Pairwise display of high dimensional information via Eulerian tours and Hamiltonian decompositions

C.B. Hurley and R.W. Oldford[*]

April 3, 2008

Abstract

A graph theoretic approach is taken to the component order problem in the layout of statistical graphics. Eulerian tours and Hamiltonian decompositions of complete graphs are used to ameliorate order effects in statistical graphics. Similar traversals of edge weighted graphs are used to amplify the visual effect of selected salient features in the data. Relevant graph theory is summarized and classic algorithms are tailored to this problem. Graphics for multiple comparisons are reviewed and a new display developed that is based on graph traversal. Interaction plots are improved and new ones proposed. Improved star glyph displays of multivariate data are described. Parallel coordinate displays tailored to particular features of the data are developed. The methods and new graphical displays are made available as an R package.

1 Introduction

Graphical displays often require an ordering of their components (e.g. scatterplot matrices, glyphs, parallel coordinate plots, etc.). The ordering itself is an encoding of information that, if neglected, could hide or distort important information in the data.

Perhaps the best known example is that of Chernoff's faces. There the information perceived is very much dependent on the order in which variates are assigned to the face features (Chernoff and Rizvi, 1975). This lack of invariance to the order renders the faces of limited use in data exploration. However, post exploration, careful assignment of variates to features can make an effective presentation graphic.

Ordering has often been used to good effect, to reveal more about the data, to encourage data comparisons, and to make large datasets coherent – in short to meet Tufte's (1987) principles of graphical excellence. Cleveland's (1995) trellis display of barley data is a convincing example of the benefits of a well chosen order: here the main effect category

[*] Research supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

levels are ordered by their median, and an anomaly in the data is immediately evident. Ankerst, Berchtold and Keim (1998), Friendly and Kwan (2003) and Hurley (2003) describe methods for sorting variables so that similar variables are positioned adjacent to each other in multivariate displays such as scatterplots and parallel coordinates, thus simplifying interpretation.

In what follows, we explore how ordering might be automated and more widely applied in statistical graphics. Particular graphics addressed will be a display for multiple comparisons, interaction plots, star glyphs and parallel coordinate plots. Some of these are new displays, others are new variations on existing displays. There are no doubt many other visualization methods where our approach will apply.

We abstract the problem to one of graph traversal, and so are able to bring mathematical results and algorithms to bear on it. In some cases, traversals can be chosen to ameliorate the order effect, rendering the display more nearly invariant to the component ordering. In other cases, some traversals are chosen over others to reinforce the desired effect of the display.

Section 2 surveys the relevant graph theory and summarizes those mathematical results most applicable to the ordering problem. The section stands on its own and is applicable to any statistical problem where order is of concern, not just those of data visualization.

In Section 3, we demonstrate the ordering effects of a number of statistical graphics and show how the relevant graph theoretic results can be used to produce both new (in the case of multiple comparisons) and improved (in the case of interaction plots, star glyphs and parallel coordinate plots) statistical displays.

Section 4 describes the algorithms used in constructing the graph traversals used and some closing remarks are made in the last section.

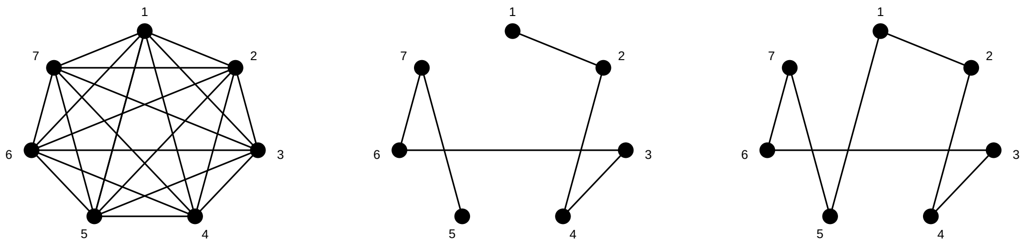
2 Graph theory

The complete graph on n nodes or vertices is an undirected graph, denoted K_n, with vertex set $V(K_n) = \{1, 2, \ldots, n\}$ and edge set $E(K_n) = \{e_{ij} | i, j \in V(K_n), i \neq j \text{ with } e_{ij} = e_{ji}\}$ (when there is no ambiguity, the edge e_{ij} might also be written ij). The cardinality of the vertex set is called the graph's *order*, here n, and that of its edge set the *size* of the graph, here $|E(K_n)| = n(n-1)/2$. Figure 1(a) shows K_7.

A complete graph is a convenient representation of n objects (the nodes) together with all possible pairings (the edges). Any path along edges of the graph simultaneously provides an arrangement of those objects identified with the nodes of the path and of the pairings identified with the edges of the path.

2.1 Hamiltonians, eulerians, and hamiltonian decompositions

A path is called a *hamiltonian path* if it visits all vertices of a graph exactly once. The hamiltonian path of Figure 1(b) orders the nodes as 1243675 (or the reverse), and is



(a) K_7 (b) A hamiltonian path (c) A hamiltonian cycle

Figure 1: K_7, euler tours, and hamiltonians.

identified with a permutation of the nodes. The set of all hamiltonian paths on a complete graph is the set of all permutations. Closing a hamiltonian path by joining its ends, as in Figure 1(c), creates a *hamiltonian cycle* which can be identified with many permutations (each being a cyclic permutation of the original). A graph G is *hamiltonian* if it contains a hamiltonian cycle and a graph G is *hamiltonian connected* if any pair of vertices are the ends of a hamiltonian path. Complete graphs K_n are hamiltonian for all n and K_n contains $(n-1)!$ distinct hamiltonian cycles.

Equivalently, a path can be regarded as providing an ordering on the edges it contains. Figure 1(b) orders its edges as $12, 24, 43, 36, 67, 75$ to which the hamiltonian cycle of Figure 1(c) adds the edge 51. Often interest lies in visiting (and hence ordering) all of the edges in a graph. A path which contains all of the edges of a graph, visiting each edge exactly once is called an *eulerian path (or eulerian trail)* and if the path is closed then the traversal is called an *eulerian tour.* A graph G which has an eulerian tour is called *eulerian.* The graph K_7 of Figure 1(a) is eulerian. An eulerian tour of a complete graph provides an arrangement of all possible pairings of the nodes. One such tour for K_7 is $T_0 = 1234567461427157352631$.

2.1.1 Many to choose from

As with hamiltonians (cycles and paths), there need not be a unique eulerian tour for a given graph. Typically there are a great many to choose from. For example, K_7 admits $129,976,320$ eulerian tours that are not cyclic permutations of one another (first determined by Reiss, 1871-3; see McKay and Robinson, 1998) while K_{21} has more than 3.4×10^{184}. (For odd $n \leq 21$, the number is available online via Sequence A007082 of the Online Encyclopedia of Integer Sequences (Sloane 2004).)

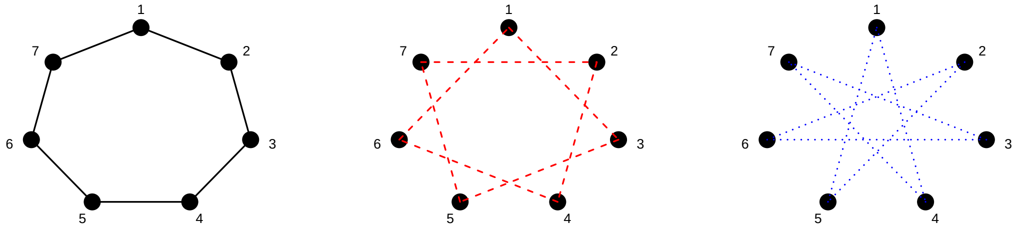
While an eulerian tour of a complete graph produces an arrangement of all possible pairings, it may be that some eulerian tours (arrangements) are preferred over others. With some measure of the value of each, the eulerian tours could, in principle, be ordered and the best selected.

For example, if each edge in the graph had a weight, we might prefer eulerians whose edge weights were by some measure as low (high) as possible in the early part of the sequence. A greedy eulerian might be one which began with the lowest (highest) weight edge, then amongst the edges available chose the next edge with lowest (highest) weight, and so on.

Alternatively, one might prefer eulerians with some particular structure.

2.1.2 Hamiltonian decomposed eulerian tours

One possibility is that the eulerian tour be composed entirely of edge-distinct hamiltonian cycles, a so-called *hamiltonian decomposition.* Figure 2 shows a hamiltonian decomposition of K_7. Note that this decomposition is also a *symmetric* hamiltonian decomposition because a node labelling exists which makes all cycles symmetric about the same node (this decomposition is in fact symmetric about every node).



(a) $H_{11} = 12345671$ (b) $H_{12} = 13572461$ (c) $H_{13} = 14736251$

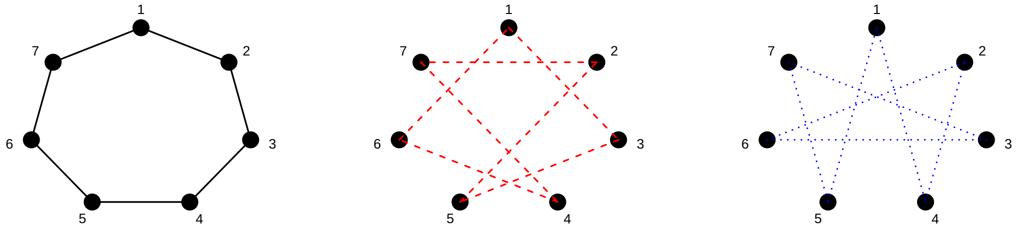
Figure 2: A hamiltonian decomposition $H_1 = H_{11} : H_{12} : H_{13}$ of K_7.

An eulerian tour can be had by joining these hamiltonian cycles, in any order, at the same node. For example, $T_1 = 1\ 234567\ 1\ 357246\ 1\ 473625\ 1$ is an eulerian tour that joins the three hamiltonians at 1 in the order of H_{11}, H_{12}, H_{13}; $T_2 = 2\ 345671\ 2\ 753164\ 2\ 514736\ 2$ joins the cycles at node 2 in the order H_{12}, H_{11}, H_{13} with the middle cycle reversed. For any hamiltonian decomposition, an eulerian tour can be constructed by varying the order of the hamiltonian cycles, varying the direction in which each cycle is traversed, and varying the point of contact between the cycles.

Moreover, the hamiltonians in Figure 2 are presented in canonical form (in terms of node labelling as given in Colbourn, 1982), so permuting the node numbers on Figure 2(a) and carrying that assignment across the hamiltonians of Figure 2(b) and (c), can produce a different hamiltonian decomposition and consequently many more eulerian tours.

