Evaluating Graph Drawing Aesthetics: defining and exploring a new empirical research area

Helen C. Purchase
Computing Science Department
University of Glasgow
Glasgow G12 8QQ
Scotland

tel: +44 141 330 4484
fax: +44 330 4913
email: hcp@dcs.gla.ac.uk
Evaluating Graph Drawing Aesthetics:

defining and exploring a new empirical research area

ABSTRACT

This chapter describes a long-term project to investigate the validity of the design principles on which many automatic graph layout algorithms are based, not from the perspective of computational efficiency, but from the perspective of human comprehension. It describes a framework for experimentation in this area, the overall methodology used throughout, as well as the details of the experiments themselves. It shows the development of the empirical ideas and methods as the project matured, and provides reflections on each experiment, demonstrating the difficulty of initiating a new experimental research area. The paper suggests how the current results should best be interpreted, as well as ideas for future work in this area.

INTRODUCTION

This chapter describes a long-term project to investigate the validity of the design principles on which many automatic graph layout algorithms are based, not from the perspective of computational efficiency, but from the perspective of human comprehension.

There are two main objectives of the chapter:
• to summarise empirical work the author has done on the effectiveness of 2D graph drawing aesthetics;
• to describe the process of initiating a new experimental research area, and of developing a framework for empirical studies within the area, with specific reference to the experimental methodology and statistical analysis issues involved.

BACKGROUND

A graph is a set of nodes (representing objects) and edges (representing relationships between the objects). Graphs may be represented in diagonal matrices, with each node being associated with both a row and a column, and each edge being represented by a positive value in the cell that links two nodes. However, graphs are more typically represented as node-edge diagrams (called graph drawings). Figure 1 represents three different renderings of the same graph structure.

![Figure 1: Three representations of the same graph](image)

The process of creating a graph drawing from an underlying graph structure is known as automatic graph layout. Many graph layout algorithms exist (Battista et al., 1994), and for many years researchers have been devising increasingly efficient and elegant algorithms for the production of graph drawings. At the annual Graph Drawing symposium, researchers present
their newest layout algorithms, or their variations on existing algorithms: this forum typically comprises approximately 30 papers, of which only a few do not present a new algorithm.

These algorithms are typically valued for their computational efficiency, and the extent to which they conform to one or more standard layout design principles (called \textit{aesthetic criteria}). Examples of such criteria include minimising the number of edge crosses, maximising the display of symmetric structures, and maximising orthogonality (the property of fixing edges and nodes to an invisible underlying unit grid). Designers of graph layout algorithms claim that by optimising these measurable aesthetics, and producing ‘nice’ graph drawings, the graph information is easier to read. Prior to 1995, these claims had not been empirically tested with respect to human studies.

The author initiated a new empirical research area in 1995: that of attempting to validate these graph drawing aesthetics, not with respect to computational efficiency, but solely with respect to human understanding. The project has so far investigated the following areas with human empirical studies:

- effectiveness of several aesthetic criteria in the performance of graph theoretic tasks;
- effectiveness of several layout algorithms in the performance of graph theoretic tasks;
- human preference for different aesthetic criteria in the layout of software engineering diagrams;
- effectiveness of several aesthetic criteria in the performance of software engineering tasks;
- effectiveness of several layout algorithms in the performance of social network analysis tasks.
In addition, to support the empirical studies, the project produced:

- definitions of computational metrics for measuring aesthetic presence, independent of the structure of the underlying graph structure.

The aim of this chapter is two-fold:

- To summarise and integrate the results of some of these empirical studies, in an attempt to give a ‘big-picture’ overview of this new research area. Six experiments are described.
- To demonstrate the process of building a new empirical research area, including the production of an empirical framework, discussion of the mistakes made, and a description of the experimental methodological and statistical analysis decisions required.

While the specific details of some of the experiments presented in this paper have been published elsewhere, this chapter also includes descriptions of the unpublished pilot studies that informed the experimental design, and reflections on the ideas and methods used as the project matured.

**EXPERIMENTAL ISSUES**

**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. The aesthetic criteria considered in this project were all extracted from the research
literature on automatic graph layout algorithms. The set of aesthetics addressed in each experiment differs: at the start of the project, only a few, common aesthetics were considered; the later experiments include a wider range.

The graph-drawing aesthetics considered in the experiments described in this paper are:

- *(c) edge crosses*: the number of edge crosses should be minimized (Reingold & Tilford, 1981)
- *(m) minimum angles*: the minimum angle between edges extending from a node should be maximised (Coleman & Stott Parker, 1996; Gutwenger & Mutzel, 1998)
- *(b) bends*: the total number of bends in polyline edges should be minimised (Tamassia, 1987)
- *(ndis) node distribution*: nodes should be distributed evenly within a bounding box (Coleman & Stott Parker, 1996)
- *(el) edge lengths*: edge lengths not be too short, nor too long (Coleman & Stott Parker, 1996)
- *(ev) edge variation*: edge lengths should be of similar length (Coleman & Stott Parker, 1996)
- *(f) flow*: directed edges should, as much as possible, point in the same direction (Waddle, 2000)
- *(orth) orthogonality*: nodes and edges should be fixed to an orthogonal grid (Tamassia, 1987; Papakostas & Tollis, 2000)
- *(sym) symmetry*: where possible, a symmetrical view of the graph should be displayed (Eades, 1984; Gansner & North, 1998)
Metrics

The measurement of these aesthetic criteria within a graph drawing is often done informally, and may differ between different algorithms. There is no standard, objective way for analysing a graph drawing with respect to the presence of different aesthetics. Trivial counting methods may be used for the simper aesthetics (for example, counting the number of intersecting edges, or the number of edges which point in the same direction), but continuous measures are necessary, so that analysing a drawing with respect to an aesthetic is not merely a binary decision. Thus, it is desirable that a drawing is not considered ‘orthogonal’ or ‘not orthogonal’, but is rather described according to the extent to which the drawing conforms to the orthogonality aesthetic (that is, the drawing may be considered to have 65% presence of orthogonality).

In defining new continuous metrics for aesthetic criteria, two particular issues arose:

- The metrics need to be scaled, so that they all lie within the same range. This enables aesthetic presence comparisons between drawings of graphs of different size and structure. For example, a drawing of a graph with 150 edges of which 5 intersect, would need a lower aesthetic presence value of the crossing aesthetic than a drawing of a graph with 10 edges of which 7 intersect.

