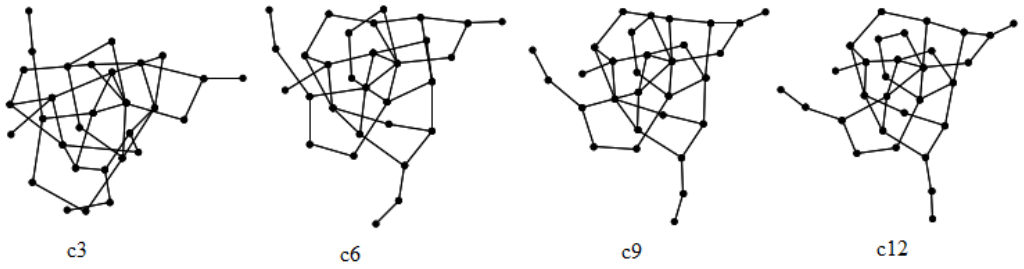

Figure 2: An example of the four condition drawings of a graph for sensibility

Figure 3: Three examples of the thirty drawings of a graph for predictability

repeated, the overall layout is generally improved. To create experimental conditions, we recorded the layout when the process had been repeated 3000, 6000, 9000 and 12000 times for each graph. As a result, four conditions were obtained: c3, c6, c9 and c12. Each graph had one drawing in each of the four conditions, and each condition had 20 drawings. The drawing quality improved across the conditions from c3, c6, c9 to c12. Figure 2 shows an example of the four condition drawings of a graph.

For predictability, we used a graph that had 39 nodes and 48 edges. This graph was drawn thirty times using a force-directed algorithm, resulting in 30 drawings in total. Each of these drawings was obtained by using a random combination of a different initial layout, a different number of iterations for convergence and a range of parameters that were used to define the forces. Figure 3 shows three examples of them.

To avoid possible fatigue or boredom caused by too many drawings, the twenty graphs for sensibility were divided into two halves (note that experiment also included drawings for predictability). Only one half was chosen in an alternating order and the corresponding drawings of the chosen half in the four conditions

were used in a trial. As a result, each subject viewed 10 (drawings) $\times$ 4 (conditions) = 40 drawings for sensibility and 30 drawings for predictability. That is, 70 stimuli in total.

4.3. Subjects

Thirty-five subjects volunteered to participate in the study. Four of them were summer vacation undergraduate students from the CSIRO ICT Centre, Australia. The rest were third-year students from the National Chiao Tung University, Taiwan. All these subjects had normal or corrected-to-normal vision at the time of the experiment.

4.4. The task

There are a range of graph reading tasks. Some involve nodes; some involve edges; some involve paths; and some involve a mix of them. Among these tasks, path searching is the one that is the most frequently required and the one that can be affected by layout the most. In the field of graph theory, a path between two nodes is a sequence of edges, in which the target node of an edge is the source node of the next edge. The length of a path is the number of edges it has. It is possible that there are two or more paths between any two nodes in a graph. The shortest path is the one that has the least number of edges. It is also possible that there are two or more shortest paths between two nodes. Finding a shortest path is often a component of path searching. Therefore, the task of shortest path search has been widely used in graph related user studies. This task was also used in our study; subjects were asked to response by indicating the length of the shortest path found.

For each graph, two nodes were randomly chosen for the task with two preconditions. The first condition was that there was only one shortest path between them. This guaranteed that whenever the same two nodes were specified, the same path was searched by subjects. The second condition was that the shortest path length was between 3 to 5 inclusive for sensibility graphs and was 4 for the predictability graphs. This limitation on the path length was to make the task neither too simple nor too complex to search. Given a graph, the same path was used across the drawings of it for a subject, and the path to be searched could be different between subjects. Using different paths for different subjects was to ensure that the impact of overall layout, rather than a specific part of a drawing, was reflected in the performance data. Further, the reason of having the length of four for predictability drawings was that in this case, each subject contributed

only one data entry to each drawing; using the same-length paths helped to remove possible impact in the performance data caused by varied lengths.

To prevent any possible bias during the task performance, subjects were not made aware of the fact that some drawings were of the same graph. Furthermore, the same path was used for the thirty drawings of the predictability block, which meant that the same path would be searched thirty times by a subject. Although the actual layout of the path was different across the drawings, we added a constraint to the random order of stimuli. That is, no two drawings from the predictability block were displayed consecutively.

