Effective information visualisation: a study of graph drawing aesthetics and algorithms

H.C. Purchase*

The Department of Computer Science and Electrical Engineering, The University of Queensland, St. Lucia 4072, Qld, Australia

Received 21 May 1999; revised 28 February 2000; accepted 5 March 2000

Abstract

Information visualisation systems which generate diagrams representing discrete relational information must consider potential users if they are to be effective. Many algorithms which render an abstract graph structure as a diagram are valued for their conformance to aesthetic criteria (e.g. reducing the number of edge crossings, maximising symmetry), or for computational efficiency. They are not usually judged on their ability to produce diagrams that maximise human performance.

This paper presents the results of experiments investigating the relative worth (from an HCI point of view) of graph drawing aesthetics and algorithms using a single graph. The results indicate that while some individual aesthetics affect human performance, it is difficult to say that one algorithm is ‘better’ than another from a relational understanding point of view. Designers of automatic layout algorithms, and the systems which embody such algorithms, can benefit from this study and this human-centred approach, by adapting their methods to focus on user concerns, rather than computational ones. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Graph drawings; Graph layout aesthetics; Graph layout algorithms; User experimentation; Relational information visualisation

1. Introduction

An increasing number of visualisation systems display relational information in a graphical form; that is, as a diagram of nodes (often represented as circles or squares) connected by edges (represented as lines between the nodes). It is important that the effects of the manner in which this information is displayed are investigated: if displaying relational information graphically is to assist in its comprehension and use, we need some understanding of human performance using these graphical displays.

* Tel.: +61-7-3365-2697; fax: +61-7-3365-1999.
  E-mail address: hcp@csee.uq.edu.au (H.C. Purchase).

0953-5438/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0953-5438(00)00032-1
Automatic graph layout algorithms enable relational data to be displayed in a graphical form, producing graph drawings comprising nodes and edges. For example, Fig. 1 shows how the relational data in a table (where * indicates null) may be displayed in two different ways. The first graph drawing (b) is based on the spring model of graph layout (Eades, 1984), while the second (c) has been formed by fixing the nodes and edges to a grid (Tamassia, 1987).

In general, viewing the data in a graphical manner makes it easier to comprehend and use than if it were viewed in tabular form, (although this assumption is not always valid (Petre, 1995; Scaife and Rogers, 1995)), and such visualisation may reveal internal relationships and structures that were otherwise hidden (Blythe et al., 1995).

A wide variety of automatic graph layout algorithms have been produced over many years: these algorithms vary in the different approaches and visualisation priorities that they favour (Di Battista et al., 1994). However, these algorithms have tended to be valued with respect to their computational efficiency or elegance, rather than because of any improvement in the human understanding and use of the information (Brandenburg et al., 1995). While these algorithms are acknowledged as being useful for information visualisation (Kamada, 1989), few studies have considered them from a human understanding point of view.1

In studying the effects of visualising information graphically, it is important that the implementational work that is being done by designers of automatic graphical layout algorithms be taken into account. The algorithms and the systems that embody them need to be based on experimental evidence indicating how they may best satisfy the needs of users.

The project reported here has so far investigated the human reading of graph drawings with respect to two separate issues: firstly in investigating the individual aesthetics that typically underlie the design of graph layout algorithms (for example, reducing the number

---

1 A notable exception is the work of Dengler and Cowan (1998), which describes a human perception experiment considering subjects’ intuitive semantic interpretation of variously laid-out graphs.
of edge crossings, maximising the display of symmetric structures); and secondly in investigating the graph drawings produced by several existing graph layout algorithms.

Two separate experiments were performed to investigate human performance in reading drawings of relational graph structures: the same experimental methodology was used for both experiments. Briefly, the experiments entailed subjects answering questions about a number of different drawings of the same graph. The difference between the two experiments was the manner in which the graph drawings were produced. In experiment 1, where the investigation centred around specific aesthetics which are typically used as the basis for graph layout algorithms, the drawings were produced by hand, with careful varying of the aesthetics as the independent variables. In the second experiment, the drawings were produced by different graph layout algorithms. In both the experiments, measurements were taken on the number of errors made and the time taken to answer the questions.