- Many of these aesthetics are difficult to measure by trivial computational methods. For example, satisfying the orthogonality aesthetic needs to be more than merely putting the nodes and edges onto an underlying grid: the resolution of the grid needs to be optimal according to the number of nodes, and the angular deviation of any edges that do not lie on the grid needs to be taken into account.
Early in the project, unscaled metrics for crosses, bends and symmetry were defined for use in experiments 1 and 2. (Purchase, Bhanji et al., 1995). These definitions were altered so that scaled values were produced, and the second version of the symmetry metric was extended to take more account of symmetric sub-graphs (Purchase & Leonard, 1996). Together with a further three metrics (orthogonal, min angle, flow), these adapted formulae were used in experiments 3, 4 and 5a (Purchase, 2002).

**EXPERIMENTAL FRAMEWORK**

It became clear during the early experiments that there were several different approaches that could be taken to experimentation in this area. As this was a new research area, there was no existing literature to which the various experiments could be related, so a new evaluation framework was defined as part of this project. This framework was based on three dimensions relevant to determining the experimental form for an investigation of graph drawing algorithms.

- *Usability measurement:* The usability of a graph drawing can be measured in many different ways: two common methods are *performance* (subjects perform better on a given task using one drawing than when using another) and *preference* (subjects express personal preference for one drawing over another).
- *The nature of the graph:* There are two types of graphs that can be used in the experiments: abstract (the information in the graph has no reference to the real world, referred to as a *syntactic* graph) and domain-specific (the graph represents a domain, for example, data-flow or transport networks, referred to as a *semantic* graph).
• **The effect being investigated:** There are two ways in which graph drawing layout can be investigated: by considering the usability effect of the individual *aesthetics*, or by looking at the usability of complete *algorithms* (which produce drawings conforming to different aesthetics to varying degrees).

This framework allows many different avenues of empirical study to be followed. While an experiment on algorithms has been conducted (Purchase, 1998) and preference data has been collected (Purchase, Carrington & Allder, 2002), this paper concentrates on performance, aesthetics experiments, both syntactic and semantic.

The form of the experiments reported in this paper are shown in Table 1.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Usability measurement</th>
<th>The nature of the graph</th>
<th>The effect being investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>performance</td>
<td>syntactic</td>
<td>3 aesthetics</td>
</tr>
<tr>
<td>2</td>
<td>performance</td>
<td>syntactic</td>
<td>3 aesthetics</td>
</tr>
<tr>
<td>3</td>
<td>performance</td>
<td>syntactic</td>
<td>5 aesthetics</td>
</tr>
<tr>
<td>4</td>
<td>performance</td>
<td>semantic (class diagrams)</td>
<td>4 aesthetics</td>
</tr>
<tr>
<td>5a</td>
<td>performance</td>
<td>semantic (UML class diagrams)</td>
<td>5 aesthetics</td>
</tr>
<tr>
<td>5b</td>
<td>performance</td>
<td>semantic (UML class diagrams)</td>
<td>7 aesthetics (perceptual)</td>
</tr>
<tr>
<td>6</td>
<td>performance</td>
<td>semantic (UML collaboration diagrams)</td>
<td>6 aesthetics (perceptual )</td>
</tr>
</tbody>
</table>

Table 1: The form of the different experiments, according to the experimental dimensions.

**Experimental method**

While the methodology for each of the experiments described below was specifically designed for the type of evaluation required, the underlying method is the same. A set of diagrams of the same
graph was produced according to different aesthetic presences (for example, one graph drawing would have no crossing edges, while another would have many). Experimental participants would be asked to perform a task using the drawings. The data collected was analysed to determine whether there was a significant difference in the performance of participants between different diagrams of contrasting aesthetic presence.

Most of the experiments were conducted using a within-subject methodology, with each subject performing a task using all of the experimental diagrams. One of the experiments uses a between-subject method (experiment 4), when participants were divided into eight groups, with each participant group being given the same one drawing as the other people in that same group.

STRUCTURE OF THIS PAPER

This first part of this paper will summarise all the experiments performed in this project so far, firstly describing the syntactic experiments, followed by the semantic ones. The order of the presentation of the experiments is roughly chronological (with the exception of experiment 4, which was performed between experiments 2 and 3). Presenting the experiments chronologically demonstrates the development of this new research area, shows how each experiment built on the results (or problems) of the prior one, and reveals the lessons learnt along the way.

The second part of this paper presents a general overview of the experimental results, reflections on the experimental processes and designs, and suggestions as to future avenues for research in this area.

THE SYNTACTIC EXPERIMENTS
Pilot experiment 1 (Purchase, Bhanji et al., 1995)

This was the first experiment that was performed, and it was intended to be only a pilot study. The aim of the experiment was to explore the relevant issues relating to aesthetic variations, experimental methodology, and graph based tasks. We hoped to gain a better understanding of what would be required in our later, more formal experiments.

The experimental drawings were defined by the following features:

- The aesthetics considered were: bends, crosses, symmetry.
- The variations for each of these aesthetics were: few, many.
- The graphs used were: dense (20 nodes, 32 edges), sparse (20 nodes, 21 edges).

There were therefore 12 drawings in total (for example, sparse-bends-few; dense-crosses-many). These drawings were drawn by hand, so that they conformed with the required aesthetic variations. These variations were measured by hand: the number of edge crosses were counted, with a three-way intersection counting as three crosses. The range of edge crosses in the drawings was 0-42. The number of edge bends were counted, with only bends with angles less than 150° being included. The range of edge bends in the drawings was 0-30. We defined a symmetry metric which identified symmetric sub-graphs around multiple axes in the drawing, and counted the edges mirrored around each axis. The range of the symmetry metric in the drawings was 0-51. Figure 2 shows the dense-crosses-many drawing.

Three different questions were asked of each of the 12 graph drawings. For example:
1. How long is the shortest path between node J and node G?

2. What is the minimum number of nodes you would need to remove in order to disconnect nodes F and B such that there is no path between them?

3. What is the minimum number of edges you would need to remove in order to disconnect nodes E and A such that there is no path between them?

The graphs were designed so that for each of the three question types, pairs of nodes could be identified that would give a range of answers. As each question type would be asked for each drawing, we needed to be sure that the answer to each question type was not always the same. The answer differed according to which node pair was used in the statement of the question. For the first question, four node-pairs could give the answers 2,3,4,5; for the second question, node pairs were identified that gave answers of 0,1,2, and for the third question, the possible answers were 0,1,2,3. The graphs included an independent smaller sub-graph of four nodes, so that the answer to questions 2 and 3 could be 0. Care was taken to ensure that in defining the node pairs corresponding to the three answers to question 3 (the minimum number of edges to be removed to disconnect two nodes), the answers to the questions were not easily determined by the degree of either of the nodes.

