Graphs as navigational infrastructure for high dimensional data spaces

C.B. Hurley and R.W. Oldford

DRAFT

November 11, 2008

Abstract

We propose using graph theoretic results to develop an infrastructure that tracks movement from a display of one set of variables to another. The illustrative example throughout is the real-time morphing of one scatterplot into another. Hurley and Oldford (2008a) made extensive use of the graph having variables as nodes and edges indicating a paired relationship between them. The present paper introduces several new graphs derivable from this one whose traversals can be described as particular movements through high dimensional spaces. These are connected to known results in graph theory and the graph theoretic results applied to the problem of visualizing high-dimensional data.

1 Introduction

The perennial challenge of visualizing high dimensional data has been addressed in many ways over the years. Typical of many approaches is to lay out small dimensional structures, either spatially or temporally, in such a way that they may be visually linked by the data analyst. In this way, by alternately focussing on low dimensional structures and then linking these together, it is hoped that higher dimensional structure might be revealed. The value of such focussing and linking has long been appreciated within the data visualization community (e.g. Buja et al 1991). In this paper we propose that graph theory can be put to good use in organizing the low dimensional structures and the links between them. The resulting graphs provide infrastructure that can then be navigated to explore high dimensional space.

To be more concrete, suppose we have a cloud of n points in $\mathbb{R}^p$ and choose to visualize this structure by examining all $\binom{p}{2}$ scatterplots of two variables. How might we appreciate its higher dimensional structure? We might cycle through all scatterplots in place on the screen, one after another (e.g. an option offered in `GGobi`, see Swayne et al, 1998), perhaps having the points coloured to aid linking. For relatively small p, a scatterplot

matrix which lays out these 2d projections in a spatial array could be a better choice – then data features can be more leisurely connected by visually scanning along rows and columns. A parallel coordinate plot does something similar, laying out univariate dot plot displays in some order and linking identical points by line segments between displays. Grand tour methods (e.g. Buja et al 1988) move a projection plane through $\mathbb{R}^p$, displaying the projected points on the screen with their positions changing over time as the projection plane moves smoothly about the high dimensional space. Points and structures are easily followed over time. Randomly selecting the planes would ensure some probabilistic coverage of the high dimensional space; alternatively, projection pursuit methods try to optimize the projection by having the projection plane at each step move in a direction that is more "interesting".

Each of these methods is an attempt to navigate through high dimensional space. Each projection is a region of that space and a sequence of projections a trail connecting the regions. Projection pursuit is a local movement strategy that tries to find a more interesting region than the one we are at. An analogy is driving a car in a strange city with no map. The roads are there and we can easily travel from one street to another, but we have no idea which regions of the city are the most interesting to visit. A random drive, if long enough, might ensure that we see the interesting neighbourhoods with some probability, but it will also entail visiting a lot of uninteresting parts as well. Pursuing the most interesting street at each intersection could also be considered, though there is no guarantee that this would lead to a truly interesting neighbourhood. What would be nice to have would be a map, with well marked routes that showed interesting regions and tours.

In an earlier paper (Hurley and Oldford, 2008a) we showed how complete graphs on variables can provide a structure for organizing display components. By associating variables with nodes of a graph, and variable pairs with edges, we were able to turn the problem of ordering one dimensional displays (and 2d transitions between them) into travelling along paths on the complete graph of variables. Immediately, graph path concepts such as Hamiltonian paths, Euler tours, and Hamiltonian decompositions become relevant. Both Hurley and Oldford (20008a) and (2008b) show the application of this graph theoretic framework to improve existing visual displays (e.g. star glyphs, parallel coordinate plots, interaction plots) and to propose new ones (e.g. multiple comparison plots, model selection plots). The algorithms we use and some of the plots we developed are available as an `R` package called `PairViz`.

In the present paper, we further pursue the idea of bringing a graph theoretic approach to bear on problems in data visualization. A number of new graphs are introduced as being relevant to providing navigable infrastructures for high dimensional visualization. Section 2 lays out the basic ideas in the simplest of high dimensional challenges – when the dimension is four. The famous iris flowers of the Gaspé peninsula in Canada provide a four dimensional data set which needs no introduction. In this section, the complete variable-pair graph, the 3d transition graph, the 4d transition graph, and the 3d-space graph are all introduced. With these understood in this simplest case, Section 3 moves on to five and