By construction, these different decompositions will be isomorphic to one another (two hamiltonian decompositions H and H' are isomorphic if there is a one to one mapping of the nodes of the graph onto themselves which maps each hamiltonian cycle of H onto a hamiltonian cycle of H') and will sometimes be identical (e.g. the decomposition produced by mapping the nodes 1234567 of Figure 2 to 2715436 is identical to that of mapping 1234567 to 4675321). In this way the hamiltonian decomposition of Figure 2 generates a class of decompositions. It does not, however, generate all hamiltonian decompositions of K_7.

There is only one other set of isomorphic hamiltonian decompositions of K_7 which is not isomorphic to that of H_1 from Figure 2. The canonical form for this set is H_2 of Figure 3 (see Colbourn, 1982).



(a) $H_{21} = 12345671$ (b) $H_{22} = 13527461$ (c) $H_{23} = 14263751$

Figure 3: $H_2 = H_{21} : H_{22} : H_{23}$ is the canonical form of the second (and only other) set of hamiltonian decompositions of K_7.

H_2 is also a symmetric decomposition, though one with many fewer symmetries than H_1 (i.e. only about node 1 in H_2). The fewer symmetries result in a smaller group order of the automorphisms (6 for H_2 vs. 42 for H_1) and consequently many more distinct (though isomorphic) decompositions (viz. $7!/42 = 120$ for H_1, $7!/6 = 840$ for H_2).

As before, the cycles of each distinct decomposition can be arranged in many ways to produce different eulerian tours. Using H_1, there will be thousands of distinct hamiltonian decomposed eulerian tours for K_7; using H_2 there will be seven times as many to choose from.

For larger orders of complete graphs, the number of non-isomorphic classes of hamiltonian decompositions is huge. There are 122 non-isomorphic decompositions of K_9 and more than $45,000$ for K_{11} (Colbourn, 1982, stopped computing more after finding this many).

While in principle it is possible to order the hamiltonian decompositions according to some preference, it is rarely practicable. Even choosing the single hamiltonian having the smallest total edge weight (i.e. the travelling salesman problem) is NP hard.

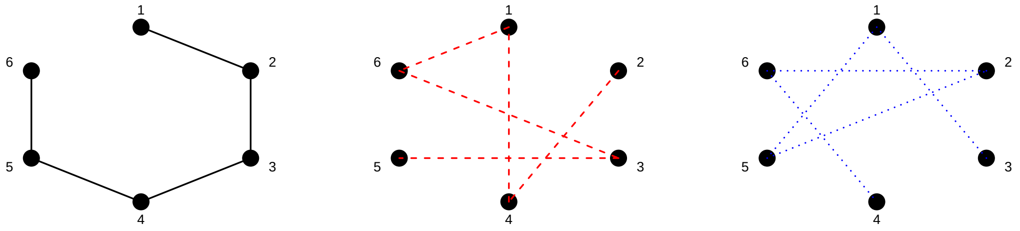
2.2 General results for complete graphs

If G is a connected graph, G is eulerian if and only if it is an even graph (i.e. every vertex has an even number of edges), or equivalently if and only if G has a cycle decomposition. Since complete graphs of odd order are connected and even, eulerian tours and hamiltonian decompositions exist.

The same notions can be extended to the complete graphs of even order through the following well known results which have been attributed to Walecki (by Lucas, 1892; e.g. see Alspach, et al 1990):

Decomposition of complete graphs. K_n can be decomposed as follows:
For $n = 2m + 1$, into either
m hamiltonian cycles, or
m hamiltonian paths and an almost-one factor.
For $n = 2m$, into either
m hamiltonian paths, or
$m - 1$ hamiltonian cycles and a 1-factor (or perfect matching).

The hamiltonian cycle decomposition for the case of odd n has already been illustrated. When n is even, the analogous decomposition of K_{2m} is into hamiltonian paths rather than cycles. Figure 4 shows one such decomposition for K_6. This was had directly from the hamiltonian cycle decomposition of Figure 3 by deleting node 1 and relabelling nodes $2 - 7$ as $1 - 6$; one might just as easily have used Figure 2.

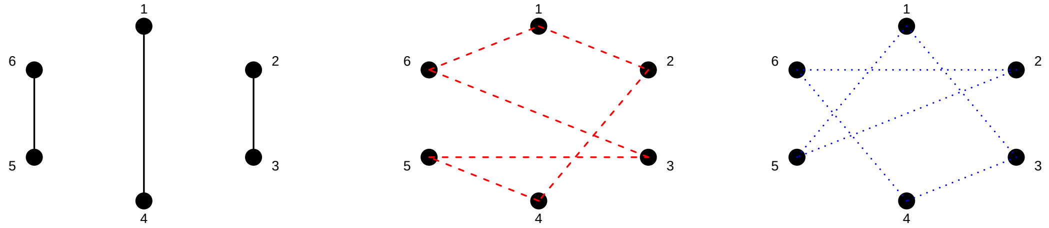


(a) $h_{31} = 123456$ (b) $h_{32} = 241635$ (c) $h_{33} = 315264$

Figure 4: $H_3 = h_{31} : h_{32} : h_{33}$ is a hamiltonian *path* decomposition of K_6.

Alternatively K_6 can be decomposed into a 1-factor (or perfect matching) and two hamiltonian cycles as shown in Figure 5. Similarly, K_{2m+1} is decomposable into m hamiltonian paths and an "almost 1-factor" (i.e. a 1-factor perfectly matching $2m$ points plus a single isolated vertex).

Although K_{2m} is not even, and hence not eulerian, m edges can be added to produce a



(a) 1-factor (or perfect matching) (b) 1245361 (c) 1346251

Figure 5: K_6 decomposed into a 1-factor and two hamiltonian cycles.

graph that will be eulerian and will have hamiltonian cycle decompositions. For example, simply close the hamiltonian paths of Figure 4 producing double edges 16, 25, and 34. If only an eulerian path is required, only $m-1$ edges need be added – an eulerian path exists for any connected graph having exactly two nodes of odd degree, so the $m-1$ edges added must be such as to satisfy this condition. For example, in Figure 4, add only extra edges 25, and 34; then an eulerian path will begin and end at the endpoints of the remaining hamiltonian path h_{31} of 4(a) and have the hamiltonian cycles (from the extra edges) be toured at any node along the path h_{31}.

Alternatively, one could start with the hamiltonian cycle decomposition of K_{2m} (e.g. $m = 3$ in Figure 5) and add $m-1$, or m, edges to the perfect matching to create a hamiltonian path, or cycle, respectively that will in turn permit an eulerian path, or cycle on the augmented graph.

Because complete graphs of even order can always be augmented to achieve eulerian paths, etc., it will be convenient to have a single notation for both K_{2m+1} and the $m-1$ edge augmented graph of K_{2m}. Denote by

$$K_n^e = \begin{cases} K_{2m+1} & \text{if } n = 2m+1 \\ K_{2m} + G(K_{2m}) & \text{if } n = 2m \end{cases}$$

where $G(K_{2m})$ is a subgraph of K_{2m} having $m-1$ edges chosen so that the graph resulting from the sum has exactly two odd nodes.

It will also be convenient to refer to an *eulerian of* K_n^e to mean an eulerian *tour* of K_n^e when $n = 2m+1$ and an eulerian *path* of K_n^e when $n = 2m$. Similarly a *hamiltonian decomposition of* K_n^e will refer to a hamiltonian *cycle* decomposition when $n = 2m+1$ and a decomposition into $m-1$ hamiltonian cycles plus one hamiltonian path when $n = 2m$.

There will of course be many hamiltonian decompositions, and many more eulerians, of K_n^e to choose from.

3 Applications to statistical graphics

In this section we illustrate the use of eulerians and hamiltonians to order the components of several graphical displays.

We begin by illustrating the use of an eulerian tour in the context of the pairwise comparison of treatment groups. We produce a simple but powerful new display for this classic problem.

Our second example addresses another classic problem, that of displaying the interactions between two factors from an experiment. Here hamiltonian paths and eulerian tours are used.

Next the problem of glyph construction whose purpose is the visual clustering of multivariate data is considered. In particular, we look at improving star glyphs by ordering variables according to eulerian tours and hamiltonian decompositions. The results are dramatic improvements over the standard stars.

Finally, parallel coordinate displays are examined. Here we show how hamiltonian paths, eulerian tours, and hamiltonian decompositions might all be used to construct different parallel coordinate displays, each suited to a different purpose.

3.1 Pairwise comparisons

In the classic one-way anova situation, several conditions are compared at once for differences in some outcome of interest. Also of interest are all pairwise comparisons, with correction for the problem of multiple comparisons.

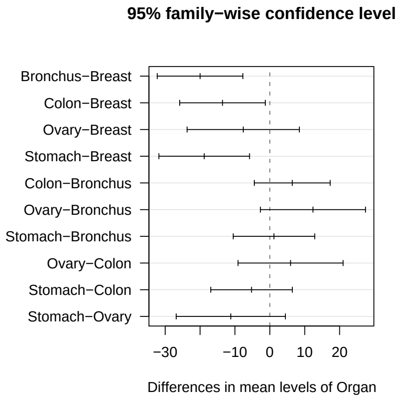


Figure 6: A standard layout of 95% confidence intervals for differences of mean survival times (square root scale), corrected for multiple comparisons.

To be concrete, we take data on the survival times of terminal patients with different types of cancer – viz. Breast, Bronchus, Colon, Ovary, or Stomach from a study reported in Cameron and Pauling (1978). The square root of the survival times are used to better approximate normality. Figure 6 shows the 95% simultaneous confidence intervals for the pairwise difference in means, using "Tukey's honest significant differences". This tidy little layout is fairly standard for multiple comparisons (e.g. it is the default plot method for `TukeyHSD`, Bates, 1997+).

Comparisons whose intervals do not overlap the vertical zero line are statistically significant (e.g. Bronchus-Breast) at a simultaneous 5% level and those which do overlap are not statistically significant (e.g. Ovary-Breast). Each interval estimates the magnitude of the corresponding difference (at a 95% simultaneous confidence level). The magnitude of the individual means is absent from this display.

This and other multiple comparison plots are critically examined by Hsu and Peruggia (1994) who also introduce an interactive "mean-mean" scatterplot designed to address the shortcomings of the existing plots. Figure 7, shows the Heiberger and Holland (2006) static version of the Hsu and Peruggia (1994), or mean-mean, multiple comparison plot.