4.5. The online system

A custom-built system was used to display the drawing stimuli. The system was designed to highlight the two pre-specified nodes as red. For each drawing, one of the two nodes was displayed first. Subjects were asked to look at the node and hit the space key on the keyboard. Once the key was hit, a drawing screen was shown, on which the whole drawing was displayed on the left hand side. According to the three-stage search model of Korner [16], people locate and identify the relevant graph nodes before reasoning their relationships. It was possible that before a path search was started, some subjects located the two nodes quickly by chance, while others took some time to find them. Having one node displayed first was to reduce such discrepancy between subjects and to ensure that the path search was always started with one of the two red nodes. Once the whole drawing was displayed, subjects started looking for the answer. Once the answer was found, subjects were required to hit the space key immediately. Once the key was hit, the time spent for the answer was recorded and an answer screen was shown.

As shown in Figure 4, there were two sets of boxes on the right hand side of the answer screen. One set of six boxes were above the other set of nine smaller boxes. There was a number just above each box. The numbers for the above six boxes represented possible answers to the task, while the numbers for the nine smaller boxes indicated possible levels of mental effort devoted for the task (from 1 being the lowest to 9 being the highest). Subjects were required to respond by clicking on one box for each set using the mouse. If a box was clicked, a red "$\times$" sign appeared in the box clicked. After the answers had been given, subjects hit the space key on the keyboard to have the answers recorded by the system and to continue with the next drawing. If a subject gave a wrong response to the path search task, the system would emit a beep to remind the subject to be more careful. This process was repeated until of the last drawing had been viewed.

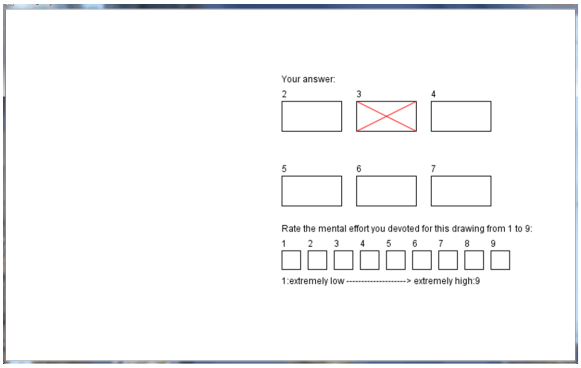

Figure 4: The answer screen

In this system, the design of displaying drawings on the left hand side and answer boxes on the right was based on the user feedback and observations of our pilot studies. Since mouse clicks were used only when the answer screen was shown, displaying the answer boxes and drawings in the middle of the corresponding screens would result in the mouse cursor appearing on the drawing when the system switches from the answer screen to the drawing screen. Subjects would have to move the mouse cursor away first to be able to see the drawing clearly, which interfered with the task performance. Displaying them on the different sides of the screens avoided such interference.

4.6. Experimental documents and procedure

The experimental documents included participation consent forms, information sheets and tutorial materials. The tutorial materials included documents that explained concepts about graphs, shortest paths and node-link diagrams, described the testing system, the procedure, the task and requirements for performing the task. Written Questions and examples were also provided as part of the documents for the purpose of self-test to ensure that everything was understood by subjects.

Subjects were given time to read the documents, sign the form, understand the graph concepts and the task, conduct the self-test, practice with the system. The pictures used for practice were different from those used in the formal experiment. Subjects were also free to ask questions before the experiment. They were told to perform the task as quickly as possible without compromising accuracy.

Once ready, subjects indicated to the experimenter and the experiment started. At any time when the answer screen was displayed, subjects could have a break

Table 1: Mean Values of Dependent Variables

Variable	C3	C6	C9	C12
Time (sec.)	9.91	9.51	7.19	7.11
Effort	3.60	3.26	3.27	3.09
Accuracy	0.69	0.75	0.76	0.76
Efficiency	-0.74	-0.22	0.28	0.48
Overall quality	-2.14	0.04	1.02	1.08

as long as they wished before hitting the space key to move to the next drawing. Therefore, the pace of the experiment was controlled by subjects in order to prevent fatigue. There was also a compulsory 2-minute break when a half of the stimuli had been viewed, allowing them to recover their concentration for the rest of the experiment. After the online task was completed, subjects were asked about their experience and debriefed about the study purposes. The whole session took about 40 minutes on average for each subject, including tutoring and break time. Drinks and snacks were provided after the study.