This experiment was paper-based: the subjects were given a booklet of the following structure:

- instructions;
- six pages of two graphs drawn three times each, with all three questions asked of all six drawings; these drawings were for practice only;
- a small word puzzle that served as a filler task;
• 12 pages showing the 12 experimental drawings, with all three questions asked of all twelve drawings: these drawings were in a different random order for each subject.

The data collected was the correctness of the subjects’ answers, and the length of time it took for them to complete the questions for each experimental diagram.

As this was only a pilot test, only five subjects took part in the experiment, and the data was not analysed statistically. Five different experimental booklets were produced, each with a different random order for each subject. The questions themselves were randomised too: although the same three question types 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 varying difficulty of the questions.

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 inclusion of practice graphs, and the random order of the graph drawings.

Most of the answers that the subjects gave to the three questions were correct. Although the accuracy and time data was not analysed, some qualitative data was collected from the subjects in the form of general comments: these revealed that the crosses-many drawings were considered more difficult than the bends-many drawings. There were no comments about symmetry.

The lessons we learned from this pilot study was how important it is for the experimental instructions to be very clear and unambiguous, and for examples of the required tasks (with their
correct answers) to be provided at the start of the experiment. We also discovered that timing by observation is very difficult and inaccurate. Most importantly, we learned about the importance of controlling confounding factors. For example, we realised how important it was that if we wished to compare subjects’ performance on the *sparse-bends-few* drawing with their performance on the *sparse-bends-many* drawing, then it is important that both these diagrams had the same number of crosses, and the same symmetry metric measurement.

The ‘graphs’ used in this experiment were really two graphs – we included an independent smaller sub-graph of four nodes (B,H,I and J in Figure 2), so that the answer to questions 2 and 3 could be 0. The main problem with this was that if one of the nodes referred to in node in question 2 was in this smaller independent graph, and it was also referred to in question 3, the answer was obvious. Many subjects also commented that they found his sub-graph very confusing, and none of the graphs used in the further experiments included an independent sub-graph.
Experiment 2 (Purchase, Cohen & James, 1997)

This experiment built on the experience of the pilot study: it considered the same three aesthetics and used adapted versions of the drawings produced for the pilot study. However, an additional aesthetic variation was introduced, and interim some measurement. Thus, for each of the three aesthetics, for the sparse graph, three drawings were produced (for example: sparse-bends-few, sparse-bends-some, sparse-bends-many). There were therefore 18 experimental graph drawings. The same methods for measuring the aesthetic variations were applied, the drawings were controlled for confounding factors (all other aesthetics being kept at zero), and the ratio between few/some/many was the same for all three aesthetics. Figure 3 shows the sparse-symmetry-some drawing, and indicates the axes around which local symmetry was measured. The same three graph-theoretic questions as were used in the pilot experiment were used as the experimental task.

As we were running the experiment formally for the first time, we needed to clearly define our hypotheses:
- Increasing the number of edge bends in a graph drawing decreases the understandability of the graph.
- Increasing the number of edge crosses in a graph drawing decreases the understandability of the graph.
- Increasing the local symmetry displayed in a graph drawing increases the understandability of the graph.

The methodology was similar to that used in the pilot experiments, although this time we wished to fix the time that the subjects spent answering the questions for each diagram, and collect accuracy data only. We used trials of this method to determine an appropriate time limit: as this time limit was different depending on whether the sparse graph (30 seconds) or the dense graph (45 seconds) was used, we separated the experimental diagrams and ran two experiments: one for the sparse graph and one for the dense graph (for which there were nine experimental drawings each).

For each of the two experiments, the subjects were given a booklet of the following structure:

- instructions, including definitions and worked examples;
- six practice drawings, with all three questions asked of each;
- a small word puzzle that served as a filler task;
- nine pages showing the nine experimental drawings, with all three questions asked of each.

There were 49 subjects for the dense graph, 35 for the sparse graph. Like before, the order of the experimental drawings was randomised for each subject. There were the same constraints on the nodes selected for each question as in the pilot study, except that as the independent sub graph had been removed, no question had 0 for its correct answer.
Like the pilot test, a within-subjects analysis method was used. The control variables were the graphs themselves and the time allowed for each drawing. The independent variables were the number of bends, the number of crosses and the value of the symmetry metric. The dependent variable was the accuracy data.

We collected accuracy data from both experiments. As the few/some/many variations measurements were ordinal and not continuous (i.e., the numerical difference in aesthetic variation between few and some is not necessary the same as the difference between some and many), we could not use parametric statistical methods. Instead, we ranked each student’s accuracy on each experimental drawing on a 1 to 3 scale, giving the drawing that they performed best on a score of 1, and the drawing that they performed least well on a score of 3 (with tied values recorded where necessary).

For our statistical analysis, we used the Friedman two-way analysis of variance by ranks method, which is used when a number of matched ordinal samples are taken from the same population. The aesthetic variations were ordinal, and the samples for each metric were matched as all values of each metric were tested on every subject. The Friedman test was used to determine the $\chi^2$-squared value for each aesthetic, for each of the two graphs, and from that, the probability that the ranked data was produced by chance. For those aesthetics for which the probability of the ranked accuracy data having been produced by chance was less than 0.05, observation of the average accuracy for the relevant drawings could tell us whether the direction of the accuracy difference followed that predicted by our hypotheses. Figures 4 and 5 show the charts representing the average accuracy and response times.
For both the sparse and dense graphs, our bends and crosses hypotheses were confirmed. There was no significant result for symmetry. There was very little variation in the average accuracy between the few/some/many symmetric variations, as most students got very few errors on the symmetric diagrams. This meant that the time limit given for these symmetry diagrams was too long: the students had sufficient time to answer the questions correctly (this is known as a ‘ceiling effect’).

One of the main lessons learned from this experiment was regarding controlling opposing potential confounds. In controlling the diagrams for the experiment, we avoided potential interaction between the aesthetics by having no variation of two aesthetics while varying the third. For example, all the graph drawings that varied the number of bends had no crosses and a zero value for symmetry. There is a potential conflict here between the nature of our three hypotheses: with bends and crosses, the hypothesis was that increasing their number would result in increased errors. However, the symmetry hypothesis was that decreasing the amount of symmetry would increase the errors. By keeping the number of crosses and bends at zero for the symmetrical graphs drawings, the drawings were being made simpler. However, keeping the symmetry metric at zero made the bends and crosses drawings more complex, it is therefore no surprise that a ceiling effect was observed with the reading of the symmetry drawings, when the same time period was being used for reading all the graphs of the three aesthetics.