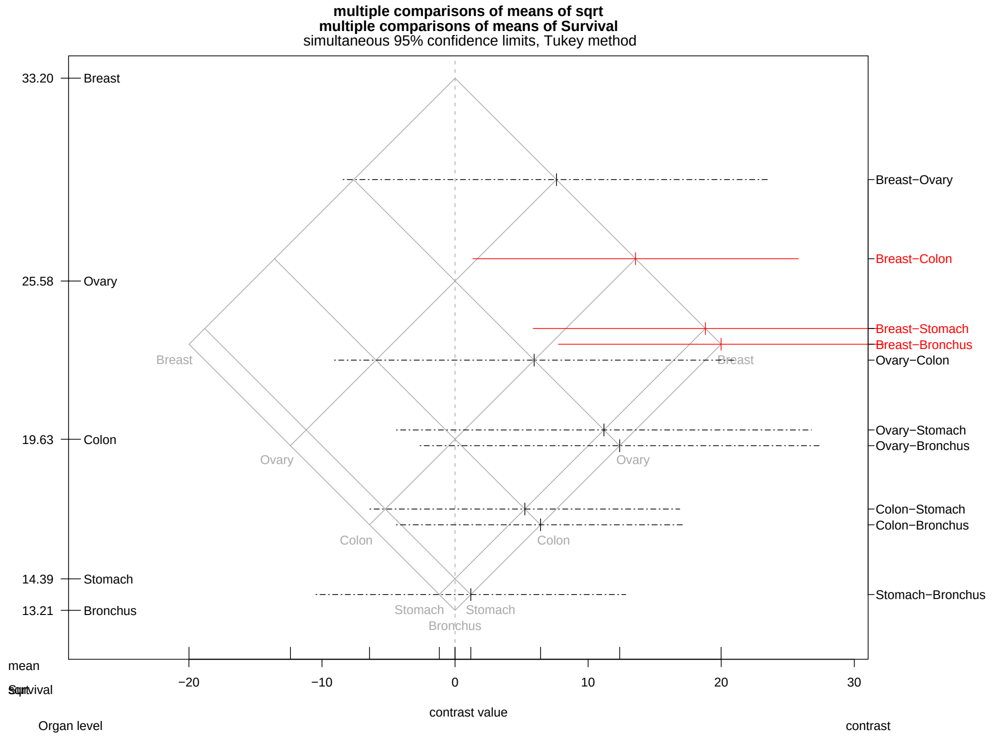


Figure 7: Mean-mean multiple comparison plot: 95% simultaneous confidence intervals, from Heiberger's HH package in R. See Heiberger and Holland (2006). Significant differences are shown as a solid red interval. The Hsu and Peruggia (1994) original is an interactive and colour coded version of this, without the left axis.

As with the standard plot, a vertical zero line is used for the simultaneous confidence intervals. For example, the Breast-Colon is found to be significant and the Breast-Ovary is not. Unlike the standard plot, the intervals are located vertically by the average of the two means being differenced. The rotated square in the background preserves the proportional distances between the different cancer types along the ±45 degree lines so that the size of each mean effect can be determined from the grid (along the ±45 degree lines). Grid locations can also be used to identify the mean difference of each interval. In the Hsu and Peruggia (1994) original, different interval are highlighted to different effect by mouse interaction with the plot.

In the static version, Heiberger and Holland (2006) use colour to further distinguish the significant pairwise comparisons (via red solid lines) and add the leftmost axis showing the levels of each mean. They also show how simultaneous confidence intervals for contrasts of these means might also be located vertically.

Some drawbacks include the possibility of overstriking intervals or grid lines or both whenever pairwise averages or original averages or both are identical (or nearly so). For example, had (Breast + Bronchus) equalled (Ovary + Colon) the intervals for the contrasts (Breast - Bronchus) and (Ovary - Colon) would have been overlaid. Similarly, if the average for Bronchus equalled the average for Stomach, the corresponding grid lines would have been indistinguishable and intervals involving these two and any other cancer, for example (Breast-Stomach) and (Breast - Bronchus), would also be overlaid (as is nearly the case in Figure 7). When overlaying occurs, Heiberger and Holland (2006) also show standard multiple comparison plots as in Figure 6 to display the overstruck intervals, which they call “tiebreaker plots” in this context.

It is also not clear that the information added by the background grid (originally used to motivate and construct the plot) merits the amount of ink it is given. Much of it is redundant given the left axis and the right labels. In the static version the grid can be used to help locate all comparisons which involve any given cancer (i.e. one versus each of the others comparison) by following the grid line of that cancer only (e.g. follow the “Ovary” grid lines to collect the relevant (Ovary - Other) intervals). The cost, however, is a visually more complicated plot, one which might appear needlessly mysterious to many viewers.

An important feature of this plot is that it shows the sample means themselves in addition to their differences. In most applications, having identified the significant differences one is interested in the actual size of each effect being compared. Indeed, we would argue that a comparison of the entire distribution of each group, not just their group means, would be highly desirable.

3.1.1 Boxplots with pairwise testing

Sample distributions can be displayed as histograms, boxplots, density estimates, and so on, which in turn can be compared along a common scale in a variety of ways: possibly overlaid (e.g. densities), or placed back to back (e.g. histograms, densities) or simply laid out side by side (e.g. boxplots, histograms, density estimates). Here we will use a boxplot for the distribution of each group and lay them out side by side to facilitate their pairwise comparison.

Figure 8 shows variable width boxplots of the (square-root transformed) survival times for each cancer type. The left axis gives the values for the boxplots and the horizontal axis

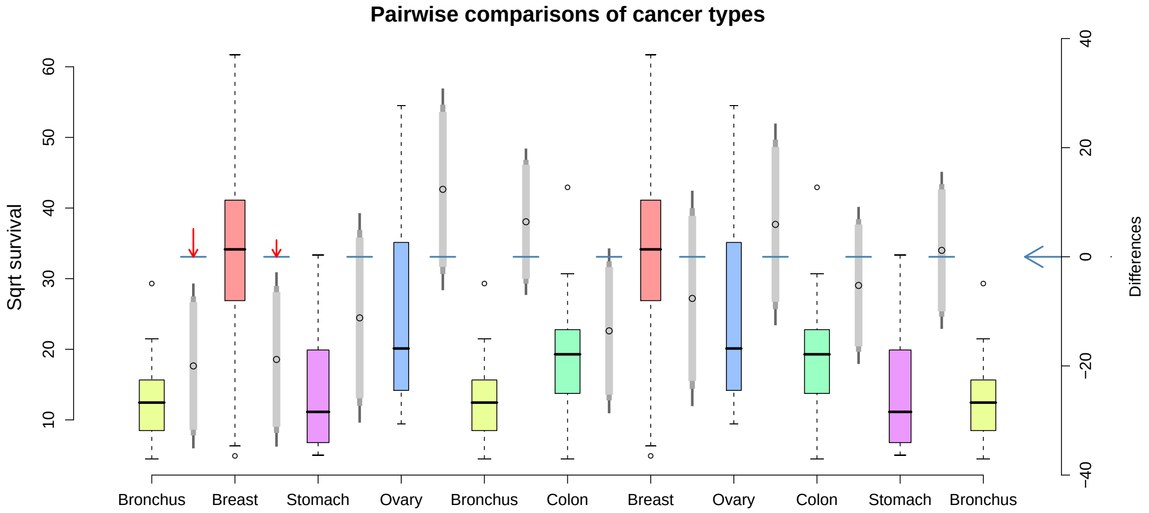


Figure 8: Boxplots and pairwise comparisons of vitamin-C treated cancer patients. The left axis and boxplots refer to square-root transformed survival times, the right axis and gray scale vertical bars refer to confidence intervals for pairwise differences of means. Red lines indicate comparisons significantly different from 0.

the label for each boxplot.

The cancer types (with their boxplots) are repeated along the horizontal axis in such a way that every cancer type appears directly beside every other cancer type exactly once. From left to right, this is an eulerian tour of K_5, where each cancer type corresponds to a node. Because every pair appears together, a fairly rich comparison of survival distributions can be made via the boxplots (e.g.. location, spread, quartiles, symmetry, tail weight, sample size, outliers, etc.). Each cancer type's boxplot is uniquely coloured to facilitate directed comparisons. For example if interest lies primarily in comparing the survival distribution of Ovary cancer to that of the others, simply look for each occurrence of Ovary's thin light blue boxplot and compare it with the boxplots on either side.

Between each pair of boxplots is a gray vertical strip. Each strip is a Tukey HSD confidence interval for the difference in means between the distributions on either side of it, each circle indicating the point estimate of that difference. All confidence interval values can be read from the vertical axis of differences at the right of the plot. Just as the boxplots are the nodes of K_5, the ten gray confidence intervals between them are the edges of K_5. Moving from left to right, from boxplot to confidence strip to boxplot to confidence strip and so on, is an eulerian tour traversal of K_5 from node to edge to node to edge, respectively.

Design features of this plot are chosen to help the user switch visual focus between the mutely coloured boxplots and the gray confidence intervals as need be. This is much like the "layering" of information, simple examples of which have been described by Tufte (1991). For example the right axis is for the confidence intervals and a blue arrow from this axis anchors a horizontal dashed blue line across the plot from its zero. The dashes of this line appear only across the space between boxplots which is reserved for the confidence intervals – the line never interferes with the boxplots themselves.

As with the other multiple comparison plots, inference is had by determining whether the zero line cuts across a confidence interval. If it does (e.g. between Stomach and Bronchus at the right) that difference is not found to be statistically significantly different from zero. Conversely, if it fails to cut through a confidence interval (e.g. between Bronchus and Breast at the left side of the plot) then the difference is significantly different from zero. When this occurs, a vertical red arrow is drawn pointing towards the confidence interval (and on the opposite side of the horizontal line) to draw attention to the interval. Moreover, the greater the length of the arrow is, the greater is its significance (i.e. the smaller its "p-value").

Note that unlike the previous plots for this data only two differences are seen to be significantly different from zero, namely (Breast - Bronchus) and (Stomach - Bronchus). The reason (Colon - Breast) does not show up here is that this plot shows confidence intervals for several levels simultaneously and the largest value here is 99% not 95% as in the other plots.

Careful examination of the vertical confidence intervals of Figure 8 will reveal that they progressively narrow and become a darker shade of gray at the ends. In the figure each interval has three widths and three shades of gray corresponding to three confidence levels: 90%, 95% and 99%. A close look at the confidence interval between the Colon and Breast boxplots shows that the horizontal zero line cuts through the 99% confidence interval, but not the 95%. The difference is significant at the 5% level just as in the other plots but, as this plot indicates, is not significant at the 1% level. Multiple confidence levels are user determined parameters used to produce the plot.