4.7. Results

4.7.1. Sensibility test

Firstly, we test how performance measures performed in reflecting actual drawing quality. Thirty-five subjects each viewed 10 (drawings) $\times$ 4 (conditions) = 40 drawings and on each drawing, a shortest path task was performed. Task completion time, responses to the task and mental effort were recorded during the experiment. Based on the obtained data, visualization efficiency (E) was computed using the following equation [20].

$$\mathrm{E} = \frac{z_A - z_T - z_{ME}}{\sqrt{3}} \tag{4}$$

In equation 4, z_A, z_T and z_{ME} are standardized z-scores of accuracy (A), time (T) and mental effort (ME), respectively. Visualization efficiency offers us insight into the overall task performance. In the end, we obtained 40 (drawings) $\times$ 35 (subjects) = 1440 experimental data entries for each of the four dependent variables: time, effort, accuracy and efficiency. For each condition, the mean value was computed across the drawings within the condition. The results are shown in Table 1.

Table 2: Results of ANOVA with Post-Hoc Comparisons

Variable	F-statistics	p	Condition pairs with a significant difference
Time	2.804	0.048	(c3, c9), (c3, c12)
Effort	7.491	0.000	(c3, c6), (c3, c9), (c3, c12)
Accuracy	2.118	0.108	none
Efficiency	7.442	0.000	(c3, c6), (c3, c9), (c3, c12)
Overall quality	28.596	0.000	all pairs but (c9, c12)

It can be seen that the subjects generally spent less time, exerted less effort and were more accurate while the drawing quality improved from c3 to c12. The performance efficiency also showed a clear increase across the conditions valued at -0.74, -0.02, 0.28 and 0.48, respectively. In other words, the means of the performance measures were in good agreement with the pre-known overall quality.

To test whether these trends of change were statistically significant at the significance level of 0.05, we ran repeated ANOVA tests with post-hoc comparisons of the Least Square Difference (LSD) method [22] on each of the dependent variables. The results are shown in Table 2.

It can be seen that there was a significant main effect on time, $F(3,57) = 2.804$, $p = 0.048$, on effort, $F(3,57) = 7.491$, $p < 0.001$, and on efficiency, $F(3,57) = 7.442$, $p < 0.001$, but not on accuracy, $F(3,57) = 2.118$, $p = 0.108$. Post-hoc comparisons revealed that these dependent variables had different levels of capacity in detecting condition differences. More specifically, out of the six condition pairs in total, time data only found that two pairs of the conditions were different; effort found three; and efficiency found three, while no difference was shown in accuracy between any pair of the conditions.

Secondly, the overall quality of each drawing was computed using equation 3 that we defined in Section 3. The mean value was computed across the drawings in each condition. The results are shown in the bottom row of Table 1.

It can be seen that the measured overall quality increased while the pre-known quality increased across the conditions from c3 to c12. To see whether this trend of increase was statistically significant, we ran a repeated ANOVA with post-hoc comparisons. The results are shown in the bottom row of Table 2. The repeated ANOVA indicated that there was a significant main effect on overall quality, $F(3,57) = 28.596$, $p < 0.001$. Post-hoc comparisons indicated that all