higher dimensional data. Five dimensional data graphs are used for illustration though the discussion is general and expressed in terms of an arbitrary number of p variates. The same graphs as introduced in Section 2 are discussed in this more general setting. Graph theoretic properties which are useful in determining when Eulerians, Hamiltonians, and Hamiltonian decompositions exist are discussed. The more general setting allows consideration of rich set of relevant graphs which we call k-space graphs to be introduced. In both Sections 2 and 3, the starting position was a complete graph of all variables. Section 4 departs from this and shows that the same constructions can be made from other graphs, graphs which better reflect the interests of the analyst in exploring the data. Section 4.1 takes this a step further, begins with two graphs, each of which represents meaningful structure, and explores how these graphs may be usefully combined according to a number of graph products. Graph theoretic results for these graph products, in particular the existence of Hamiltonian decompositions, are summarized there. Concluding remarks, as well as some indication of where we might go from here, are given in the final section.

2 Four dimensional example: Gaspé Irises

The Iris data, first published by Anderson (1935) and made famous by Fisher (1936), have measurements on four variables: `PetalWidth`, `PetalLength`, `SepalWidth`, and `SepalLength`. The six possible pairs of variables can be laid out as nodes of a complete graph as in Figure 1. Nodes represent some display involving the two named variables and edges between

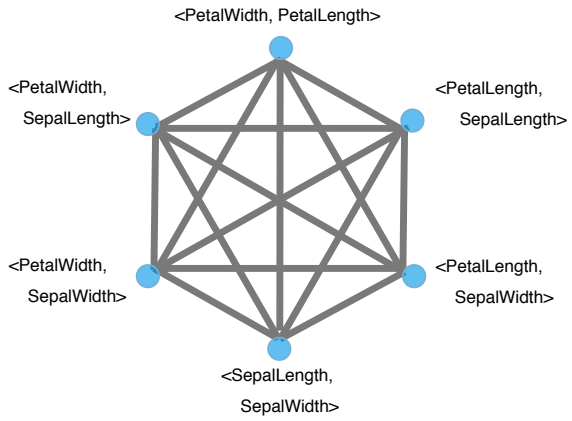


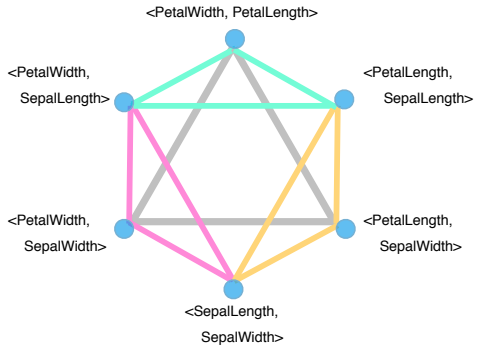
Figure 1: Complete graph on variate pairs from the Iris data.

nodes indicate a transition from the display of one variable pair to that of another.

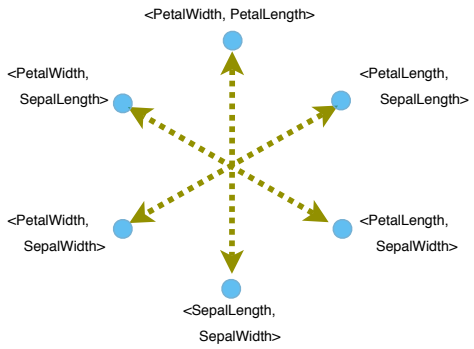
To be concrete, suppose that each node is a scatterplot and that the edge between them indicates a real time transition of one scatterplot morphing into the next. For example, at

the left top node of Figure 1 is the scatterplot of the pair $<$ `PetalWidth`, `SepalLength` $>$ and moving up and right along that edge to the top node means rotating the `SepalLength` axis into the `PetalLength` axis to arrive at the scatterplot of $<$ `PetalWidth`, `PetalLength` $>$. Each node defines a 2d plane upon which the data points are projected and moving along the edge between nodes corresponds to a sequence of 2d projection planes which smoothly morph the scatterplot of one node into that of another.

All edges in the complete graph connect variable pairs which either have one variable or no variables in common. Decomposing the complete graph into these two (as in Figure 2) separates the set of transitions into those which are inherently three dimensional and



(a) 3d transition graph: edges connect nodes that share a variable.



(b) 4d transition graph: edge connects nodes that share no variables.

Figure 2: The complete graph can be decomposed into two separate graphs, each having edge transitions restricted to spaces of fixed dimension, either 3 or 4.

those which are four dimensional. Movement along edges in Figure 2(a), correspond to rigid rotations through three space. These connect with common visual experience and so are easily comprehended.