Note also that significant differences seem to appear mostly on the left side of this plot. This is had by attaching a weight to each edge of K_m^e (here K_5) and applying a greedy algorithm which selects the lowest weight edge from those available at each step. To produce Figure 8, we attached the appropriate significance level to each graph edge as its weight. A different choice of weights could produce a different eulerian tour and hence ordering.

These plots could be constructed for any contrasts (boxplot nodes) and any choice of multiple comparison confidence intervals (edges). The boxplots themselves could even be replaced by some other univariate display which highlighted other distributional features.

By carefully ordering the nodes, a relatively simple yet highly informative plot for multiple comparisons has been constructed.

3.2 Interaction plots

Interaction plots are used to explore the presence of interactions between two factors. Figure 9 shows an interaction plot for the survival time of 48 rats, each given one of four treatments A, B, C, or D and one of three poisons P1, P2, or P3 (data from Box and Cox, 1964). The average response is profiled for each poison across the treatments in the standard default order of ABCD. Interaction is detected as a lack of parallelism in the profiles.

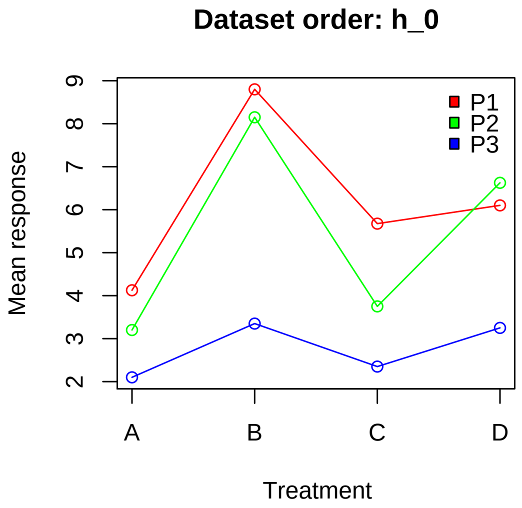


Figure 9: Interaction plots of Rat data. Plot shows are average responses by poison type with the 4 treatments in standard ABCD order.

The profiles are similar in shape, indicating relatively strong main effects – e.g. treatment B produces longer survival times than A, whatever the poison. However, the eye is drawn to possible interactions involving P1 and P2 with C and D where these profiles cross, but also to no interaction of P1 and P2 with A and B in the long nearly parallel line segments from A to B. These profiles invite overall comparison as well as pairwise comparison, each of which can be affected by the ordering of the treatments along the horizontal axis.

The order ABCD is a hamiltonian path (say h_0) on the complete graph K_4 where each node is a treatment. Any other hamiltonian would produce a different ordering and a different looking interaction plot. Figure 10 shows two such hamiltonians, h_1 and h_2 say, that together form a hamiltonian path decomposition of K_4.

An interaction plot based on any one of these hamiltonians might be interpreted differently. For example, in h_1 of Figure 10, the P1 and P2 profiles cross twice and so the visual impression of interaction could be taken to be somewhat stronger than in h_0 of Figure 9. The double crossing occurs because the profile difference changes sign at treatment D which has moved from position 4 to position 2 in the ordering. The second hamiltonian h_2 of Figure 10 might also give a different impression. There, the P1 and P2 profiles exhibit strong zig-zag patterns and with the long line segments connecting treatments B and A the overall impression is of parallelism, while by contrast the P3 profile is quite flat.

An interaction plot using a full hamiltonian decomposition as in Figure 10 ensures that the line segments can be compared for every pair of treatments in a single display and so is less susceptible to interpretation based on a chance ordering of treatments. Alternatively,

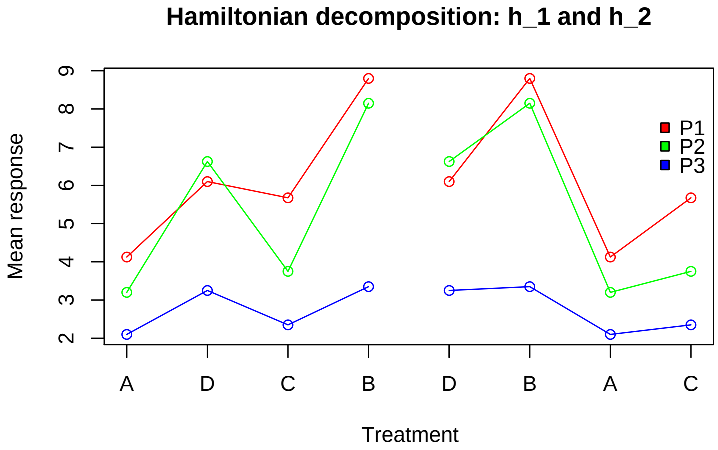


Figure 10: An interaction plot via hamiltonian decomposition (into paths since the number of treatments is even). Every pair of treatments adjacently exactly once.

Figure 11 shows an interaction plot following an eulerian path of the treatments. The eulerian in this case contains back to back hamiltonians, though this will not be the case in general. Also, because the K_4 is of even order, the "eulerian" has had to repeat one of the edges (viz. D-B, or B-D) so is only truly eulerian on the graph with this extra edge. Repeating an edge may give it undue influence in the assessment of interaction. So the hamiltonian path decomposition of Figure 10 would be typically preferable.

Experiments have shown (e.g. Rochlin, 1955) that parallelism is more easily assessed in line plots which are nearly horizontal. To this end, Figure 12 shows a hamiltonian decomposition interaction plot, but with the average profile $(\bar{y}_A, \bar{y}_B, \bar{y}_C, \bar{y}_D)$ removed. This removes the overall treatment effect and magnifies the profile differences. Now the P1 and P2 profiles are better separated and the impression of interaction is stronger that before. The P3 profile clearly has a different shape to the other two, largely because of its low survival times for treatment B. Subtracting the average profile generally reduces the tilt of the line segments, allowing vertical comparisons to be more easily made.

Alternatively, the work of Cleveland and McGill (1984) suggests that parallelism of two curves might be assessed most easily from their difference. In Figure 13, the profile differences are plotted directly. In the ideal no-interaction case, the three difference profiles are flat and and this should be easier to detect than parallelism in tilted lines (Rochlin 1955). Again, the presence of interaction is fairly clear and greatest between P2 and P3. While plotting differences has the benefit of offering a simpler visual task, the tradeoff is that for factors with more than 5 or 6 levels the number of difference profiles may be too large to be easily be distinguished by colour. Even then, the no-interaction pattern of

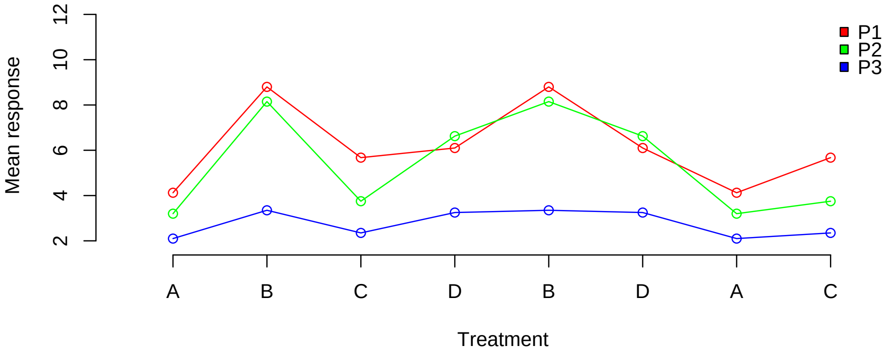


Figure 11: An interaction plot via an eulerian path.

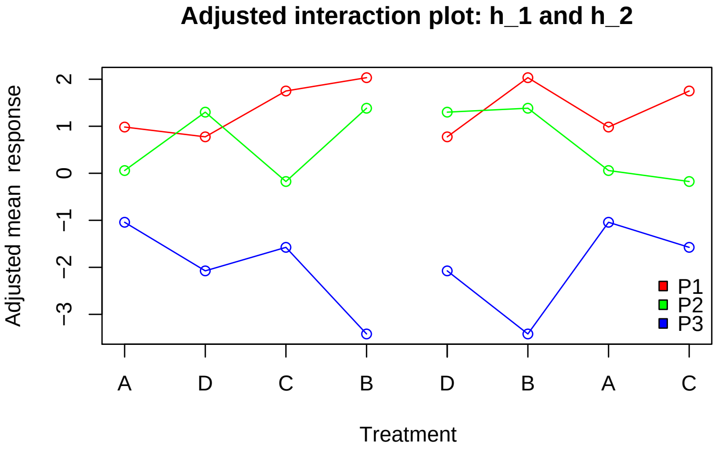


Figure 12: Interaction plots of Rat data with treatment adjusted responses, via the hamiltonian path decomposition as in Figure 10.

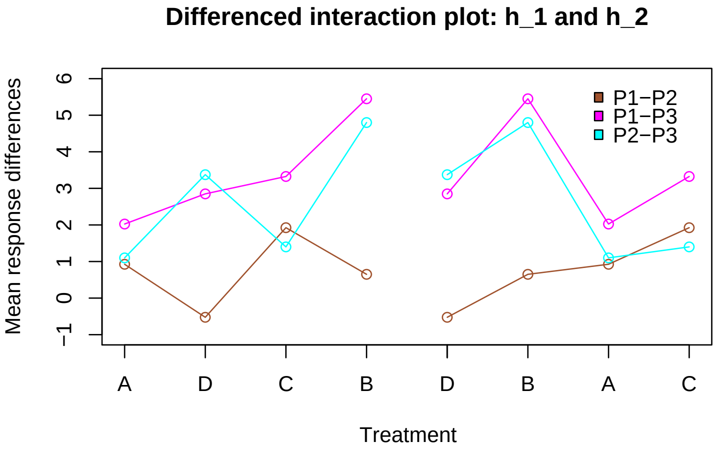


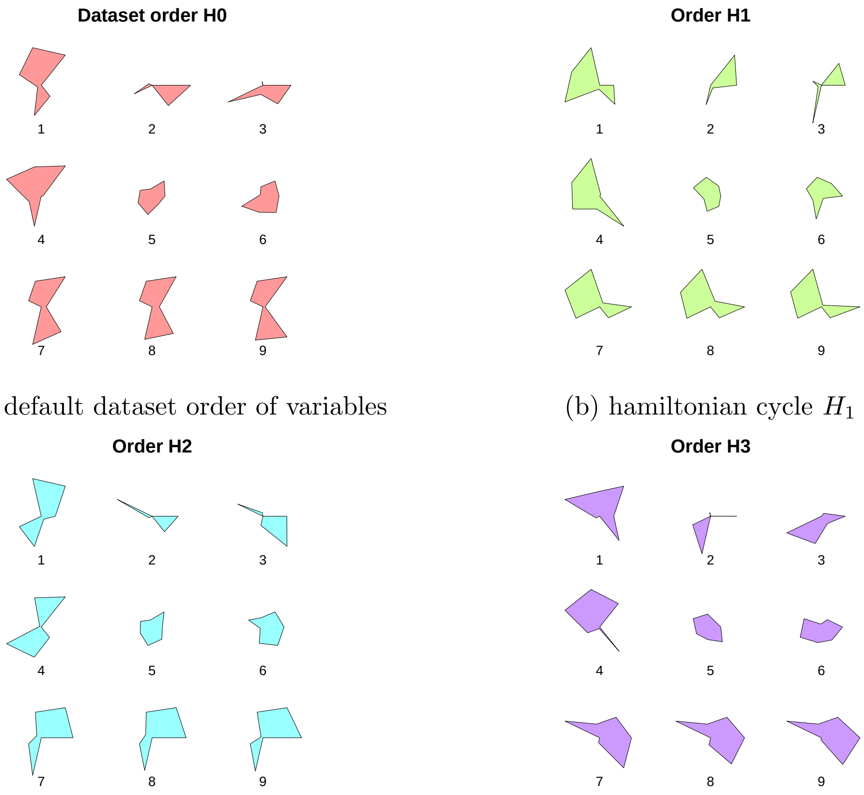
Figure 13: Profile differences for the Rat data via the hamiltonian path decomposition as in Figure 10.

horizontal lines will be easily detected.