Another important problem was revealed in this experiment. One of the subjects commented that much of the time was spent locating the two nodes whose labels corresponded with those stated in the question. This was a serious experimental design choice that should not have been followed when time was an important factor in the experiment. In subsequent experiments, no nodes were labeled, and the nodes relevant to the question were highlighted in black.
Figure 3: The *sparse-symmetry-some* drawing used for experiment 2, showing the 6 axes of symmetry used for calculating the symmetry metric.
Figure 4: The accuracy results for the dense graph in experiment 2.

Figure 5: The accuracy results for the sparse graph in experiment 2.
Other issues that arose out of this experiment was the difficulty of manual timing when a paper-based experiment is used, and we also questioned the effectiveness of the current metric used to determine the extent of symmetry in the drawing.

**Experiment 3 (Purchase, 1997)**

This third experiment considered the three aesthetics used before (crosses (c), bends (b), symmetry (s)), and introduced two more: maximizing minimum angles (m) and increasing orthogonality (o). Two variations for each aesthetic were considered: a + version (assumed to be easy to read), and a – version (assumed to be difficult to read). Only one graph was used, with 16 nodes and 28 edges. There were therefore 10 experimental graph drawings (for example, c+, o-, m-). Like the previous experiments, answering the three graph theoretic questions were the experimental task, and suitable node pairs were identified that could give a range of answers to the three questions.

One of the ways in which this experiment was different from the previous ones was in the use of a full set of aesthetic metric definitions. Aesthetic metric formulae that measured the aesthetic presence of all five of the aesthetics to be considered were defined: these could be applied to any graph drawing of any graph, and were scaled to lie between 0 and 1 to allow for numerical comparison. This set of aesthetics included an updated definition of symmetry which takes into account the perceptual affect of symmetry of bends and crosses (Purchase & Leonard, 1996).

In creating the experimental diagrams, the aesthetic variations were carefully controlled: while it was impossible to always have the metric values of the other four aesthetics identical in any aesthetic variant pair, we ensured that they lay within a reasonable ‘neutral’ range according to that
aesthetics’ distribution. We were also more careful about the definition of our extremes: the
drawings named with a + (e.g. $c+$, $o+$) were assumed to be easy to read. In some case this means
having a high aesthetic presence (e.g. orthogonality), in other cases it means having a low aesthetic
presence (e.g. crosses). We therefore introduced the terms ‘bend-less’ and ‘cross-less’ to refer to
these two aesthetics, so that all five aesthetic hypotheses could state that the increase of the
aesthetic would improve understanding. Figure 6 shows the experimental graph drawings, and their
associated aesthetic values.

The other main difference between this experiment and the former ones was the use of an
experimental online system. The methodology was similar. The experimental system presented:

- instructions, including definitions and worked examples;
- six practice drawings, with all three questions asked of each;
- a small logic puzzle that served as a filler task;
- the ten experimental diagrams: these were each displayed three times, one for each question.

All thirty instances of the diagram/question combinations were presented in random order; thus, the
three questions for the $m$- diagram were not necessarily asked consecutively. The diagrams were
also given a random orientation, apart from the highly orthogonal one which was randomly rotated
by a multiple of 90°. The nodes on the drawings were unlabelled, and the two nodes relevant to
each question were highlighted on the screen (thus removing the search time that was included in
the prior experiments). The subjects typed their answers to the questions.

There were 55 subjects in this experiment. The online system collected the subject answers to each
diagram/question pair, and the time taken for the subject to answer each question.
Figure 6: The graph drawings used in experiment 3, with their aesthetic metric values. Note that although the nodes are labeled in this figure, the nodes were blank when the drawings were displayed on the screen, with the two nodes relevant to the current question highlighted.

<table>
<thead>
<tr>
<th>graph</th>
<th>bend-less</th>
<th>cross-less</th>
<th>minangle</th>
<th>orthog</th>
<th>sym</th>
</tr>
</thead>
<tbody>
<tr>
<td>b+</td>
<td>0.96</td>
<td>0.97</td>
<td>0.38</td>
<td>0.27</td>
<td>0.75</td>
</tr>
<tr>
<td>b-</td>
<td>0.47</td>
<td>0.99</td>
<td>0.44</td>
<td>0.28</td>
<td>0.71</td>
</tr>
<tr>
<td>c+</td>
<td>0.82</td>
<td>1.00</td>
<td>0.46</td>
<td>0.33</td>
<td>0.63</td>
</tr>
<tr>
<td>c-</td>
<td>0.67</td>
<td>0.88</td>
<td>0.35</td>
<td>0.29</td>
<td>0.84</td>
</tr>
<tr>
<td>m+</td>
<td>0.71</td>
<td>0.98</td>
<td>0.62</td>
<td>0.22</td>
<td>0.74</td>
</tr>
<tr>
<td>m-</td>
<td>0.62</td>
<td>0.98</td>
<td>0.16</td>
<td>0.26</td>
<td>0.79</td>
</tr>
<tr>
<td>o+</td>
<td>0.82</td>
<td>0.98</td>
<td>0.42</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td>o-</td>
<td>0.82</td>
<td>0.98</td>
<td>0.41</td>
<td>0.21</td>
<td>0.68</td>
</tr>
<tr>
<td>s+</td>
<td>0.77</td>
<td>0.99</td>
<td>0.57</td>
<td>0.29</td>
<td>0.96</td>
</tr>
<tr>
<td>s-</td>
<td>0.87</td>
<td>0.90</td>
<td>0.44</td>
<td>0.25</td>
<td>0.00</td>
</tr>
</tbody>
</table>
This was again a within-subject methodology: the controls were the graphs and the other four aesthetics for each aesthetic variant pair, the independent variable was the aesthetic variation within each pair, and the dependent variables were the accuracy of subjects’ answers to the questions, and the time taken to answer each question.

As we had measured both accuracy and response time, we had two measures of understanding, These were treated independently in the statistical analysis. Although it may have also been useful to consider the correlations between time and accuracy to see to what extent we could consider these measures independently, this was not done for this experiment. As there were only two variations for each aesthetic, we could use a parametric test, as the measurable difference between the two variations is not important.

For both time and accuracy, we performed t-tests for each of the five aesthetics, to see which of our five hypotheses were confirmed. We found that the bends and crosses hypotheses were supported for accuracy, and that the crosses and symmetry hypotheses were supported for response time. There was no support for orthogonality, nor for maximising the minimum angle.

We also performed a Tukey pair-wise comparison for both time and accuracy on both the set of – drawings and the set of + drawings. This statistical test indicates which of an ordered set of results are significantly different from each other, and takes into account the Bonferroni correction that is usually applied when data is being used more than once in repeated statistical tests. The results of this test revealed that the c- (lots of crosses) drawing took significantly more time, and had a lower
accuracy, than all other – versions of the graph. In addition, the highly symmetric \( (s+) \) drawing took significantly less time than the minimum angle \( m+ \) and orthogonality \( o+ \) drawings.