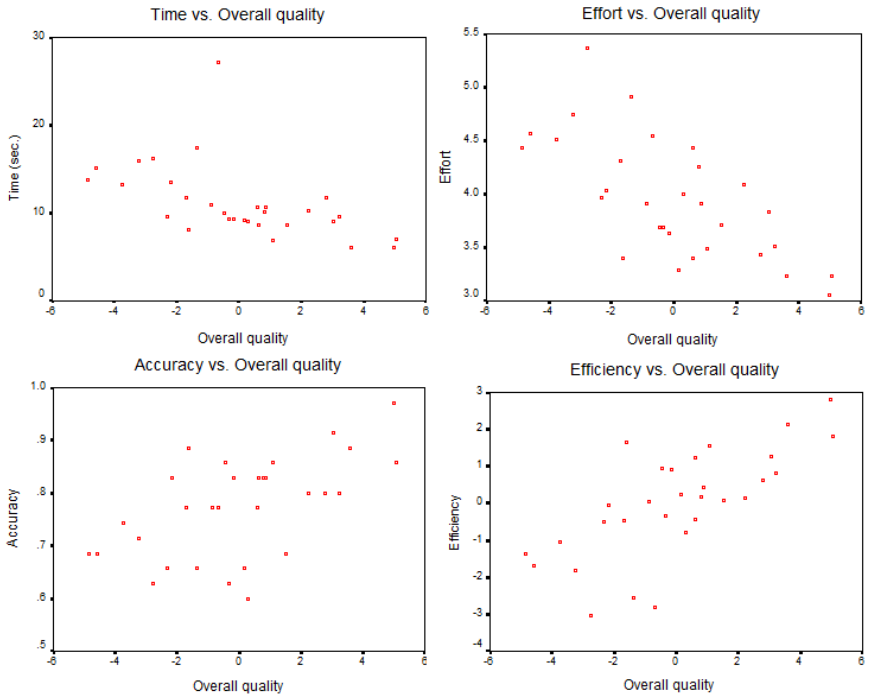

Figure 5: Scatter diagrams between dependent variables and overall quality

conditions were different from each other, except the condition pair of c9 and c12.

4.7.2. Predictability test

Thirty-five subjects each viewed 30 drawings for the shortest path task. During the study, the time they spent, their responses to the task and the exerted mental effort were recorded. For this part of the study, the dependent variables were time, accuracy, effort and efficiency, while the predictor variable was overall quality. For each dependent variable, 30 (drawings) $\times$ 35 (subjects) = 1050 data entries were obtained. For overall quality, equation 3 was used to compute the value for each drawing.

We expected that the measured overall quality was negatively correlated with time and effort, and positively correlated with accuracy and efficiency. To test our hypotheses, we first plotted the score of overall quality and the scores of each dependent variable as Cartesian coordinates to generate a scatter diagram, with overall quality on the horizontal axis and the dependent variable on the vertical axis. This was to have a general idea about the relationships between the dependent variables and the predictor. The obtained diagrams are shown in Figure 5.

Table 3: Results of Simple Linear Regression Tests

Dependent Var.	Predictor	β	F-statistics	R^2
Time	overall quality	-0.539	11.478 **	0.291
Effort	overall quality	-0.692	25.796 ***	0.480
Accuracy	overall quality	0.575	13.835 **	0.331
Efficiency	overall quality	0.717	29.625 ***	0.514

Notes: **: $p < .01$; ***: $p < .001$

In these scatter diagrams, as expected, it appeared that overall quality had a negative correlation with time and effort, and a positive correlation with accuracy and efficiency.

Then, we ran simple linear regression tests to see whether the observed correlations were statistically significant [23]. We regressed each dependent variable on overall quality, and the results are shown in Table 3. In this table, along with F statistics and p values, values of R^2 and β are also reported. R^2 represents the proportion of the variance in a dependent variable that is explained or predicted by the predictor, i.e., overall quality. β is a standardized coefficient. The size of β indicates the strength of the relationship, or the effect size that overall quality has on a dependent variable; it represents the amount of change in standard deviations for the dependent variable that is produced by one standard deviation increase in overall quality. The sign of β implies the direction of that change. According to common rules of thumb suggested by Cohen [10], effect size is small if β is less than 0.10, is large if β is more than 0.50, and is medium if β is between 0.10 and 0.50. For example, in Table 3, β is -0.539 for the regression test of time on overall quality. This means that each standard deviation increase in overall quality will lead to 0.539 of the standard deviation *decrease* in response time, indicating a strong negative correlation between overall quality and time.

Specifically, as shown in Table 3, the overall regression test of time was significant, $F(1,28) = 11.478$, $p < 0.01$. Time was negatively correlated with overall quality, $\beta = -0.539$. Overall quality explained 29.1% of the variance in time. Similarly, regression tests of effort, accuracy, efficiency were all also significant with either $p < 0.01$ or $p < 0.001$.

4.8. Discussion

Our data analysis for sensibility revealed that there was a significant overall difference shown in the data of performance measures including time, effort and efficiency, but not in the accuracy data. This indicated that the subjects had followed the instructions closely and did not compromise accuracy for speed in performing their tasks; the same level of accuracy was achieved across the four conditions, no mater whether the drawing quality was low or high. The analysis also showed that the proposed overall quality measure was able to identify the quality difference between the drawings in the four conditions as expected.

The further post-hoc pairwise comparisons revealed that the proposed measure was able to find more condition pairs being different than any of the task performance and visualization efficiency measures did (see Table 2). In particular, the measure of accuracy failed detect the difference between any of the condition pairs. This, on the one hand, indicated that the actual differences between conditions were sufficiently small for testing sensibility. On the other hand, it indicated that the proposed overall quality measure was more sensitive to quality changes than performance measures. This should not be surprising if we consider that the proposed measure measures overall quality directly, while performance measures measure indirectly. In addition, factors associated with any human experiment could negatively affect performance measures in evaluating overall quality. More specifically, among other factors, human factors, methodological issues, data analysis methods and the choice of tasks can each have a certain role in the evaluation process, affecting the measures in revealing the truth about the quality in one way or another.