The results show that while the effect of some of the individual aesthetics may be significant, when more than one aesthetic are combined in a layout algorithm, the relative effect of the combination of aesthetics is not as obvious.

This paper describes the nature of the on-line system used for the two experiments and the experimental methodology, presents and discusses the results, and compares the results of the two experiments. Experimental studies that may follow from this initial study are outlined.

2. The experiments

2.1. Definition

There are two ways in which human performance in the reading of graph drawings may be measured. A purely relational method measures the efficiency and accuracy with which people can read a graph structure and answer questions about it. Such graph-theoretic questions need to be generic and application-independent, and may include questions of the form “What is the shortest path from node A to node B?” A more application-specific method would rather consider a graph interpretation task: in this case it is more appropriate that the effectiveness of the graph drawing is measured within the context in which the application-specific graph is usually used. Thus, instead of eliciting answers to specific questions asked about the graph itself, it is more suitable to look at whether the graph has assisted the user in accomplishing a particular application task. Suitable questions for this approach would include (in the area of software engineering) “What object classes would be affected by changing the external interface to class X?”.

In this experiment, the relational reading of a graph drawing is considered, leaving the interpretive consideration of graph reading for a later study. The questions that were chosen to measure relational understandability are the ones that tend to be of interest to graph-theorists:

- How long is the shortest path between two given nodes?
- What is the minimum number of nodes that must be removed in order to disconnect two given nodes such that there is no path between them?
What is the minimum number of edges that must be removed in order to disconnect two given nodes such that there is no path between them?

2.2. Experimental methodology

The experiments were performed online using an experimental system, designed and implemented for experiments like these, and robust enough to withstand input from novices. Each student interacted with a unique experimental program.

First, the subjects were introduced to the relevant terminology \(\text{graph, node, edge, path, and path length}\), and had the three questions explained to them. A simple example graph drawing, with the three questions and their correct answers, was shown. Subjects were given an opportunity to ask any questions.

Next, the subjects answered the three questions on six “practise” graph drawings: they were not told that these graph drawings were not experimental.

The experimental graph drawings (ten in experiment 1, eight in experiment 2) were then each displayed three times, once for each question. The order of presentation of the drawings and the questions was random, as was the orientation of the drawings.

The questions themselves were randomised too: although the same three questions were asked of each drawing, the pair of nodes chosen for each question was randomly selected from a list of node-pairs. This ensured that any variability in the data could not be explained by the varying difficulty of the questions. The two relevant nodes for each question were highlighted in black on the screen, ensuring that response time did not include time taken to locate the nodes by labels. The subjects typed their answers to the questions.

The use of the online system enabled two dependent variables to be recorded: the time taken for the subject to answer the questions (the “response time”); as well as the correctness of the answer. This enabled analysis to be performed on two measures of understanding.

The experiment was therefore controlled for the questions and the graph, and the two dependent variables were the response time, and the number of errors made for each drawing. In experiment 1, the independent variable was the amount of aesthetic presence in each drawing, and in experiment 2, the independent variable was the algorithm used to create the drawings.

A within-subjects analysis method was used to reduce any variability that may have been attributable to differences between subjects (e.g. age, experience). Any learning effect was minimised by the large number of graph drawings used, the inclusion of practise graphs, and the random order of the graph drawings.

3. Experiment 1: investigating individual aesthetics

3.1. Graph drawing aesthetics

Designers of graph drawing algorithms tend to optimise certain aesthetics, and claim that by doing do, the resultant graph drawing helps the human reader to understand the
information embodied in the graph. This first experiment aimed to investigate the effect of these individual aesthetic criteria on human performance.

Five common graph-drawing aesthetics were chosen for experiment 1:

- **minimise bends** (the total number of bends in polyline edges should be minimised (Tamassia, 1987));
- **minimise edge crossings** (the number of edge crossings in the display should be minimised (Reingold and Tilford, 1981));
- **maximise minimum angles** (the minimum angle between edges extending from a node should be maximised (Coleman and Parker, 1996; Gutwenger and Mutzel, 1998);
- **orthogonality** (fix nodes and edges to an orthogonal grid (Tamassia, 1987; Papakostas and Tollis, 2000));
- **symmetry** (where possible, a symmetrical view of the graph should be displayed (Gansner and North, 1998));

In this experiment, the five primary hypotheses associated with the five different aesthetics under consideration were proved or disproved. In addition, Tukey’s WSD pairwise comparison procedure (Gottsdanker, 1978) was then used to determine if there were significant understandability priorities between the aesthetics.