The main lesson learnt from this experiment was the limitations of the formal experimental method: in order to run a well-controlled and valid experiment, the context in which it needs to be run constrains the extent to which the conclusions can be generalised. This experiment used only one set of subjects (computer science students); there were only two variations for each aesthetic; the aesthetic variations were measured by metrics that may not correspond with perception; only one graph of a particular size and structure were used; only three questions of limited scope were asked. The constraints of this experiment mean that, while the results are useful, they cannot necessarily be generalised to encompass larger graphs, or graphs of very different structure, or graphs used for different purposes. The danger was that, in presenting this work to researchers in graph layout who had never had their aesthetic assumptions experimentally challenged before, these researchers would generalise these results further than is appropriate, simply because they are the only experimental results in this area.

THE SEMANTIC EXPERIMENTS

Pilot experiment 4 (Purchase, Grundon & Naumann, 1996)

Having worked on abstract graphs, we then extended our efforts to semantic graphs, i.e., graphs which relate to some real-world domain. This pilot semantic experiment firstly explored three experimental methods for testing graph drawing aesthetics in the area of software-engineering diagrams, and then ran one experiment to test the effects of four aesthetics (bends, crosses, orthogonality, upward-flow).
The semantic notation chosen was object oriented class diagrams, based on the Booch notation (Booch, 1994) which was adapted for our purposes to suit Smalltalk code. The notation was made more graph-like (using ellipses instead of clouds, adapting the font variations used to illustrate the class type etc.) The notation was therefore sufficient to describe the code, while being uncluttered by unnecessary syntax that was irrelevant to our research questions regarding aesthetic variations. A brief tutorial handout was prepared, that could be given to all subjects, explaining the notation to be used in the experiment, and how it related to object-oriented design.

Three initial methodologies were attempted in a preliminary study which collected no data, did not use aesthetic-variant drawings, and whose aim was to identify an appropriate method for semantic aesthetics experiments.

For the first method, the task the subjects needed to perform was code adaptation. They were given the class and scenario diagrams for a system, and the matching code (system A). They were also given the class and scenario diagrams for a similar system (system B). Both systems were key-access systems with identical functionality, but with different underlying structures, the information being either centralised or distributed. The subjects’ task was to adapt the system A code so that it represented the diagrams given for system B, and to demonstrate the identical functionality when the new code is executed. The dependent variable was to be the time taken to complete the task. We tried this method with five subjects, four of whom took well over an hour to complete the task, while the fifth (an expert) took 35 minutes. We concluded that this adaptation task was inappropriate for experimental purposes, as it is too heavily influenced by the subjects’ prior coding experience, and did not rely enough on actual use of the diagrams where the aesthetic variation to be tested would be represented.
The second method required subjects to identify mistakes in given code. They were given a brief textual description of the system, a class diagram representing the structure of the system, and a code listing. Within 15 minutes, they needed to identify as many errors in the code as they can, in relation to the diagram. We hoped that this task would require more from the subjects in terms of understanding the class diagram than the prior pilot experiment had done. The dependent variable was the number of mistakes identified in the given time. We tried this method with only one subject, but this was sufficient to highlight its problems. Subjects need to look at all the code, even that containing no errors. They may also (rightly or wrongly) identify errors that were not part of our intended list, and that are unrelated to any need to understand the diagram. We felt that we still needed to find a methodology that relied more on understanding the diagrams.

The third method was the one ultimately chosen for the pilot experiment, and required subjects to match diagrams with code.

We used the two systems A and B which were designed for the first method investigated above: two systems with identical functionality, but with different underlying structures. Both of these systems were represented by a class diagram, and a code listing. Figure 7 shows the class diagram for system A. The subjects were given the two diagrams and the two code listings, and were told to match the diagrams to the code. So as not to bias their decisions, the diagrams and listing were not referred to by any ordinal scale (for example, by letters or numbers), but were colour coded. We also did not give any indication that there would be a perfect two-way match between the two diagrams and two code listings. Subjects were therefore asked to say whether the pink code corresponded with the green diagram, the blue diagram or neither diagram, and whether the yellow code corresponded with the green diagram, the blue diagram or neither diagram.
We considered four aesthetics (bends, crosses, orthogonality, upward-flow), each with two aesthetic variations (few/\text{-}, many/+). Thus, there were eight diagrams of system A ($Ab+ Ab- Ac+ Ao-$ etc), and eight diagrams for system B ($Bb+ Bb- Bc+ Bo-$ etc.). The diagrams were produced according to the existing aesthetic metric formulae, with appropriate controls.

The dependent variable was the time taken for the subject to get the correct answer (as measured by experimenter observation): subjects submitted their solution to the experimenter repeatedly until told that the answer was correct. The data from subjects who appeared to be using a quick trial-and-error method was eliminated.

As the task was more complex than the three graph-theoretic questions used in the syntactic experiments (taking approximately 30 minutes), a between-subjects methodology was used. Thus, each subject was randomly paired with a particular aesthetic variation (e.g. $b+$), and used only the two diagrams associated with that variation ($Ab+ and Bb+$). There were six subjects for each of the eight aesthetic variations, a total of 48 subjects.

The between subjects t-test statistical analysis on the data, used to determine whether the different presence of the aesthetics had any effect on the time taken to complete the task, revealed no significance at all. There was very little variation in the data, and any that there was could have been attributed to chance.

With regard the methodologies investigated in this experiment, we learned that repeatedly submitting results until getting a correct answer clouds the data, and that the more well-defined the
Figure 7: The Smalltalk application domain used in pilot experiment 4.
task, the better the experiment. Adapting code is a loosely defined task; matching two diagrams with code listings is much more well defined.

The most important lesson we learned from investigating all three methodologies, and from running the one experiment, was the difficulty of running valid between-subjects experiments. For a complex domain like understanding object-oriented programming, the expertise variability between subjects is likely to be so great that comparing their performance on the same task is clearly inappropriate. Even if a pre-experiment tutorial is provided to ensure that all subjects have the same base-level of domain-specific knowledge, for a programming task, there is still likely to be substantial subject variability that will make the data meaningless.

**Experiment 5a (Purchase, McGill, Colpoys & Carrington, 2001)**

For our first formal semantic aesthetics experiment, we used an adaptation of the matching experiment described in experiment 4. Our domain was UML class diagrams (Rumbaugh et al., 1999), and five aesthetics were considered: bends, node distribution, edge variation, flow, and orthogonality. As the prior experiments had given strong support for removing edge crosses, this aesthetic was not included in the list, and all experimental diagrams had no crosses at all.