Our data analysis for predictability revealed that each of the dependent variables had a significant correlation with the predictor variable, overall quality. And the significant correlations came with large effect sizes as shown in the β values.

In summary, our tests demonstrated that given a graph, the proposed measure was able to not only differentiate drawings based on overall quality, but also significantly predict the performance of human graph comprehension. In other words, the proposed measure is a valid measure of overall quality.

5. General Discussion

In this paper, we reviewed the current practices of quality evaluation in graph drawings and proposed a measure that measured overall quality based on aggregation of aesthetics. A user study was presented that demonstrated the sensibility

of the proposed measure in detecting quality changes and the capacity of it in predicting task performance of human graph comprehension. It was also found that the proposed measure was more sensitive to quality changes than performance measures, thus being a better option for measuring overall quality.

Given these findings, it is important to note that the comparison we made with performance measures was only to demonstrate the sensitivity of our new measure. And this should not be interpreted as an argument of one measure being used against the other. Both types of measures are related, but are essentially different serving for different purposes. On the one hand, the overall quality measure serves as a handy computational tool that can be used by visualization designers to quickly evaluate relative quality between drawings at the early stages of a visualization process. On the other hand, performance measures are used for summative evaluation so that we could understand how effective a visualization is in conveying the embedded information to end users.

One limitation of the proposed measure is that equation 3 assumes a linear relationship between overall quality and each of its component aesthetics. And the reality can be more complex. For example, it has been found that there existed a significant quadratic relationship between the size of crossing angles (*crossRes*) and the drawing quality [19]. Another limitation is that only four aesthetics were considered in our measurement. It is not known what it would be if more aesthetics were considered. Although more studies are needed to clarify these, the empirical evidence presented in this paper has shown that our proposed measurement in its current form does give valid and useful insights into the relative overall quality between drawings.

Given the fact that it has been identified as being the most important factor and found having the greatest impact on humans reading graphs [27], the aesthetic of crossings is often used to judge the layout quality out of convenience. One might argue why we need another measurement. First of all, the number of crossings does not equal to layout quality as a graph with one more crossings is not necessarily less readable. Secondly, more and more evidence has emerged from recent research demonstrating that crossings may not be as important as we normally think [15], and that in some cases, more-crossings drawings can be perceived as more readable or more desirable [24]. Thirdly, it is the joint effect of all aesthetics that affects the overall quality. Considering crossings only is likely to be misleading.

Finally, like any other empirical studies, our experiment has limitations itself [25]. For example, only the shortest path task was used; the proposed measure would have been better evaluated with a wider range of tasks. Therefore,

more studies are needed to further refine and verify the validity of our proposed measure.

References

[1] D. Archambault, H. C. Purchase, and Bruno Pinaud, "Animation, Small Multiples, and the Effect of Mental Map Preservation in Dynamic Graphs," *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539-552.

[2] E. N. Argyriou, M. A. Bekos, and A. Symvonis, "Maximizing the total resolution of graphs," *Proceedings of the Symposium on Graph Drawing (GD'10)*, Springer-Verlag, 62-67, 2010.

[3] G. di Battista, P. Eades, R.Tamassia, and I. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice Hall, 1998.

[4] G. di Battista, A. Garg, G. Liotta, R. Tamassia, E. Tassinari, and F. Vargiu, "An experimental comparison of four graph drawing algorithms," *Computational Geometry: Theory and Applications*, vol. 7, no. 5-6, pp. 303-325, 1997.

[5] U. Brandes, "Drawing on Physical Analogies," In Michael Kaufmann and Dorothea Wagner (Eds.): *Drawing Graphs: Methods and Models*. LNCS Tutorial 2025, pp. 71-86, Springer-Verlag, 2001.

[6] U. Brandes, M. Gaertler, and D. Wagner, "Engineering Graph Clustering: Models and Experimental Evaluation," *ACM Journal of Experimental Algorithmics*, 12, Article 1.1, 2007.

[7] U. Brandes, and M. Mader, "A Quantitative Comparison of Stress-Minimization Approaches for Offline Dynamic Graph Drawing," *Proceedings of the Symposium on Graph Drawing (GD'11)*, pp. 99-110, Springer-Verlag, 2012.