In fact, the 3d transition graph can itself be decomposed into components which correspond to the individual 3d spaces. Figure 3 arranges these components as nodes of a *3d space graph*, where edges indicate that it is possible to move from one 3d space to another through a rigid rotation (3d transition). The 3d space graph is not in general complete; it is only so in the case of four variables as in this example. Traversing this graph amounts to exploring a 3d space (possibly through arbitrary rotations) at each node and moving from one 3d space to another along an edge by a simple and easily comprehended 3d rigid rotation.

The transitions of Figure 2(b) are less familiar but only slightly more complicated. Moving along an edge here represents a sequence of 2d planes within the space of the 4 node variables. The sequence could be chosen by selecting the variable basis vectors for

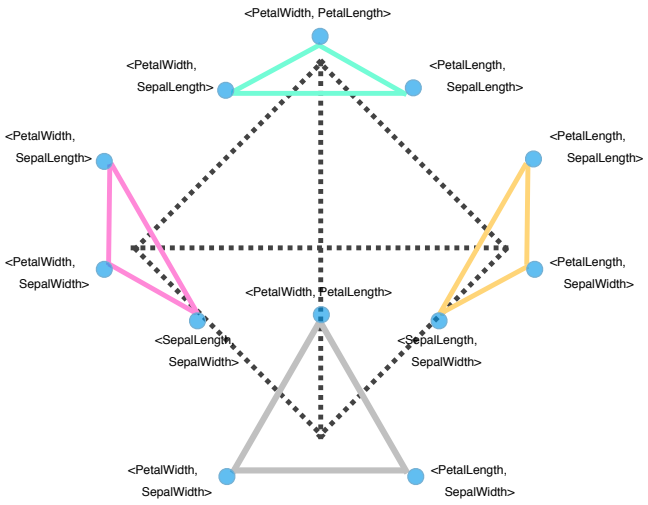
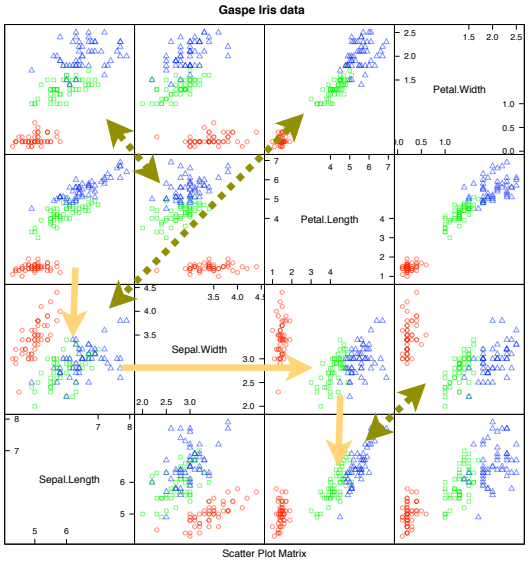


Figure 3: The graph of 3d spaces. Each node is a 3d space component of the 3d transition graph.

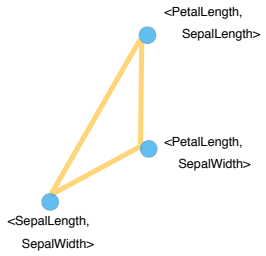
the source and destination planes (nodes) and smoothly interpolating one orthogonal basis set into the other, say along a geodesic path as described in the grand tour (e.g. Buja et al, 1988, Cook et al, 1995) and as implemented, for example, in `XGobi` (Swayne et al, 1998, now `GGobi`). The movement is no longer a rigid rotation and we lose this grounding of common visual experience. However, because each axis of the screen plane has one variable being morphed into one other variable, point movements are still comprehensible within the context of the variables (cf. a grand tour).

These two graphs provide substantively different route patterns for exploring the same higher dimensional space. In the scatterplot matrix of this data of Figure 4 we see that the 3d rigid rotations are equivalent to moves within rows or within columns of the scatterplot matrix. The 4d transitions cannot be achieved by a move along a single row or column. Of course, as is seen from the scatterplot matrix of Figure 4 and even more easily from the complete graph of Figure 1, any 4d transition can be effected as two 3d transitions (with some, presumably minor, loss of information).