3.3 Star glyphs

In a star glyph display, variables are assigned to equispaced radii, (scaled) observations are plotted on each radius, and lines are drawn connecting the case values. High-dimensional features of the data are quickly compared across cases by comparing glyphs, individual radii and overall shape – an example of what Tufte (1991) describes as the use of "small multiples". Suppressing the radial rays from the display, focus is on comparison of shapes rather than of variable values across cases – a distinction which has been usefully described as that between integrable and separable dimensions by Wilkinson (2005, p. 269). Arranged in an array of glyphs corresponding to the cases in a dataset, such a star glyph display invites visual clustering by shape. Figure 14 shows four star glyph displays of a subset of car models from the `mtcars` dataset found in the R datasets package. Here variables are assigned to radial axes starting from the 3 o'clock position and moving counter-clockwise thereafter. Each display uses the same seven variables, they differ only in the assignment of variables to the axes. Imagining the variables as nodes of a complete graph, the order of assignment of the seven variables to the radial arms of the star is equivalent to the choice of a hamiltonian cycle from the complete graph K_7. Not surprisingly, the shapes of the star glyphs vary considerably from one ordering or hamiltonian to another.

Let's attempt to use the first ordering, the hamiltonian cycle H_0, to cluster the cars. The glyphs for models 7, 8 and 9 look very similar to each other, and quite similar to the glyph for model 1. Models 5 and 6 are both represented by medium-sized blobs which look



(a) H_0: default dataset order of variables

(b) hamiltonian cycle H_1

(c) hamiltonian cycle H_2

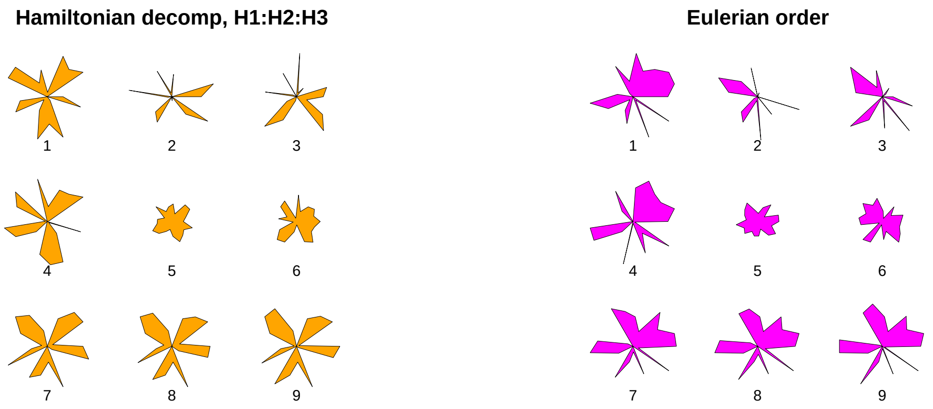
(d) hamiltonian cycle H_3

Figure 14: Star plots of 9 models from the `mtcars` data using different hamiltonian cycles. H_1, H_2, and H_3 together form a hamiltonian decomposition.

roughly similar. The model 4 glyph looks different to all others.

Other orderings tell a different story; in H_1 model 4 looks like it belongs to the $\{1, 7, 8, 9\}$ cluster, while in H_2 and H_3 we have two separate clusters consisting of models $\{1, 4\}$ and $\{7, 8, 9\}$. Clearly visual clustering based on star glyph displays is order dependent.

Intuitively, if we replaced the sequence of variables used in the star glyph by a longer sequence where all variables appear adjacently, we should remove some of the dependence on variable order. Figure 15 shows two different eulerian sequences of variables. The first plot uses the sequence $H_1 : H_2 : H_3$, which is a concatenation of the sequences appearing in Figure 15(b), (c) and (d), and was constructed using weighted hamiltonians via WHam (Algorithm 3). The second plot uses a weighted eulerian sequence constructed via GrEul (Algorithm 2) which favours high correlation pairs of variables appearing early on in the

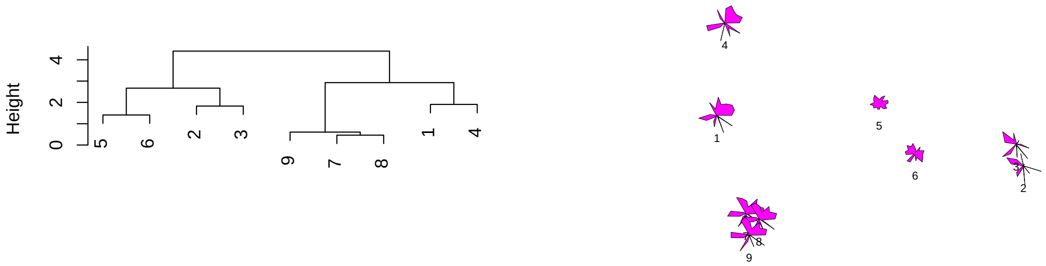


(a) Order by hamiltonian decomposition

(b) Correlation ordered eulerian

Figure 15: Star plots of 9 models from the `mtcars` data.

sequence. We notice that the shapes of the star glyphs vary less with the sequence used that in Figure 14. This occurs because in Figure 14 the star vertices are rearranged in each hamiltonian, whereas in Figure 15, it is the star edges that are rearranged. Visual clustering based on either of the sequences shown in Figure 15(a) and (b) gives the same results; it appears there are four clusters, made up of models $\{1, 4\}$, $\{2, 3\}$, $\{5, 6\}$ and $\{7, 8, 9\}$. These findings are verified by the dendrogram (shown in Figure 16 a) obtained from average



(a) Dendrogram: Average linkage

(b) Glyph scatterplot on first two principal components

Figure 16: Clustering the cars.

link clustering (single and complete linkage dendrograms were identical) and reinforced by plotting the symbols in the space of the first two principal components (correlation matrix) as shown in Figure 16(b). This experiment suggests eulerian variable sequences on star glyphs for reliable visual clustering.

3.4 Parallel Coordinate Plots

Parallel coordinate displays (Inselberg 1985, Wegman 1990) are multivariate data displays where n variables are assigned to parallel, equispaced axes, observations are plotted on each axis and lines are drawn connecting observations belonging to each case. These displays are useful for detecting clusters, outliers and correlation between pairs of variables. Once again, choosing a variable ordering amounts to selecting a hamiltonian path on the complete graph with the variables as nodes. However, as demonstrated by Wegman(1990), there are strong reasons for displaying all pairwise variable relationships in a parallel coordinate display, not just the $n-1$ pairwise relationships determined by a particular choice of hamiltonian.

Here we use parallel coordinate displays to revisit the `mtcars` data. Figure 17 shows

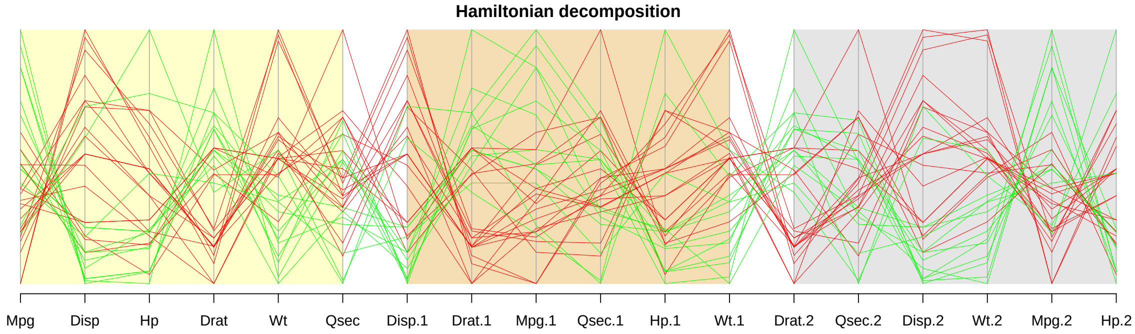


Figure 17: Parallel coordinate plots of the `mtcars` data. This shows a hamiltonian decomposition, panel colors distinguish the three hamiltonian paths. Line color shows transmission type.

a parallel coordinate display of six performance measures. The display has three sections, each highlighting a different hamiltonian path, which together constitute a hamiltonian decomposition on K_6^e. The first six axes (the yellow section) show the variables in the order in which they are listed in the dataset, corresponding to the hamiltonian 123456. Here we see that the first two variables, Mpg and Disp, are negatively correlated, but the association between the first and third variables, Mpg and Hp is not so obvious until we look at the last panel in the grey section and discover that they are also negatively correlated. The dataset has a cluster of unusually heavy cars which, we discover from the Wt-Disp panel in the grey section, also have high displacement.