3.2. The graph and graph drawings

The graph for experiment 1 was carefully designed so that the node-pairs could be identified which gave a suitable range of values for the three questions. It was important that the answers to the three questions that the subjects would be asked were not the same for each version of the graph. Thus, the node-pairs were defined that would give correct answers to the first question (the shortest path) of either 2, 3, 4 or 5, the node-pairs were defined that would give correct answers to the second question (the number of nodes to remove) of either 1 or 2, and node-pairs were defined that would give correct answers to the third question (the number of edges to remove) of either 1, 2 or 3. The graph has 16 nodes and 28 edges.

To manipulate the “amount” of any aesthetic in the creation of the experimental drawings, a means of measuring aesthetic presence was required. Metrics for all five aesthetics were defined (Purchase and Leonard, 1996). These are scaled to lie between 0 and 1, where 0 represents an amount of the aesthetic that it is assumed makes the drawing difficult to read (e.g. not much orthogonality), while 1 represents an amount of the aesthetic that it is assumed makes the drawing easy to read (e.g. not many crosses). The metric for symmetry takes into account global and local symmetries, weighting them by their area, and considers the effects of crosses and bends on perceptual symmetry.

Ten experimental graphs were created, two for each of the aesthetics (representing a strong or weak presence). The graph drawings are named after the relevant aesthetic:

- **b** (minimise bends);
- **c** (minimise edge crossings);
- **m** (maximise minimum angles);
The suffix + or − indicates the strength of the aesthetic, where + indicates an aesthetic value close to 1 (i.e. assumed easy to read), and − indicates an aesthetic value close to 0 (i.e. assumed difficult to read). Thus, the s+ drawing has a symmetry metric value closer to 1 than the s− drawing. Fig. 2 shows the ten graph drawings for experiment 1. Note that these drawings have been variously scaled for the purposes of effective presentation: in the

<table>
<thead>
<tr>
<th>Graph drawing</th>
<th>b+ b- c+ c- m+ m- o+ o- s+ s-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Errors</td>
<td>0.24 0.53 0.29 0.80 0.36 0.36 0.36 0.29 0.31</td>
</tr>
<tr>
<td>Time</td>
<td>67.18 81.40 66.39 139.78 76.55 68.17 71.37 76.71 55.58 67.74</td>
</tr>
</tbody>
</table>

Fig. 3. Experiment 1: the average number of errors and average response time for each graph drawing.
experiment they were displayed with all nodes being of equal size, within a screen area of approximately 16 cm × 16 cm.

In the drawings varying a particular aesthetic, it was important that the values of the other four aesthetics were kept constant, to ensure no confounding of variables. It is difficult, and in some cases impossible, to use the extremes of 0 or 1 as the constant value for the other four aesthetics: e.g. a metric value of 0 for the bend aesthetic would imply an infinite number of bends; a metric value of 1 for minimum angle aesthetic would mean that all the nodes in the drawing have the optimum angles between its edges (impossible for any cyclic graph). For this reason, a “neutral range” was defined for each aesthetic (based on perception), and for the drawings which varied one aesthetic, the values of the other four aesthetics were kept within these specified ranges.

Fifty-five second-year computer science students took part in experiment 1, for a reward of $10.

3.3. Results: testing the five individual hypotheses

The average number of errors and the average response time for the ten experimental graph drawings are shown in both tabular and chart form in Fig. 3.

To test the five primary hypotheses, one for each aesthetic, first the significance of the effects of the level of difficulty (the +/− dimension) needed to be confirmed. After this confirmation that the +/− dimension had indeed affected the error and response time data collected, each individual aesthetic was then tested for its contribution to this overall effect. This analysis was performed for both errors and response time.