[8] U. Brandes, and C. Pich, "An Experimental Study on Distance-based Graph Drawing," *Proceedings of the Symposium on Graph Drawing (GD'08)*, pp. 218-229, Springer-Verlag, 2009.

[9] M. Burch, N. Konevtsova, J. Heinrich, M. Hoferlin, and D. Weiskopf, "Evaluation of Traditional, Orthogonal, and Radial Tree Diagrams by an Eye Tracking Study." *IEEE Trans. Vis. Comput. Graph.*, 17(12): 2440-2448, 2011.

[10] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Routledge Academic; 2nd edition, 1988.

[11] W. Didimo, G. Liotta, and S. Romeo. "Topology-driven force-directed algorithms," *Proceedings of the Symposium on Graph Drawing (GD'10)*, Springer-Verlag, 2010, pp. 165-176.

[12] P. Erdos and A. Renyi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290-297, 1959.

[13] T. Fruchterman, and E. Reingold, "Graph drawing by force-directed placement," *SoftwarePractice & Experience*, vol. 21, no. 11, pp. 1129-1164, 1991.

[14] S. Hachul, and M. Junger, "An experimental comparison of fast algorithms for drawing general large graphs," *Proceedings of the Symposium on Graph Drawing (GD'05)*, Springer-Verlag, 235-250, 2005.

[15] S. G. Kobourov, S. Pupyrev, and B. Saket, "Are Crossings Important for Drawing Large Graphs?" *Proceedings of the Symposium on Graph Drawing (GD'14)*, Springer-Verlag, 234-245, 2014.

[16] C. Korner, "Eye movements reveal distinct search and reasoning processes in comprehension of complex graphs," *Applied Cognitive Psychology*, Volume 25, Issue 6, pages 893-905, 2011.

[17] C. Korner, and D. Albert, "Speed of comprehension of visualized ordered sets," *Journal of Experimental Psychology: Applied*, 8, 57-71, 2002.

[18] W. Huang, *Handbook of human centric visualization*, Springer, 2014.

[19] W. Huang, P. Eades, and S.-H. Hong, "Effects of crossing angles," *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis'08)*, IEEE Press, 2008, pp. 41-46.

[20] W. Huang, P. Eades, and S.-H. Hong, "Measuring effectiveness of graph visualizations: a cognitive load perspective," *Information Visualization*, vol. 8, no. 3, pp. 139-152, 2009.

[21] W. Huang, P. Eades, S.-H. Hong, and C.-C. Lin, "Improving multiple aesthetics produces better graph drawings," *Journal of Visual Languages & Computing*, Volume 24, Issue 4, August 2013, Pages 262-272.

[22] D. C. Howell, *Statistical Methods for Psychology*, Sixth Edition, Thomson Wadsworth, 2007

[23] T. Z. Keith, *Multiple Regression and Beyond*, Published by Allyn & Bacon, 2005.

[24] P. Mutzel, "An Alternative Method to Crossing Minimization on Hierarchical Graphs." *Proceedings of the Symposium on Graph Drawing (GD'96)*, Springer-Verlag, 318-333, 1996.

[25] H. C. Purchase "A healthy critical attitude: Revisiting the results of a graph drawing study." *J. Graph Algorithms Appl.* 18(2): 281-311, 2014.

[26] H. C. Purchase, "Metrics for graph drawing aesthetics," *Journal of Visual Languages and Computing*, vol. 13, no. 5, pp. 501-516, 2002.

[27] H. C. Purchase, "Which Aesthetic has the Greatest Effect on Human Understanding?" *Proceedings of the Symposium on Graph Drawing (GD'97)*, Springer-Verlag, 248-261, 1997.

[28] M. Taylor and P. Rodgers, "Applying graphical design techniques to graph visualisation," *Proceedings of the International Conference on Information Visualisation (IV'05)*, IEEE Press, 2005, pp. 651-656.

[29] C. Ware, H. Purchase, L. Colpoys, and M. McGill, "Cognitive measurements of graph aesthetics," *Information Visualization*, vol. 1, no. 2, pp. 103-110, 2002.

[30] F. Zaidi, D. Archambault, and G. Melanon, "Evaluating the Quality of Clustering Algorithms Using Cluster Path Lengths," *In Proceedings of the industrial conference on Advances in data mining (ICDM'10)*, Springer-Verlag, 42-56, 2010.