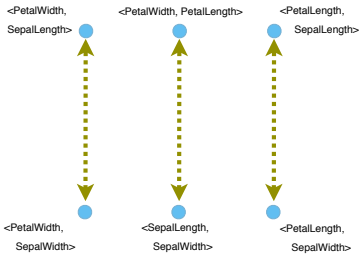
Hopping along rows and columns of a scatterplot matrix, but not both simultaneously (i.e. from row i column j to either row i column k or to row k column j but not to row k column m), is equivalent to following a path on the 3d transition graph of Figure 2. Following a path on the 3d transition graph is the same as viewing one scatterplot after another via 3d rigid rotations. If the path is also Hamiltonian then, like a scatterplot matrix, we will be assured that every pair of scatterplots has been viewed. To follow a Hamiltonian cycle on the 3d transition graph would be to "cycle" through all scatterplots (preferably smoothly via 3d rigid rotations) in the fewest number of steps as opposed to,



(a) Routes on the splom



(b) 3d transitions



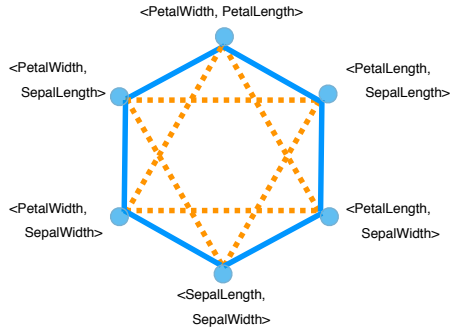
(c) 4d transitions

Figure 4: Routes on the scatterplot matrix of the Iris data. Solid arrows of (a) correspond to 3d transitions of (b), dashed arrows of (a) to the 4d transitions of (c).

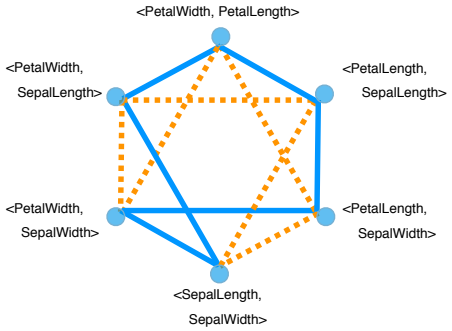
say, cycling across all rows of the scatterplot matrix (e.g. cycling of static XY-plots in `GGobi`, 2008).

Of course not all Hamiltonian cycles are created equal. The solid line cycles shown in Figures 5(a) and 5(b) are both Hamiltonians. However that of Figure 5(a) contains no transition in the 3 space defined by `PetalWidth`, `PetalLength`, and `SepalWidth` – there is no edge in the path from the triangle of these three variables. Every other 3d space will have two 3d rigid rotations presented. By contrast the solid line Hamiltonian of Figure 5(b) contains at least one such rotation for every possible 3d space, though now only two of the four 3d spaces will have two 3d rigid rotations presented. Moreover, the dashed edges of Figure 5(b) also form a Hamiltonian cycle and this too contains at least one 3d transition in every 3d space. The latter is characteristic of Hamiltonians from any Hamiltonian decomposition of the 3d transition graph.

If all 3d transitions are to be presented, then an Eulerian path (one which visits all edges) is called for. These are easily found via general algorithms such as Hierholzer's (1873) or Fleury's (1883) (see also Hurley and Oldford, 2008a). A Hamiltonian decomposition as in Figure 5, however, would allow an Eulerian to be easily constructed from first following one Hamiltonian and then following the next. The resulting presentation would consist



(a) A Hamiltonian cycle (solid lines).



(b) A Hamiltonian decomposition.

Figure 5: Examples of Hamiltonian cycles on the 3d transition graph for 4 variables.

of two separate blocks of all six scatterplots and each block would contain 3d rotational transitions within every one of the four 3d spaces with no transition repeated anywhere.

If the graph has weighted edges, say some cognostic measure of how interesting the transition might be to view, then the `GrEul` algorithm of Hurley and Oldford (2008a) could be employed to find an Eulerian that tends to have interesting transitions appear earlier in the sequence. As its name suggests `GrEul` is a greedy Eulerian algorithm and so is not guaranteed to produce the best such ordering.

2.1 Construction

These graphs are easily constructed. Begin with the individual variables as nodes of a complete graph (shown in Figure 6). For any graph G the *line graph* of G, $L(G)$ is the

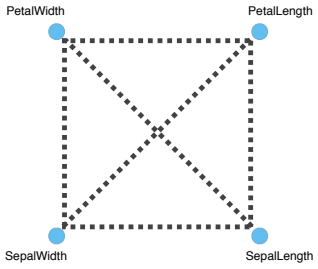


Figure 6: The complete variable graph.

graph whose nodes correspond to edges of G and has an edge e between nodes of $L(G)$

whenever these nodes, as edges of G shared a node in G. It follows then that if G is the variable graph as in Figure 6, then $L(G)$ is the 3d transition graph of these variables. The 4d transition graph is the complement of the 3d transition graph in the complete graph of the variable *pairs*.