The hypotheses for our experiment were:

- Increasing the number of bends in the representation of associations in a UML class diagram decreases the understandability of the classes and the relationships in the application.
o Placing the visual objects representing classes in a UML class diagram such that they are evenly distributed over the drawing area increases the understandability of the classes and the relationships in an application.

o Representing associations in a UML class diagram with lines of similar length increases the understandability of the classes and the relationships in an application.

o Representing the direction of associations and inheritance in a UML class diagram with arrows that point in the same direction increases the understandability of the classes and the relationships in an application.

o Placing the visual objects representing associations, inheritance and classes in a UML class diagram on an underlying unit grid increases the understandability of the classes and the relationships in an application.

The application domain was a simple UML class diagram of 13 objects and 17 edges, and modeled a small information technology company. A textual specification in simple English was produced which matched this class diagram. The experimental task was to match this specification against a set of experimental diagrams, indicating whether each diagram matches the specification or not.

For each of the five aesthetics, a "low-effect" (-) and a "high-effect" (+) version of the diagram was produced. To ensure that there were no confounding factors between the aesthetics, the ranges were controlled as much as possible. For example, to remove any confounding factors in a diagram pair for a particular aesthetic, the measurement of all other aesthetics were kept within a "middle-effect" range. A control diagram that conformed to a "middle-effect" range for all the aesthetics as much as possible was also created. There were therefore a total of 11 correct experimental diagrams.
Aesthetic presence was measured using computational metrics, and Figure 8 shows the aesthetic values for each of the experimental diagrams.

<table>
<thead>
<tr>
<th>Diagram</th>
<th>Aesthetic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bends (b)</td>
</tr>
<tr>
<td>b+</td>
<td>1</td>
</tr>
<tr>
<td>b-</td>
<td>0.71</td>
</tr>
<tr>
<td>o+</td>
<td>0.85</td>
</tr>
<tr>
<td>o-</td>
<td>0.85</td>
</tr>
<tr>
<td>ev+</td>
<td>0.85</td>
</tr>
<tr>
<td>ev-</td>
<td>0.85</td>
</tr>
<tr>
<td>n+</td>
<td>0.85</td>
</tr>
<tr>
<td>n-</td>
<td>0.85</td>
</tr>
<tr>
<td>f+</td>
<td>0.85</td>
</tr>
<tr>
<td>f-</td>
<td>0.85</td>
</tr>
<tr>
<td>control</td>
<td>0.85</td>
</tr>
<tr>
<td>example</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 2: The computational aesthetic values for the experiment 5a diagrams.

Ten incorrect diagrams were created by randomly changing the origin or destination of one relationship per diagram. The layouts of the incorrect diagrams were visually comparable to those of the correct diagrams: as we did not intend to analyse the responses to the incorrect diagrams, their layout was not important. However, it was, of course, important to include incorrect diagrams in the experimental set (so that the correct answer to each diagram presented was not the same), and for these incorrect diagrams to be visually comparable to the correct diagrams (so they could not be identified by mere visual pattern matching).

The first phase of the experiment was preparation: subjects were given a brief UML tutorial which explained the meaning of UML class diagrams, and, using a simple example, described its semantics. Subjects were not expected to have any prior knowledge of UML and this tutorial provided all the UML background information that they required for the experimental task. A
worked example demonstrated the task that the subjects were to perform, presenting a small specification with four different diagrams, and for each diagram indicating whether it matched the given specification or not. Care was taken to ensure that neither the tutorial nor the worked example would bias the subjects towards one layout over another.

As part of this preparation, subjects were also given an UML class diagram (which conformed to a ‘middle effect’ for all the aesthetics) of the application domain, and the accompanying textual specification: this was the specification against which they would need to match the experimental diagrams. The subjects were asked to study this specification closely, and memorise it if possible.

The second phase of the experiment was conducted online: UML diagrams were presented in turn on the screen, and subjects needed to press ‘Y’ or ‘N’ depending on whether they thought the diagram matched the specification or not. A copy of the specification was propped up against the computer for quick reference. Pilot tests had revealed that 50 seconds was a suitable cut off time for each diagram, so, if a subject had not responded within 50 seconds, the next diagram was presented, and an incorrect answer recorded. The practice diagrams helped the subjects get used to this cut off time. The diagrams were presented in blocks of eight, with a rest break between each block (the length of which was controlled by the subject).

The experimental diagrams were presented in random order for each subject: the set of diagrams included two instances of each of the 11 correct diagrams, and one instance of each of the incorrect diagrams: 32 in total. 16 practice diagrams were presented at the start of the experimental session: the data for these was not collected.
A within-subject methodology was used to reduce any variability that may have been attributable to differences between subjects: thus, each subject’s performance on one layout was compared with his or her own performance on alternative layout. The practice diagrams and random order of presentation helped counter any learning effect.

Thirty subjects took part in this experiment. Both response time and accuracy data for the 11 correct diagrams was collected. We used a t-test to analyse both the response time and accuracy data for each of the five aesthetics independently, using the ‘middle effect’ diagram as the middle variation point for all of the aesthetics. We realised after analysing the data that a non-parametric test would have been more appropriate, as the three variation measures were ordinal and not continuous, similar to experiment 2 above.

There was no significance in the accuracy data. In the response time data, the only significant results were that the middle effect diagram had the best performance for bends and edge variation, and for flow, the upward direction produced the quickest response. The results were confusing and contradictory.

In reassessing the diagrams that we used for this experiment, we felt that perhaps the problem was in the measurement of the presence of the aesthetics. The computational metrics, while useful for measuring the aesthetics from a computational point of view, may be less useful for measuring perceptual aesthetic presence from a human point of view. For example, the orthogonality metric measures the extent to which the nodes and edges are placed along an underlying unit grid, but the human perception of orthogonality in a diagram may not match the numerical value produced by the metric. This phenomenon may particularly be the case for aesthetics which are global; that is,
require an overall assessment of the entire diagram, for example, orthogonality, symmetry, or node distribution.

We therefore decided to run the experiment again, but this time with a different set of diagrams, created according to humans' perception of the presence of each aesthetic in the diagrams, rather than according to the defined metrics.

**Experiment 5b (Purchase, McGill, Colpoys & Carrington, 2001)**

The experimental methodology for this experiment was identical to that used in experiment 5a, and the same UML application domain was used. The difference was in the way in which aesthetic presence within the experimental diagrams was determined.