3.3.1. Results

The 2 × 5 within-subject analysis of variance showed that:

- The main effect of the level of difficulty (the +/− dimension) was significant for both errors ($F(1, 54) = 14.89, \alpha = 0.05$) and response time ($F(1, 54) = 40.67, \alpha = 0.05$).\(^2\)
- The simple effect of the bends metric was significant for errors ($F(1, 54) = 14.49, \alpha = 0.01$) but only approaches significance for response time ($F(1, 54) = 5.84, \alpha = 0.01$).
- The simple effect of the crosses metric was significant for both errors ($F(1, 54) = 24.25, \alpha = 0.01$), and response time ($F(1, 54) = 87.98, \alpha = 0.01$).
- The simple effect of the minimum angle metric was not significant for both errors ($F(1, 54) = 0.09, \text{NS}$) and response time ($F(1, 54) = 3.05, \text{NS}$).
- The simple effect of the orthogonality metric was not significant for both errors ($F(1, 54) = 0.00, \text{NS}$) and response time ($F(1, 54) = 1.44, \text{NS}$).
- The simple effect of the symmetry metric was not significant for errors ($F(1, 54) = 0.09, \text{NS}$), but was significant for response time ($F(1, 54) = 7.57, \alpha = 0.01$).

---

\(^2\) The statistical analysis used here is a standard ANOVA analysis, based on the critical values of the F distribution: \(\alpha\) is the level of significance, and results that are not significant are indicated by NS.
3.4. Results: prioritising the aesthetics

To determine the relative effect of the aesthetics, and attempt to place a priority ordering on their importance, both the set of + drawings and the set of - drawings needed to be tested for the overall effect of the aesthetics. Those sets of drawings for which the effect of the aesthetics were significant were then subjected to a Tukey’s pairwise comparison to determine which aesthetics differed significantly from one another.

3.4.1. Results

The $2 \times 5$ within-subject analysis of variance showed that:

- The main effect of the aesthetics dimension was significant for both errors ($F(4, 216) = 4.16, \alpha = 0.05$) and response time ($F(4, 216) = 28.49, \alpha = 0.05$).

The - drawings

- The simple effects of the five different aesthetics were significant for the error data ($F(4, 216) = 9.60, \alpha = 0.025$). The Tukey’s WSD pairwise comparisons procedure showed that, for the error data, crosses were significantly different from all other aesthetics: for bends ($F(5, 216) = 9.11, \alpha = 0.05$), minimum angle ($F(5, 216) = 22.05, \alpha = 0.05$), orthogonality ($F(5, 216) = 24.20, \alpha = 0.05$), symmetry ($F(5, 216) = 30.01, \alpha = 0.05$). There were no other significant pairwise differences.

- The simple effects of the five different aesthetics were significant for response time ($F(4, 216) = 50.89, \alpha = 0.025$). The Tukey’s WSD pairwise comparisons procedure showed that, for the response time data, crosses were significantly different from all other aesthetics: for bends ($F(5, 216) = 95.09, \alpha = 0.05$), minimum angle ($F(5, 216) = 143.07, \alpha = 0.05$), orthogonality ($F(5, 216) = 110.98, \alpha = 0.05$), symmetry ($F(5, 216) = 144.79, \alpha = 0.05$). There were no other significant pairwise differences.

The + drawings:

- The simple effects of the five different aesthetics were not significant for the error data ($F(4, 216) = 1.02$, NS).
- The simple effects of the five different aesthetics were significant for the response time data ($F(4, 216) = 4.68, \alpha = 0.025$).

The Tukey’s WSD pairwise comparisons procedure showed that, for the response time data, symmetry was significantly different from the minimum angle ($F(5, 216) = 17.14, \alpha = 0.05$), and orthogonality ($F(5, 216) = 9.72, \alpha = 0.05$).

3.5. Analysis

3.5.1. Errors

The error chart in Fig. 3 shows that the average number of errors for the - versions of the drawings (i.e. the “difficult” drawings) was greater than the average number of errors
for the + versions, in all cases except orthogonality when the averages were the same. The statistical analysis shows that the level of difficulty of the drawings (as measured by errors) was only significant for \textit{bends} and \textit{crosses}.

The Tukey’s pairwise comparison for the error data showed that the average number of errors for the \textit{crosses} \(c\) drawing was significantly greater than the errors in all other \(c\) versions of the aesthetics, and that there were no significant pairwise orderings for the + drawings.