3 Five and higher dimensions

As dimensions increase the graphs become increasingly complex. Nevertheless, certain structure persists regardless of dimension.

3.1 3d transition graph

Figure 7(a) shows the variable graph for 5 variables labelled A,B,C,D, and E; the 3d

(a) G, the variable graph for 5 dimensions.

(b) $L(G)$, the 3d transitions graph.

Figure 7: Variable graph and the 3d transition graph for 5 variables: A,B,C,D,E.

transitions graph is given by its line graph in Figure 7(b). As with the Iris data, the 3d transitions graph for 5 variables is even.

More generally, a 3d transition graph, $L(K_p)$, has $p(p-1)/2$ nodes and $p(p-1)(p-2)/2$ edges and every node has even degree, namely $2(p-2)$. $L(K_p)$ is sometimes called the *triangular graph*; it is $2k$-regular (for $k = (p-2)$; regular means all nodes have the same degree) and, being even, is also Eulerian, can be decomposed into cycles, and has an odd number of cycle decompositions (e.g. see Gross and Yellen, 2004, p. 215). If we had weights on the edges of the graph, we could find an Eulerian where smallest (or largest) appeared early in the tour (see Hurley and Oldford, 2008ab). $L(K_p)$ is also Hamiltonian; it contains a spanning cycle.

The graph $L(K_p)$ has many possible cycle decompositions or 2-factorizations. A 2-factor of a graph G is a set of subgraphs of G such that each 2-factor spans G, no two

subgraphs of a 2-factor share an edge, and the degree of each vertex is two. Clearly every 2-factor is a set of cycles if G is a simple graph (no loops, no multi-edges). A 2-factorization is a collection of 2-factors whose edge intersection is null and whose union is G – it is a decomposition of a graph into 2-factors.

Cycle decompositions of $L(K_p)$ provide non-intersecting routes for visiting different regions of p-space via 3d transitions between 2d spaces. Each cycle of a two-factor is a 3d transitional tour through a particular "neighbourhood" of 2d spaces. For example, the cycle decomposition of Figure 5(a) has one two factor that provides two 3-cycle routes, one through the 3 dimensional "neighbourhood" of {`PetalWidth`, `PetalLength`, `SepalLength`} and one through the 4 dimensional "neighbourhood" of all variables. Its second two factor is a Hamiltonian cycle and so provides a longer tour through the four dimensional space. The two factorization of Figure 5(b) provides two different detailed tours through the whole space, a Hamiltonian decomposition. Hamiltonian decompositions are of particular interest precisely because each 2-factor is a single unique route that tours the entire p-space. Different Hamiltonians provide different views of the entire space.

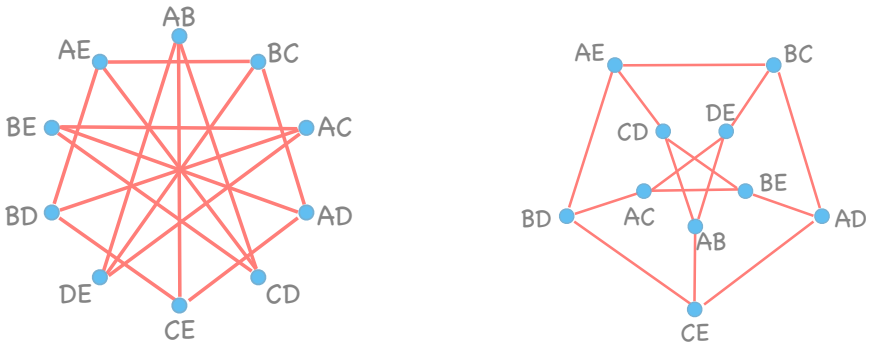
Every Euler tour of a graph G corresponds to a Hamiltonian cycle on its line graph $L(G)$. Heinrich and Verrall (1997) and Verrall (1998) construct a "perfect set" of Euler tours on K_{2k+1} and on $K_{2k} + I$, respectively (or K_n^* for odd or even n, in the notation of Hurley and Oldford 2008a), which perfectly partition the set of 2-paths (pair of adjacent edges) on these graphs. Each Euler tour of a perfect set shares no 2-path with another and is therefore a distinct Hamiltonian on the line graph. By partitioning the 2-paths, the perfect set of Euler tours on K_p^* becomes a cycle decomposition of Hamiltonians on its line graph. The result is that the 3d-transition graph, $L(K_p)$, is Hamilton decomposable for all p.