Seven aesthetics were considered in this experiment: bends, node distribution, edge variation, direction of flow, orthogonality, edge lengths, and symmetry, each with three variations. For each aesthetic, three diagrams were created by hand: low-effect (-), middle-effect (0) and high-effect (+). To confirm that these diagrams had an appropriate amount of low-, middle- and high-effect of the aesthetics, and that the aesthetics were appropriately controlled, simple perception experiments were performed with 10 subjects. The subjects were asked to rank sets of three diagrams according to the presence of the aesthetic. For example, a subject was shown the n+, n0 and n- diagrams and asked to rank them according to the extent of even node distribution in the diagrams. The possible confounds of symmetry and orthogonality were also addressed in these interviews. For example, the subjects were asked to rank the n+, n0 and n- diagrams according to symmetry, the desired result being that they would find it difficult to do so. We needed to ensure that a difference in performance on the node distribution diagrams could not be attributed to differences in symmetry
and orthogonality. The bends and flow aesthetics were not perceptually tested in the production of
the diagrams, as their presence is better assessed computationally (for example, by counting the
number of bends or counting the number of edges pointing upwards). However, the bends and flow
diagrams were tested for the possible symmetry and orthogonality confounds. Figure 9 shows the
flow+ (consistent direction of flow) diagram.

Note that the decision to measure aesthetic presence by perception rather than by computational
measures introduces a new feature into the three-dimension experimental framework described at
the beginning of this chapter (usability measurement, the nature of the graph, the effect being
investigated). If the effect being investigated is that of individual aesthetics, then the method by
which that aesthetic presence is measured is another dimension of the experiment.
Figure 8: The flow+ class diagram used in experiment 5b.
The same methodology was used as in experiment 5a, with 35 subjects taking part, and the same preparatory materials and online system being used. The same 10 incorrect diagrams from experiment 5a were used. The 21 correct and 10 incorrect diagrams were each presented once in the online task: a total of 31 experimental diagrams. Pilot tests indicated that a reduced cut off time of 40 seconds would be more appropriate than the 50 seconds used previously.

Like before, both response time and accuracy data for the 21 correct diagrams was collected. We used t-tests for each aesthetic (although a non-parametric test and pair-wise analysis would have been more appropriate).

The only significant results obtained supported the hypothesis that reducing the number of bends improves both time and accuracy. There were some confusing results regarding edge variation that showed the middle effect variation to be worse (with respect to response time only).

While it is tempting to say that no aesthetics matter (apart from reducing the number of bends, which matters a little bit), our view is that there are other issues that need to be considered, in particular, the notion of semantic grouping, where related (but unconnected) objects are placed in close proximity.

One of the main lessons we learnt from these two experiments was the limitations in the production of the experimental materials: we considered (and tried to control for) seven aesthetics, but there are many other visual features regarding diagram layout that we did not control for. For example, some subjects said that they found the task easier if the classes in the diagram were presented in the
same vertical order as shown on the textual specification. We think that visual proximity of semantically related objects might also affect performance (Petre, 1995); this and other visual features (for example, the amount of white space, ratio of diagram width to diagram height, left-to-right placement of nodes etc) were not considered at all in our controls. It became clear that it would be impossible to identify (and control for) all the possible layout aesthetics that may affect comprehension of the diagram, and this was revealed as a limitation of using a formal experimental method that requires strict controls.

 Statistical lessons that we learnt included the necessity to identify whether a parametric or non-parametric test is required, and the usefulness of using a pair-wise test if more than two variations have been used: multiple t-tests need to be adjusted according to the Bonferroni correction; pair-wise tests (for example, Tukey), automatically take that correction into account.

 We also considered more closely than before the potential interaction between accuracy and response time (as indicated by linear correlations): if there is a significant correlation between these two measures of understanding for any of the experimental diagrams, then the importance of any significant results for that aesthetic in either of the measures is reduced.

 **Experiment 6 (Purchase, Colpoys, McGill & Carrington, 2002)**

 Our most recent experiment uses the same overall methodology as experiment 5b, but in the domain of UML collaboration diagrams. Six aesthetics were considered: node distribution, edge variation, direction of flow, orthogonality, edge lengths, symmetry. Each of these had three variations (+, 0,-, for example, *nd*+ represents a diagram with a high node distribution), although flow had four variations, depending on the position of the actor in the diagram: a total of 19
experimental diagrams. Figure 10 shows the UML collaboration diagram with even node
distribution.

The UML collaboration diagram used in the experiment represented a simple preferential voting
system, and comprised an actor, 10 objects, 17 edges and 22 messages. Perception pre-tests were
used with eight subjects, to validate the hand-drawn diagrams, to ensure their conformance with the
required aesthetic variations (in the same manner as in experiment 5b). Ten incorrect diagrams were
produced by randomly changing the origin or destination of one link per diagram. The layouts of
the incorrect diagrams were visually comparable to the correct diagrams.

The same preparatory and online experimental process was followed: this time the subjects needed
to match a pseudo-code listing with the collaboration diagram. Thirty-five subjects took part. There
were 16 practice diagrams, and 29 experimental diagrams (19 correct and 10 incorrect). The cut off
time was 60 seconds.

The statistical method used in this experiment was more appropriate for ordinal variations, and for
pair-wise analysis of data. We used the Friedman test for each aesthetic (for both response time and
accuracy) to determine whether there was any variation in the data, and if there was, we used a non-
parametric version of the Tukey pair-wise comparison test to determine where the differences lay:
the Nemeyi test (Zar, 1999).

Figures 11 and 12 represent the data from the experiment. The only significant results we obtained
were with edge variation, and these were contradictory ($ev0$ is slower than $ev+$, and $ev-$ is less
accurate than both $ev+$ and $ev0$). However, there was also a significant linear correlation between
Figure 9: The UML collaboration diagram used in experiment 6, with a balanced node distribution ($nd^+$)
Figure 10: The response time results for experiment 6.

Figure 11: The accuracy results for experiment 6.
time and accuracy for the ev0 diagram, which makes the data for edge variations difficult to interpret.

One point that was revealed quite strongly in this experiment was the problem of trying to cover more than one hypothesis in a single experiment. In all our experiments, we have addressed more than one aesthetic, and in this particular one, we were looking at six in total. While this is an efficient way to consider several aesthetics, it can cause problems with methodology (for example, choice of time cutoff, choice of errors introduced into the incorrect diagrams) and the interpretation of results.