3.5.2. Response time

The response time chart in Fig. 3 shows that \(c\) versions of the bends, crosses, orthogonality and symmetry drawings all took longer than the + versions. The statistical analysis shows that the level of difficulty of the drawings (as measured by the response time) was only significant for \textit{crosses} and \textit{symmetry}. The unexpected reversal of average response time for the two minimum angle drawings is not significant, and can therefore be attributed to chance.

The Tukey’s pairwise comparison for the response time data showed that the \textit{crosses} \(c\) drawing took significantly more time than all the other \(c\) versions of the aesthetics. In addition, the \textit{symmetry} \(s\) drawing took significantly less time than the “easier” \textit{minimum angle} \(m\) and \textit{orthogonality} \(o\) drawings.

3.6. Discussion

There is no doubt that the evidence is overwhelmingly in favour of the reduction of crosses as being the aesthetic that affects human relational graph reading the most, as suggested by the results of the two Tukey pairwise comparison tests performed on the \(c\) drawings. The effect of crosses was not noticeable, however, in the + drawings, implying that crosses are only more problematic than the other aesthetics when there are a large number of them.

The results of the other aesthetics are less certain: the bends and symmetry hypotheses were supported either for response time or errors, but not both. Orthogonality and minimum angle had no effect on the subjects’ relational graph reading at all. The Tukey test for the response time data for the + drawings showed that symmetry took significantly less time than the minimum angle and orthogonality, suggesting that symmetry only has a more positive effect than the other aesthetics when it is at a maximum value.

An unusual result was that for the easy drawings, the different aesthetics had no significant effect on the number of errors (even though there was an effect on response time). This suggests that the subjects tended to give correct answers on all aesthetics if the drawings were easy, but they used all the time necessary, requiring different amounts of time for the different aesthetics. On the other hand, for the difficult drawings, subjects still took the amount of time necessary (which differed for the different aesthetics), but the difficulty of the drawings meant that the number of errors was also differentially affected for different aesthetics.
4. Experiment 2: investigating layout algorithms

4.1. Graph drawing algorithms

Many algorithms have been defined for the process of depicting a relational graph structure as a diagram of nodes and edges; the algorithms investigated in this experiment are all implemented in the GraphEd system (version 4.1.7-beta (Himsolt, 1990)). Eight algorithms were used, resulting in a set of eight experimental graph drawings. Much of the following information about the algorithms implemented in GraphEd has been provided by Himsolt (1997), and has been documented elsewhere (Brandenburg et al., 1995; Himsolt, 1995). The names given to the algorithms are, as much as is possible, the same as those used in the papers by Brandenburg et al. (1995) and Himsolt (1995).

FD-FR: This force directed algorithm by Fruchterman and Reingold (1991) is based on an original idea by Eades (1984). It produces a drawing that attempts to display symmetric structures, with few edge crossings.

FD-K: This is another force directed algorithm, by Kamada and Kawai (1989) based on Eades. The drawing produced by this algorithm is similar to that produced by FD-FR, with a more even distribution of nodes.

POGB: This is an implementation of an algorithm for planar orthogonal grid drawing with bends minimisation, designed by Tamassia. GraphEd provides two implementations of this algorithm, with differences in the manner in which the co-ordinates of the nodes are assigned. Himsolt (1997) describes the differences in the two algorithms as being related to the starting conditions: the first algorithm (POGBa) generates the planar embedding from the drawing; while the second one (POGBb) uses a planarity test. He notes that the result is not deterministic.

PG: This planar grid drawing algorithm is based on the one designed by Woods (1982). Unlike the product of the POGB algorithm, the PG drawing has many sloped edges.

PGS: The planar grid drawing with straight line edges is based on the algorithm by de Fraysseix et al. (1990) with improvements as suggested by Chrobak and Payne (1990).

SEIS: This algorithm is documented in a thesis by Seisenberger (1991). The first step is always the PGS algorithm. This is followed by repeated compression of the drawing in both the x and y directions (Himsolt, 1997). This algorithm is not deterministic.

Tu: The algorithm used here is the incremental algorithm designed by Tunkelang (1992). The product is a drawing that has a similar even distribution of nodes to that expected from a force-directed algorithm.