The main argument against all-pairs parallel coordinate displays is that the number of panels (i.e., the number of edges in the eulerian on K_n^e) is $O(n^2)$. (From the discussion in Section 2.2, the number of edges is $\binom{n}{2}$ when n is odd, and $\binom{n}{2} + (n-2)/2$ when n is even.) Figure 18 shows an all-pairs parallel coordinate display for the `sleep` data which has

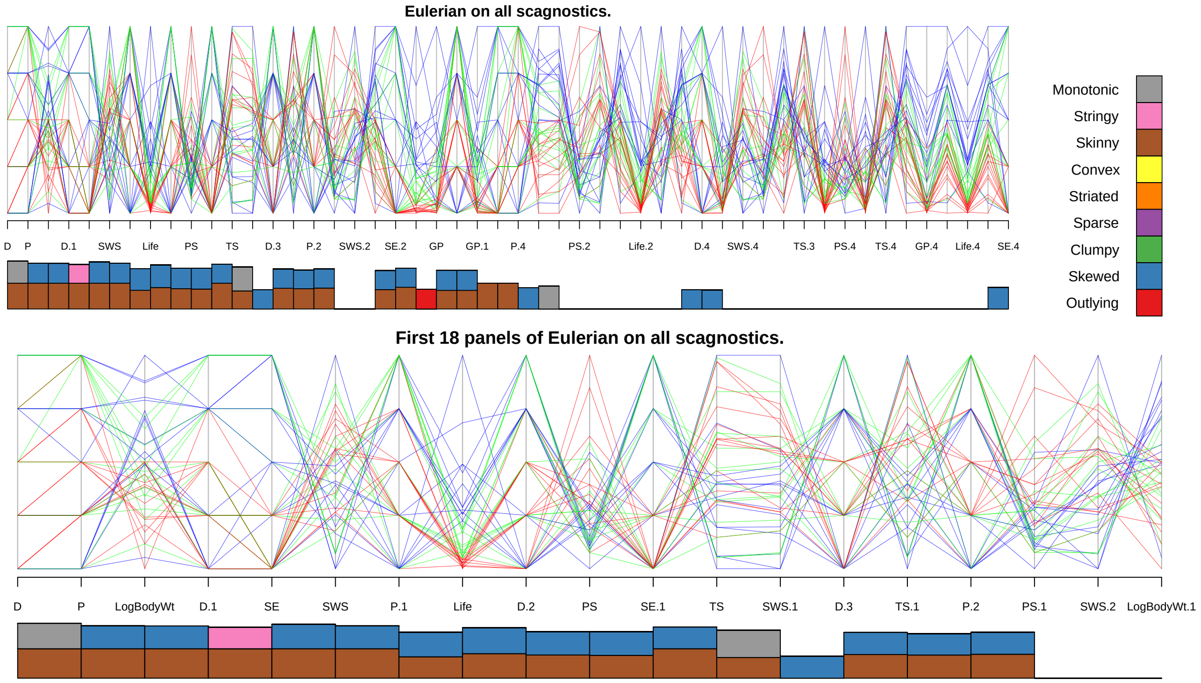


Figure 18: Parallel coordinate plots of the `sleep` data. The top display shows an eulerian with panels ordered by the total of scagnostic values. The lower display shows the first 18 panels only. Line colors are assigned using the life expectancy variable. The barcharts show scagnostic index levels for each panel.

$n = 10$ measurements on 62 mammal species (Allison and Cicchetti, 1976). The eulerian has 49 edges and it is difficult to see patterns on a standard computer screen or page. To ameliorate this, we use the GrEul algorithm (Algorithm 2 of Section 4.1.1) to construct an eulerian where “interesting” panels appear early on in the sequence. The lower parallel coordinate plot of Figure 18 zooms in on the first 18 panels, which in some sense is the most interesting portion.

Here we use scagnostic indices to measure the “interestingness” of each panel. These indices were designed by Tukey and Tukey (1985), revisted by Wilkinson et al (2005) and recently implemented in the R scagnostics package (Hofmann et at, 2007). Scagnostics evaluate different features of a bivariate scatterplot, such as convexity and monotonicity. (Possibly other indices could be developed focusing on characteristics of a parallel coordinate display.) In Figure 18 the accompanying barcharts show scagnostic indices of each panel (note index values less than 0.7 are ignored). The overall interestingness of a panel is then measured by the sum of the scagnostic indices, so that the first panel (D versus P) is the most interesting overall, having large values of both the monotonicity and skinniness indices.

Note that about half of the panels exhibit either considerable skewness or skinniness index or both. None of the panels score highly on the convex, striated, sparse or clumpy indices. The lower display zooms in on the first several panels; these few important panels permit patterns and relationships to be more easily seen, with the scagnostic barchart as a guide to interpretation. (Recall from the results given in Section 2.2, when n is even, construction of an eulerian requires that $n/2-1$ extra edges are added to K_n.) Coupled with ordering and zooming, eulerian parallel coordinate displays can be practical and informative for datasets with n=10 variables and more.

Alternatively, rather than zoom in on interesting portions of eulerian parallel coordinate displays we could focus instead on interesting hamiltonians, or even on several from a hamiltonian decomposition. For example, parallel coordinate plots are often used to find groups of data points in high dimensions, and for outliers. The top left display of Figure 19 shows the best hamiltonian, that which maximizes the sum of the indices "outlying", "clumpy", and "sparse". The result is a parallel coordinate plot tailored to find clusters and outliers in this data. The first few panels are dominated by clumpy and sparse, the last few by outlyingness. Striated data is a particular form of clustering and so the best

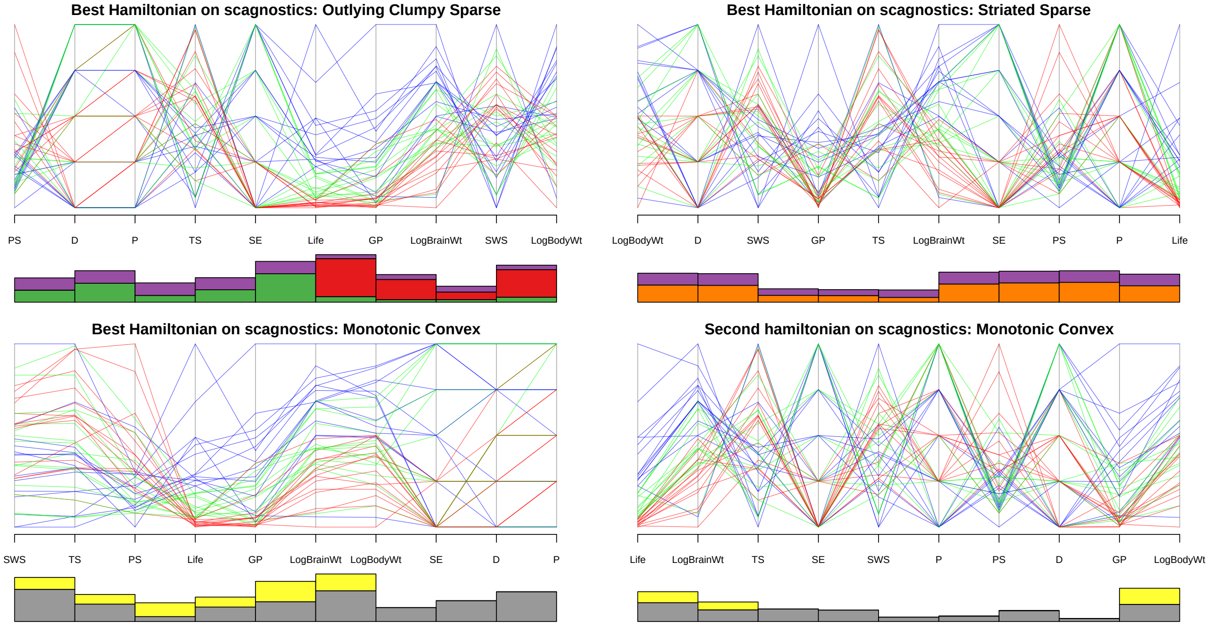


Figure 19: Hamiltonian parallel coordinate plots of the `sleep` data. Line colors are assigned using the life expectancy variable. The barcharts show scagnostic index levels for each panel (all values larger than 0 appear, and on a common scale).

for “striated” + “sparse” is shown in the top right display of Figure 19.

Similarly, parallel coordinate plots are often used for examining correlational structure; the bottom row of Figure 19 shows the first two hamiltonians from the WHam algorithm (Algorithm 3 of Section 4.2.3) for maximizing convexity and monotonicity (together they should indicate pairs of variables which are correlated, i.e. monotonic, and whose scatterplots are convex – together, ideal conditions for parallel lines and strong crossings to appear). While neither of these latter two hamiltonians share any pair of adjacent variables (being from the same decomposition), they might share adjacent pairs with other hamiltonian displays, for example the D-P pair of the two leftmost displays.

These two methods of zooming in on selected interesting subsets of all-pairs parallel coordinate plots allows the analyst to focus on different features of the data. If these were also dynamically linked so that brushing could occur across these displays, they would constitute a very powerful exploratory data tool indeed.

4 Graph traversal algorithms

In this section, we describe algorithms for constructing various graph traversals which were used in the applications to statistical graphics of Section 3. First we present the algorithms for constructing eulerian paths. Specifically, we recall the standard algorithm due to Hierholzer (1873) and modify it for weighted graphs. We then move on to constructions for hamiltonian decompositions, specifically on complete graphs. Finally, we present a new algorithm which is useful for building hamiltonians on complete graphs that are weighted.

4.1 Constructing eulerian paths

4.1.1 Hierholzer’s algorithm

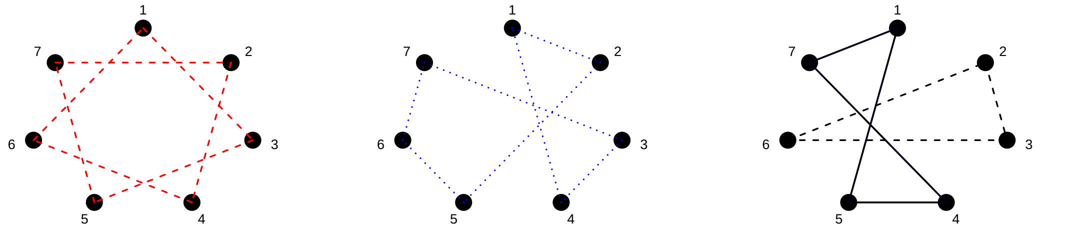
Algorithm 1 (Hierholzer 1873) constructs eulerian tours; another well known algorithm is due to Fleury (1883). Recall from Section 2.2 that eulerian tours exist for even graphs, but with a minor adaptation Hierholzer’s algorithm constructs an open eulerian path or trail for graphs with exactly two odd nodes. Fleury’s algorithm is essentially the same (e.g. see Fabràga and Fiol, 2004) and could be adapted analogously.

Hierholzer’s method has many arbitrary choices – the choice of the vertex v in line 1 and at each step of the path constructed in lines 2 and 5, the choice of w in line 4, and if w appears more than once in T at step 6, the choice of which occurrence of w in T to use to splice path D into T (though the most recent is suggested).