RESULTS AND REFLECTIONS

These experiments have covered the investigation of graph drawing aesthetics in several ways, and have produced both useful and confusing results (see Table 3). It became clear very early on in the project that reducing the number of edge crosses was by far the most important aesthetic criterion, and that reducing the number of bends was also important. The results for other aesthetics are less convincing: edge variation and flow gain some support within the UML domain, but the evidence is far from overwhelming.
<table>
<thead>
<tr>
<th>Experiment</th>
<th>The nature of the graph</th>
<th>Aesthetic</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>syntactic</td>
<td>Pilot experiment only: insufficient data points for analysis</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>syntactic</td>
<td>bends</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crosses</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>symmetry</td>
<td>no result</td>
</tr>
<tr>
<td>3</td>
<td>syntactic</td>
<td>bends</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crosses</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>minimum angle</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>orthogonality</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>symmetry</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td>4</td>
<td>semantic (class diagrams)</td>
<td>Between subjects analysis: subject variation produced invalid data</td>
<td></td>
</tr>
<tr>
<td>5a</td>
<td>semantic (UML class diagrams)</td>
<td>bends</td>
<td>contradictory result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>edge length variation</td>
<td>contradictory result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>direction of flow</td>
<td>hypothesis confirmed for upward direction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>node distribution</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>orthogonality</td>
<td>no result</td>
</tr>
<tr>
<td>5b</td>
<td>semantic (UML class diagrams)</td>
<td>bends</td>
<td>hypothesis confirmed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>edge length</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>edge length variation</td>
<td>contradictory result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>direction of flow</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>node distribution</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>orthogonality</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>symmetry</td>
<td>no result</td>
</tr>
<tr>
<td>6</td>
<td>semantic (UML collaboration diagrams)</td>
<td>edge length</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>edge length variation</td>
<td>contradictory result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>direction of flow</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>node distribution</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>orthogonality</td>
<td>no result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>symmetry</td>
<td>no result</td>
</tr>
</tbody>
</table>

Table 3: A summary of the results of the experiments.

We believe that there are other influences that need to be considered:

- We think that the notion of a possible ‘critical mass’ for any of the aesthetics should not be ignored. It may be that a small number of edge crosses may be acceptable if, by including
them, a high degree of symmetry can be portrayed, but that once the number of edge crosses exceeds a critical mass, then the crosses aesthetic takes a higher priority over symmetry.

- There are many other visual properties that it would be impossible to control to ensure that the data collected can be considered valid within the confines of the formal experimental method. We controlled for features like node size, edge width etc., but did not also consider other syntactic features like white space, nor semantic features like node proximity. It would be impossible to identify all the potential visual influences, and then be able to control them effectively.

The results presented here do not include results from preference data collected in experiments that also form part of this project. (Purchase, Carrington & Allder, 2002) The results from the preference experiments indicated that subjects’ preference does not always match performance: the aesthetics that subjects prefer are not always the aesthetics that result in improved performance.

Although what has been learned about graph drawing aesthetics may be considered limited in its scope and generalisability, much has been learning about experimenting in this area, and about the process of defining a new empirical research area. We made many mistakes along the way, and all our results need to be interpreted in the context of the particular method used in the experiments. The most important experimental issues that emerged include:

- the importance of providing preparatory materials that are easy to understand, that include very clear definitions of all the concepts required for the task, and that present a clear worked example of the task that the subjects are to perform (with the correct answer and explanation);
• the importance of having a very clearly defined experimental task that does not require
  skills unrelated to the research question; for example, a task that requires subjects to write a
  program is much less clearly defined than a task that requires subjects to answer yes or no to
  a question.);
• the importance of defining the experiment clearly when writing it up for publication, with
  respect to hypotheses, variables (independent, dependent and control), within-subject or
  between-subject design, quantitative or qualitative data;
• the importance of choosing the correct statistical tests: the distinction between parametric
  and non-parametric data, and the use of pair-wise comparison tests when more than two
  variations are used.

Most importantly, we realised the importance of presenting the results of these experiments in such
a way that their results are not generalized beyond what is appropriate. It was important to state
clearly the limitations of each experiment, and to present the analysed data separately from our own
interpretation of it. By making a clear distinction between the actual data that was obtained from the
experiment, and what this data meant to us, we allowed the readers of our publications to consider
the factual data independently of our opinion, and therefore gave them the opportunity to form their
own conclusions.

FUTURE TRENDS

The results of this project have been welcomed by the graph drawing community who seek
empirical validation of their assumptions. However, it is plain that the results of each of these
experiments are limited by the necessary constraints of formal experimentation, and that this work
could continue in the same vein for a long time to come. However, there are two important alternate routes that this research may take:

- Perceptual and Cognitive issues: The main problem with all these experiments is that they are purely concerned with behaviour. Thus, while the experiments show that crosses are indeed detrimental to the understanding of a graph drawing, we cannot say why this is the case from the point of view of what is currently known about both perceptual visual processes nor about cognition. There is a substantive literature on perceptual and cognitive models: integrating this theory with these experimental results would not only enable them to be explained, but may provide a predictive model by which a user’s performance in a graph that has a high presence of a particular aesthetic may be predicted with reference to the model. Some initial work has recently been performed in this area: Ware, Purchase et al. (2002) recently ran an experiment that hypothesized the effect of the angular continuity of a path in a shortest path task, based on knowledge of perception of continuity. They found that the angular continuity was indeed an important factor.

- Real-world tasks: The experiments have all been run in limited environments, with artificial tasks. For this type of research to be truly useful for the development of practical software systems that include automatic graph layout algorithms, investigations of real world use ‘in the field’ are important. Real-world experiments are more difficult to run: they are more time consuming on the part of both experimenter and participants, and the data collected is less quantitative, less easy to characterize, and therefore less easy to analyse. There is less chance of recruiting a sufficient number of participants for statistical analysis, and it may therefore be more difficult to convince others of the validity and generalisability of the results.
This project has revealed a great deal, both about the worth of existing graph drawing aesthetics, and about the process of empirical studies. The former is of interest to the designers of graph drawing algorithms, the latter is of interest to experimental psychologists. Two other communities that could also be served if the research were to be focused differently are cognitive scientists, and the designers of domain-specific diagram support software.

ACKNOWLEDGEMENTS

With thanks to: Peter Eades, who initially suggested this research area, and who has been supportive throughout; people who assisted with the overall approach and aims of the experiments (Robert Cohen, David Carrington); people who prepared experimental materials and ran experiments (Shelina Bhanji, Murray James, Steven Grundon, Daniel Naumann, Jo-Anne Allder, Linda Colpoys, Matthew McGill); people who assisted in defining aesthetic metrics (Shelina Bhanji, David Leonard); people who took part in the experiments (many students in the School of Information Technology and Electrical Engineering at the University of Queensland over several years); institutions who provided resources for this project (The Australian Research Council, The University of Queensland).

REFERENCES