4.2. The graph and graph drawings

The graph for this experiment was based on the graph used in the first experiment, and also needed to be carefully designed so that the node-pairs could be identified which gave a suitable range of values for the three questions. Thus, a set of node-pairs was defined that would give correct answers to the first question (the shortest path) of either 2, 3, or 4; a set of node-pairs was defined that would give correct answers to the second question (the
number of nodes to remove) of either 2 or 3; and a set of node-pairs was defined that would give correct answers to the third question (the number of edges to remove) of either 2 or 3.

In addition, the graph structure was limited by the algorithms that were to be applied to it: it needed to be undirected, with maximum degree 4. The graph has 17 nodes and 29 edges.

Fig. 4 shows the eight drawings of this graph produced by the eight algorithms. Note
that these drawings have been variously scaled for the purposes of effective presentation: in the experiment they were displayed with all nodes being of equal size.

Fifty-five third-year computer science students took part in experiment 2, for a reward of $10.

4.3. Results: prioritising the algorithms

The average number of errors and the average response time for the eight experimental graph drawings are shown in both tabular and chart form in Fig. 5.

The within-subject analysis of variance showed that the main effect of the algorithms was significant for errors \( F(7, 378) = 2.856, \alpha = 0.05 \) but not significant for response time \( F(7, 378) = 1.464, \text{NS} \).

To determine the relative effect of the algorithms on the average number of errors, and attempt to place a priority ordering on their difficulty, the set of drawings were subject to a Tukey’s pairwise comparison to determine which algorithms differed significantly from one another in terms of human performance.

The Tukey’s WSD pairwise comparisons procedure showed that, for the error data, the SEIS drawing produced significantly more errors than the FD-FR \( F(8, 378) = 10.12, \alpha = 0.05 \), Tu \( F(8, 378) = 11.486, \alpha = 0.05 \) and FD-FK \( F(8, 378) = 14.63, \alpha = 0.05 \) drawings. There were no other significant pairwise differences.

4.4. Analysis

The response time chart in Fig. 5 shows little variability between the performance of the subjects on the eight graph drawings; unsurprisingly, statistical analysis of this response time data revealed no significant effect. However, the error chart in Fig. 5 shows variability between the performance of the subjects on the eight graph drawings: the within-subject analysis of variance showed that this effect was significant, with the SEIS drawing produced significantly more errors than the FD-FR, Tu and FD-FK drawings.
4.5. Discussion

The average response times for the products of the eight algorithms were not significantly different, implying that the subjects did not perceive the drawings of this graph to be of varying difficulty. But the average number of errors for the drawings were significant, indicating that, despite this perception, there was indeed a difference in the difficulty of the drawings. Further analysis revealed where this variance lay: the SEIS drawing produced significantly more errors than the two force-directed drawings and Tukelang’s incremental algorithm.

From an aesthetic presence point of view, it is not clear why the SEIS algorithm produced such poor results. Fig. 6 shows the relative aesthetic values (as calculated by the defined aesthetic metrics (Purchase and Leonard, 1996)) for the SEIS graph drawing, and those drawings from which its performance was significantly different (FD-FR, Tu and FD-FK). The only noticeable difference is in the amount of symmetry, where the SEIS drawing is more symmetrical than the others. Intuitively, this difference ought to imply better performance from the SEIS drawing, not worse.

This observation may indicate that the symmetry metric definition itself may need to be reassessed. At present, the metric identifies symmetric sub-graphs and weights them according to their area, as a proportion of the total area occupied by the drawing. In an attempt to measure perceptual symmetry (as opposed to pure geometric symmetry), a tolerance factor is used when determining whether two pairs of co-ordinates are symmetric around an axis. It may be that the metric needs to be more tolerant towards small deviations than at present in order to provide a better indication of perceptual symmetry.

Another possible explanation for this phenomenon is the presence of other aesthetics that have not been included (and therefore not measured or compared) in the experiments reported here. Examples include node proximity and uniformity of edge lengths (Colemen and Parker, 1996).

These results therefore indicate that, apart from the SEIS algorithm, all the algorithms produce drawings that are of comparable difficulty. While the chart of the average number of errors in Fig. 5 seems to indicate that the force-directed algorithms have better performance than the grid-based ones, there is no statistical evidence to support this.

This result indicates that it is difficult to say that one algorithm is ‘better’ than another from a relational understanding point of view.