Figure 20 shows how an application of Hierholzer’s method might create an eulerian tour for K_7. Starting at node 1 the selection of edges is such as to produce the hamiltonian cycle 13572461 of Figure 20(a), followed by a second hamiltonian cycle 12567341 of Figure 20(b), and finally by the short cycle 15471 of Figure 20(c). At this point, node 1 has no further unused edges and $T = 1357246\ 1\ 256734\ 1\ 547\ 1$. Path D (of Algorithm 1 line 5)

Algorithm 1 Hierholzer 1873 (adapted to find an eulerian tour or open eulerian path)

```
Require: A connected graph $G$ that is even or that has exactly two odd vertices.
1: Choose a vertex $v$. If $G$ is even, $v$ can be any vertex, otherwise $v$ is one of the two odd vertices.
2: Starting at $v$ construct a path $T$ in $G$, stopping when a vertex is reached without an unused edge.
3: while there are edges of $G$ not already in path $T$ do
4:     Choose any vertex $w$ in $T$ that is incident on an unused edge.
5:     Starting at $w$, construct a path $D$ of unused edges stopping when a node is reached without any unused edges.
6:     Enlarge $T$ by splicing path $D$ into $T$ at vertex $w$.
7: end while
8: return $T$
```



(a) First hamiltonian cycle (b) Second hamiltonian cycle (c) Two non-hamiltonian cycles

Figure 20: An application of Hierholzer's method to K_7 which happens to follow one hamiltonian cycle after another.

is the dashed cycle 2362 of Figure 20(c), which line 6 of the algorithm allows to be spliced into T at node 2. The resulting eulerian tour can be either 1357 2362 46125673415471 or 13572461 2362 5673415471.

Hierholzer's method applies to the graph K_n^e for all n. When $n = 2m + 1$, $K_{2m+1}^e = K_{2m+1}$ is even and it will yield an eulerian tour. When $n = 2m$, K_{2m}^e is an augmented version of K_{2m}, adding $(m - 1)$ extra edges to produce a graph with exactly two odd nodes, and the result is an open eulerian path.

4.1.2 Eulerians on weighted graphs

If the graph G is a weighted graph (e.g. the weights represent some kind of distance or other dissimilarity measure between the vertices), we might prefer an ordered eulerian T with low weight edges occuring early in the sequence and with weights tending to increase as the sequence progresses. As the discussion in Section 2.2.1 illustrates, the number of distinct eulerian tours is typically immense, and finding the overall "best" tour is not a practical option. However, a greedy algorithm which attempts this is easily had by exploiting the arbitrary choices available in Hierholzer's method. The necessary minor modifications of Algorithm 1 are given below as the greedy eulerian or GrEul of Algorithm 2. Note that the choice of starting vertex is limited to the two odd vertices when constructing eulerian paths, but when constructing K_{2m}^e one can always ensure that a particular start vertex v has odd degree.

Algorithm 2 GrEul: Greedy Eulerian.

```
Require: A connected graph $G$ that is even or that has exactly two odd vertices.
1: Choose a starting vertex $v$ from one of the odd vertices connected by the lowest weight
   edge, using the next lowest weight edge in their vertex sets to decide between them.
2: Starting at $v$ construct a path $T$ in $G$, always moving to the lowest weight unused edge,
   stopping when a vertex is reached without an unused edge.
3: while there are edges of $G$ not already in path $T$ do
4:    Choose the last vertex $w$ in $T$ that is incident on an unused edge.
5:    Starting at $w$, construct a path $D$ of unused edges, always moving to the lowest
      weight unused edge and stopping when a node is reached without any unused edges.
6:    Enlarge $T$ by splicing path $D$ into $T$ at vertex $w$.
7: end while
8: return $T$
```

For example, suppose the edge weight for the edge connecting two vertices i and j is min(i, j). Then the construction of an ordered eulerian tour on K_5 proceeds as follows. Choose $v = 1$. Line 2 of the modified algorithm will produce $T = 12314251$. Line 5 starts at vertex 5, and builds $D = 5345$, which then replaces the '5' in T, yielding a tour of 12314253451.

We note that constructing eulerian trails is an $O(|E|)$ task, where $|E|$ is the size of the graph, and so the algorithm given above constructs trails on K_n^e in $O(n^2)$ time. The cost associated with constructing an ordered eulerian must include the cost of an edge sort at each vertex, and so has overall order on K_n^e of $O(n^2 \log n)$.

4.2 Lucas-Walecki hamiltonian decompositions for K_n^e

While the adapted Hierholzer method will produce an eulerian for any K_n^e, the eulerian need not be hamiltonian decomposable. Even if node choices were restricted so that the algorithm first constructed one hamiltonian followed by another, the result need not be a hamiltonian decomposition. Figure 20 shows just such a situation.

Fortunately, the special structure of K_n^e can be exploited to write down explicit formulas for eulerians and Hierholzer's method need not be used. This method has the added advantage that for $n = 2m$, the eulerian is composed of m hamiltonian paths, while for odd $n = 2m+1$, it is composed of m hamiltonian cycles. The constructions given here have been attributed to Walecki by Lucas (1892), and are sometimes described as *Lucas-Walecki* constructions (Bailey et al, 2003). A disadvantage is that these necessarily generate only a single class of isomorphic decompositions which, though potentially huge, cannot include those produced from possibly thousands of other hamiltonian decomposition classes that are non-isomorphic to this one.

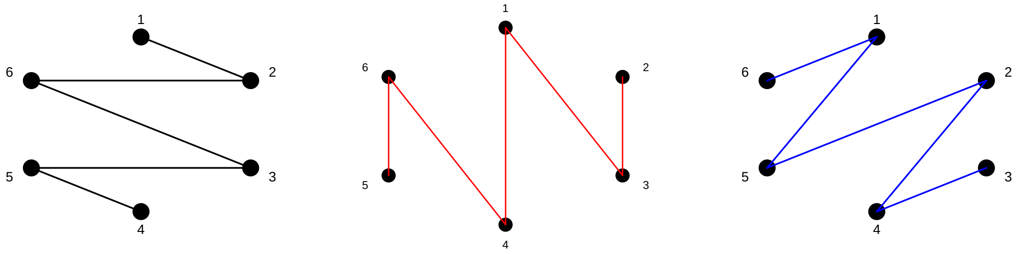
4.2.1 Hamiltonian decompositions, n even

As before, let $n = 2m$ and define

$$\begin{array}{rcl} H[1,1] & = & 0 \\ H[1,j] & = & H[1,j-1] + (-1)^j(j-1) \pmod{n}, \;\; j = 2,\ldots,n, \\ H[k,j] & = & H[k-1,j] + 1 \pmod{n}, \;\; k = 2,\ldots,m \text{ and } \;\; j = 2,\ldots,n,. \end{array}$$

Finally, increase each element of H by 1, and set $T_{n(k-1)+j} = H[k,j], \quad j = 1,\ldots,n, \;\; k = 1,\ldots,m$. That is, form T by listing the elements of H row-wise. The resulting path T is an eulerian trail on K^e_{2m}.

When the vertices of K_{2m} are arranged clockwise around a circle, the first row of H visits all vertices in a zig-zag pattern. This is shown for K_6 in Figure 21(a). Each successive row of H follows another zig-zag starting one position clockwise away from the start of the previous row.



(a) $h_{41} = 126354$

(b) $h_{42} = 231465$

(c) $h_{43} = 342516$

Figure 21: A hamiltonian path decomposition $H_4 = h_{41} : h_{42} : h_{43}$ of K_6.

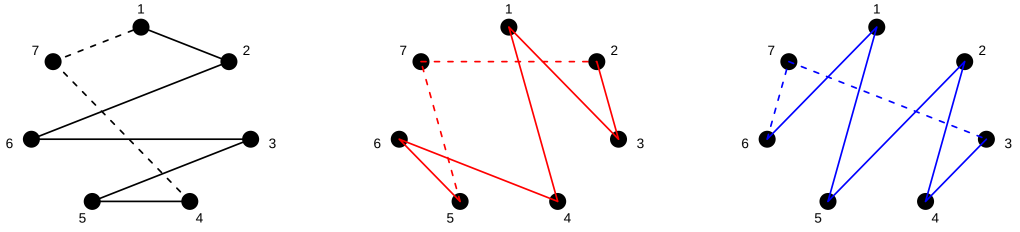
The rows of H form a hamiltonian path decomposition of K_{2m} and consequently every pair of vertices appears consecutively in exactly one of the rows. When the rows are glued together to form the T-sequence, the edge pairs contributed by $H[i,n]$ and $H[i+1,1]$ are duplicates. These are the edges j $(j+m-1)$(i.e. $e_{j,(j+m-1)}$) for $j = 2,\ldots,m$. The resulting

T-sequence, having duplicate edges, is a decomposition of K_{2m}^e into m hamiltonian cycles, where in this case K_{2m}^e is formed from K_{2m} by adding the additional edges between nodes j and $j+m-1$ for all $j = 2, \ldots, m$.

Figure 21 illustrates the process for K_6^e. Each panel shows a hamiltonian path from a row of H, and these paths are joined up to give $T = 126354\ 231465\ 342516$. In this sequence the edges 4 2 (or 24) and 5 3 (or 35) are duplicates. Note also that this decomposition is isomorphic to the decomposition H_3 for K_6 given in Figure 4.

Wegman (1990) used this Lucas-Walecki construction to list m different permutations of $2m$ variables where each pair of variable adjacencies appears exactly once. Following Wegman (1990), we will use the more evocative name, "zig-zag method", to refer to this construction. For $n = 2m+1$ the zig-zag method lists m permutations of variables, where each pair of variables appears adjacently at least once, but with some pairs appearing twice. The result will obviously not be a hamiltonian decomposition.

4.2.2 Hamiltonian decompositions, n odd



(a) $H_{41} = 1263547$ (b) $H_{42} = 1327564$ (c) $H_{43} = 1524376$

Figure 22: A hamiltonian decomposition $H_4 = H_{41} : H_{42} : H_{43}$ of K_7. For K_{2m+1}, the paths for K_{2m} are constructed and then the point $2m+1$ joined to the ends to complete the cycle.

An easy way of generating a hamiltonian decomposition for $n = 2m+1$ uses a minor modification of the zig-zag method just described. Start with the H matrix used in the construction of the path for $2m$ vertices, and create the augmented matrix H^* by prepending a column of n's to H. Row-wise listing the elements of H^* and adding a final n produces an eulerian tour for K_{2m+1}.

Each row of H^* has the form $n, j, \ldots, (m+j)$, so we have inserted the required edges nj at the beginning of each row and $(m+j)n$ at the end, for $j = 1\ldots, m$. The extra n at the end of the T-sequence contributes the edge $(n-1)n$. For example when $n = 7$, we transform the $n = 6$ sequence of 126354 231465 342516 to 7 126354 7 231465 7 342516 as illustrated in Figure 22. Note that this decomposition is isomorphic to the decomposition H_2 for K_7 given in Figure 3.

In general, each row of H^* is a hamiltonian path, and since each row begins with n, we have formed a decomposition of K_{2m+1} into m hamiltonian cycles.

Interestingly, constructions of hamiltonian decompositions on K_{2m+1} have applications in experimental design. Bailey et al (2003) call these decompositions a *round-dance neighbour design*, where an odd number n of objects is arranged in $(n-1)/2$ rings so that each pair of objects are adjacent in exactly one ring. They also give a number of other constructions for such designs, and relate them to Latin and Tuscan squares.

4.2.3 Hamiltonian decompositions on weighted graphs

For weighted graphs our goal is an ordered eulerian T where weights tend to increase as the sequence progresses. Here we will build such paths out of hamiltonians. For a given hamiltonian (path or cycle) decomposition $H = H_1 : H_2 : \cdots : H_m$, it is clear from the discussion of Section 2.1 that since the labelling of vertices is arbitrary, any sequence of vertices can be chosen as the first (or any other) hamiltonian in the hamiltonian decomposition, but then the other hamiltonians in the path must follow the same labelling scheme. The order in which the hamiltonians appear in constructing the eulerian can be permuted and each component path or cycle H_i can be oriented arbitrarily to form the eulerian composed of these hamiltonians.

These operations open up a huge number of possible paths, far too many to attempt to find the overall winner based on some merit measure using edge weights. However some preferences can be made algorithmically. For example, given an eulerian T composed of a hamiltonian decomposition $H = H_1 : H_2 : \cdots : H_m$ (e.g. arrived at by applying the zig-zag algorithm) we could choose to order the hamiltonians within the decomposition those with smaller total edge weight precede those with larger total edge weight. Then within each hamiltonian we could choose to orient the path (or cycle) so that smaller weights tended to appear earlier in the path (cycle) than larger weights (a strict ordering will not likely be possible). If no decomposition is in hand, we could first choose a hamiltonian with smallest total weight out of all possible hamiltonians. This is essentially the travelling salesman problem (tsp) and so typically only an approximate solution is guaranteed. It would be nice to think that we could do this recursively, always getting the next best hamiltonian from the remaining graph, but as the examples of Figures 20 and 5 show, it is possible to produce several disjoint hamiltonians in sequence without arriving finally at a full decomposition. So recursing in this way will only be useful for some number of hamiltonians. If a full decomposition is desired, then we will choose only the first hamiltonian to be 'best' and then apply the zig-zag algorithm from this starting point to ensure that a full decomposition results.

These ideas are put together as the WHam (or weighted hamiltonian) algorithm outlined below as Algorithm 3.

Algorithm 3 WHam: Weighted Hamiltonian Ordered

```
Require: A weighted $K_n^e$.
1: For $H_1$, find the hamiltonian (path for even $n$, cycle for odd $n$) with the smallest total
   weight.
2: Let $C(P)$ be a measure of the tendency for the edge weights in a path $P$ to decrease.
3: Using the criterion $C$, pick the best starting point and path orientation for $H_1$. (For
   open paths, there are only two possible starts, for cycles there are $n$).
4: Apply this node labelling to the other hamiltonians $H_2, \ldots, H_m$ in the sequence.
5: Use criterion $C$ again to find the best orientation for each of $H_2, \ldots, H_m$.
6: Permute $H_2, \ldots, H_m$ in order of increasing total weight, and relabel the hamiltonians.
7: return $T = H_1 : H_2 : \cdots : H_m$.
```

Note that line 1 of Algorithm 3 is essentially the "Travelling Salesman Problem" or TSP. While finding the optimal solution is NP-hard, there are many approximate solutions that work well in practice.

5 Concluding remarks

The appearance and resulting interpretation of many data visualizations depends on the ordering of their components. Our goal is to identify good orderings which reveal the data, make large datasets coherent and encourage data comparisons and so promote graphical excellence (Tufte 1987). We approached the ordering problem using graph traversals, and presented algorithms for constructing hamiltonian decompositions and eulerians, which enumerate all pairwise comparisons in a systematic way. Aside from Wegman (1991) who used one of the constructions given here (see Section 4.2.1) to construct parallel coordinate displays, we know of no other application of these techniques to statistical graphics.

In Section 3 we explored applications of these methods in data visualization, devised a new multiple comparisons display which facilitates easy comparison of treatment groups, constructed improved star glyph displays for better visual clustering, and modified parallel coordinate displays and interaction plots to reveal more data patterns. We have also investigated applications to profile glyphs, with results similar to that for star glyphs, and Andrews' (1972) curves (which also suffer from a variable order effect). More generally, our methods are applicable to any statistical technique or visualization that relies on a particular sequencing of variables, cases or factor levels.

The main drawback to using hamiltonian decompositions and eulerians in constructing data visualizations is that the length of the decomposition or eulerian path is roughly $n^2/2$. Here our solution is to construct hamiltonians and eulerians on weighted graphs and in Section 4 we presented new algorithms designed for this purpose. The resulting visualization can then give prominence to relevant features of the data. In an interactive setting, the user could select an interesting data feature or features and immediately zoom in on a subsequence of the associated weighted eulerian.

In this paper we focused on complete graphs, as these are widely applicable in visualization problems. But we could also envisage visualization applications of incomplete graphs. For example consider the canonical correlation setting where there are two groups of variables, $\{X_i, i = 1, \ldots, a\}$ and $\{Y_j, j = 1, \ldots, b\}$ and we wish to construct a parallel coordinate display where X and Y variables appear adjacently. Here we construct a bipartite graph where edges connect X and Y variables only. If a and b are both even an eulerian exists, and our modified Hierholzer (Algorithm 1) or GrEul (Algorithm 2 for weighted graphs) give a construction. (If a and b are not both even, extra edges must be added to the graph so that only two vertices are odd.)

Finally, the algorithms and new graphical displays introduced here are available as the contributed R package `EulerViz`.

References

Allison, T. and Cicchetti, D. (1976). "Sleep in Mammals: Ecological and Constitutional Correlates", *Science*, 194, pp. 732-734.

Andrews, D.F. (1972). "Plots of High-Dimensional Data" *Biometrics*, 28, pp. 125-136.

Alspach, B, J.-C. Bermond, and D. Sotteau (1990), "Decomposition into cycles I: Hamilton decompositions", in *Cycles and Rays* (eds. G. Hahn, G, Sabidussi, and R.E. Woodrow), Kluwer Academic Publishers, Boston.

Ankerst, M., Berchtold S. and Keim D. A. (1998), "Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data", *Proceedings: IEEE Symposium on Information Visualization*, pp. 52-60.

Bailey, R.A., M.A. Ollis, and D.A. Preece (2003), "Round-dance neighbour designs from terraces", *Discrete Mathematics*, 266, pp. 9-86.

Bates, D. (1997+), "`TukeyHSD`: Tukey's Honest Significant Difference", *The R Project*, http://www.r-project.org

Cameron, E. and L. Pauling (1978),"Supplemental ascorbate in the supportive treatment of cancer: Re-evaluation of prolongation og survival times in terminal human cancer", *Proc. Nat. Acad. Sci., USA*. 75, No. 9, pp. 4538-4542.

Chernoff, H and M. H. Rizvi 1975). "Effect on classification error or random permutations of features in representing multivariate data by faces." *Journal of American Statistical Association*, 70, pp. 548-554.

Cleveland, W.S. and R. McGill, (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods" *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 531-554.

Colbourn, C.J. (1982), "Hamiltonian decompositions of complete graphs", *Ars Combinatoria*, 14, pp. 261-269.

Fabràga, J. and M.A. Fiol (2004),"Connectivity and Traversability", Chapter 4, pp. 193-

339 of *Handbook of Graph Theory* (eds. J.L. Gross and J. Yellen), CRC Press, Boca Raton.

Fleury (1883), "Deux problèmes de géométrie de situation", *Journal de mathématiques élémentaires*, pp. 257-261.

Friendly, M. and Kwan, E. (2003), "Effect Ordering for Data Displays", *Computational Statistics and Data Analysis*, 43, 509-539.

Heiberger, R.M., and P. Holland. (2006), "Mean-mean multiple comparison displays for families of linear contrasts.". *Journal of Computational and Graphical Statistics*, 15, pp. 937-955.

Hierholzer, C. (1873), "Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren". *Math. Annalen*, VI, pp. 30-32.

Hofmann, H., Wilkinson, L., Wickham, H., Temple Lang, D. and A. Anand (2007) "The scagnostics package", http://www.r-project.org.

Hsu, J. and M. Peruggia (1994), "Graphical representation of Tukey's Multiple Comparison Method", *Journal of Computational and Graphical Statistics*, 3, pp. 143-161.

Hurley, C. (2004), "Clustering Visualizations of Multidimensional Data", *Journal of Computational and Graphical Statistics*, vol. 13, (4), pp 788-806, 2004.

Inselberg, A. (1985), "The plane with parallel coordinates", *The Visual Computer*, 1, pp. 69-91.

Lucas, D.E. (1892), *Recréations Mathématiques, Vol. II*, Gauthier Villars, Paris.

McKay, B. D. and R.W. Robinson (1998), "Asymptotic enumeration of eulerian circuits in the complete graph", *Combinatorics, Probability and Computing*, 7, pp. 437-449.

Reiss, M. (1871-3), "Evaluation du nombre de combinaisons desquelle les 28 dés d'un jeu du domino sont susceptibles d'après la règle de ce jeu", *Ann. Mat. Pura. Appl.*,5, pp. 63-120.

Rochlin, A.M. (1955), "The Effect of Tilt on the Visual Perception of Parallelness", *The American Journal of Psychology*, 68, pp. 223-236.

Sloane, N.J.A. (2004), "Sequence A007082" from the *Online Encyclopedia of Integer Sequences*, http://www.research.att.com/∼njas/sequences/.

Tufte, E.R. (1987), *The Visual Display of Quantitative Information*, Graphics Press, CT.

Tufte, E.R. (1991), *Envisioning Information*, Graphics Press, Cheshire, CT.

Wegman, E.J. (1990), "Hyperdimensional data analysis using parallel coordinates", *Journal of the American Statistical Association*, 85, pp. 664-675.

Wilkinson, L. (2005), *The Grammar of Graphics (Second Edition)*, Springer, New York.

Wilkinson, L., Anand, A. and Grossman, R. (2005), "Graph-theoretic scagnostics", *Proceedings of the IEEE Information Visualization 2005*, pp. 157-164.