5. Experimental comparison

The results of experiment 2 indicate that, despite the differing aesthetic bases for the algorithms, they produce comparable human performance results. This is a surprising result, particularly given the results of experiment 1, when there was overwhelming evidence to support the reduction of the number of edge crossings over all other aesthetics, partial support for minimising edge bends and maximising symmetry, and no support for orthogonality and maximising the minimum angle.

The outcome of experiment 2 shows no difference between the force-directed algorithms FD-FR, FD-K and Tu (which tend to maximise symmetry at the expense of...
crosses, with no bends), and the grid-based algorithms **POGBa** and **POGBb** (which maximise orthogonality, have no crosses, and minimise the number of bends). Given the strong support for reducing the number of crosses in the prior experiment (with only partial support for symmetry) the expectation may be that the grid-based algorithms would produce better performance. This was not the case.

There are three possible explanations for this outcome. One factor may be the difference in the nature of the students: third year students (experiment 2) may have had more exposure to graph drawings than second year students (experiment 1); the latter may therefore have been more sensitive to aesthetic variations.

Second, may be that it is not possible to separate the effect of individual aesthetics, as was attempted experiment 1, and that the interactions between the aesthetics are also important. For example, the combination of both reducing the number of bends and maximising orthogonality may produce a more significant effect than merely reducing the number of edge crossings. Further experimentation will need to be performed in order to investigate the interaction effects of the aesthetics.

Third, the result may be explained intuitively by considering that, in experiment 2, the force-directed algorithms produced drawings which included only very few crossings. It may be that the number of edge crossings only has a significant effect on understandability when there are many of them: perhaps a “critical mass” of crossings needs to be reached before their number creates a serious understandability problem. This issue of critical mass may, of course, also be relevant to the other aesthetics: perhaps it is the case that orthogonality only has a positive effect if the orthogonality measure of the drawing is greater than a certain amount.

6. Conclusions

The aim of experiment 1 was to indicate to the designers of graph drawing algorithms the most effective aesthetics to use from the point of view of human reading of relational information. The results show that there is strong evidence to support minimising crossings, and weaker evidence for minimising the number of bends and maximising perceptual symmetry. Maximising the orthogonal structure of the drawing, and maximising the minimum angles between edges leaving a node, appear to have little effect.

The aim of experiment 2 was to investigate the graph drawing algorithms, to indicate to the designers of systems which algorithms are best from a human readability point of view. The experiment has shown that for a simple graph, only one of the eight algorithms gives a poor performance when compared with the others.

These results indicate that despite the range of aesthetic bases for these layout algorithms, it is difficult to say that one algorithm is ‘better’ than another from a relational understanding point of view.

This conclusion must, of course, be interpreted within the limitations of a formal, controlled experiment. It is common knowledge that all formal experiments are limited by their parameters (Gottsdanker, 1978): this is an inevitable consequence of the controlled experimental method. In particular, this preliminary study has considered the
performance effect on two graphs, with three specific, relational questions, and the generalisability of these results are therefore limited to within these clearly defined parameters. These experiments therefore open a wide, new field of empirical investigation, and there are many avenues for further studies which may either corroborate or invalidate the results presented here. In particular:

- Would the same aesthetic and algorithmic comparative results be forthcoming on a set of different graphs (for example, graphs with different structures and sizes)?
- Would the same aesthetic and algorithmic comparative results be forthcoming with different relational performance measurements (for example, if subjects were asked to trace paths between nodes, determine the maximum degree of the graph, or identify sub-graphs)?
- Would the same aesthetic and algorithmic comparative results be forthcoming if an interpretive approach were taken (for example, if subjects were asked to read the graph in the context of application domains like object-oriented design diagrams or data-flow diagrams)?

By presenting this methodology and these results, it is hoped that this new area of investigation may be adopted more widely, so that the answers to these important issues may be resolved. Studies like the one reported here can guide the production of information visualisation systems to ensure that they serve users effectively.

Acknowledgements

I am grateful to Robert Cohen (who assisted with the experimental definition), to Michael Himsolt (for the descriptions of the algorithms as implemented in GraphEd), to Murray James (who developed the online system), to David Leonard (who helped define, and implemented, the aesthetic metrics), to Daniel Naumann (who assisted with the experiments), and to Julie McCreddon (who guided the statistical analysis).

References