Other cycle decompositions of $L(K_p)$ also exist. For example, Cox and Rodger (1996) show that m-cycle decompositions of $L(K_p)$ exist for certain values of m and p. In particular, if $m = 2^i$ then there exists an m-cycle decomposition of $L(K_p)$ if and only if $p \equiv 1 \pmod{2m}$ or $p \equiv 0$ or $2 \pmod{m}$. These and other decompositions could be helpful for large p.

3.2 4d transition graph

The 4d transition graph for $p = 5$ is shown in Figure 8(a). It is a 3-regular graph and hence is not Eulerian. The graph is isomorphic to the Petersen graph and usually drawn as in Figure 8(b). Though the graph is not Hamiltonian, it does contain many Hamiltonian paths. For example, begin at AB of in either graph of Figure 8, move to CD and follow the edges of the pentagram to DE. From DE move to BC and then follow the edges of the pentagram clockwise until reaching AE. The result is a Hamiltonian path that visits all 2d spaces through 4d transitions.

For any p, the 4d transition graph, $\overline{L(K_p)}$, is a kind of *Kneser* graph, namely $KG(p, 2)$. More generally, a Kneser graph $KG(p, m)$ is a graph whose vertices are the $\binom{p}{m}$ subsets



(a) 4d transition graph for $p = 5$.

(b) Petersen graph.

Figure 8: The 4d transition graph for 5 variables: A,B,C,D,E. It is $\overline{L(K_5)}$ complement of $L(K_5)$.

of size m with edges between vertices whose subsets do not intersect. The complement of a Kneser graph is sometimes called a Johnson graph (providing yet another name for the 3d-transition graph). The 4d transition graph is a k-regular graph for $k = \binom{p-2}{2}$ and connected whenever $p > 4$. Chen (2000) has shown that $KG(p, m)$ graphs are Hamiltonian for $p \geq 3m$ and so 4d transition graphs have a Hamiltonian cycle whenever $p \geq 6$.

Moreover, 4d transition graphs are even whenever $(p-2)(p-3) \equiv 0 \pmod 4$. A generative formula is $p = (5 + \sqrt{1 + 4a(a+1)})/2$ for integer $a > 2$ with $a \pmod 4 \in \{0, 3\}$ (as shown in the Appendix). Because it is connected, whenever the 4d transition graph is even it is also Eulerian. So at least half the time we can employ a greedy Eulerian algorithm to tour the 2d spaces via 4d transitions.

3.3 3d spaces graph

Separating the 3d transition graph into 3d space subgraphs yields the 3d space graph for 5 variables of Figure 9(a). The obvious isomorphism with the 3d transition graph is peculiar to the $p = 5$ case. In general the 3d space graph has $\binom{p}{3}$ nodes, each of degree $3(p-3)$ which happen to match those of the 3d transition graph when $p = 5$. Similarly, when $p = 5$, the complement of the 3d space graph (Figure 9(b)) is isomorphic to the 4d transition graph although this is not generally the case.

The 3d space graph is reminiscent of the Kneser graph. The nodes correspond to all $\binom{p}{3}$ subsets of three variables from the p available, but instead of edges between disjoint subsets edges exist between nodes that share a single variable. In general, if the complete

(a) The graph of 3d spaces (S_2).

(b) Complement of the 3d space graph (S_1 since $p = 5$).

Figure 9: 3d variable spaces for $p = 5$ variables.

graph on the 3d space nodes is K_n where $n = \binom{p}{3}$ with vertex set V and edge set E, then this graph can be decomposed into 3 distinct graphs S_0, S_1, and S_2 where S_i has vertex set V but edge set E_i and $e \in E_i$ if and only if the vertices of e share exactly i variables.

S_2 is the 3d space graph constructed from the 3d transition graph. Movement along edges here correspond to 3d transitions. For example, moving along the edge from ABC to BCD might be achieved by rotating A into D on one axis while holding the other fixed as either B or C. Alternatively, the semantics of the common pair of variables might be something quite different altogether. The common pair might identify variables to be conditioned on, so that edge indicates that an interest in the joint behaviour of the distinct variables given the common variables.

S_1 is a 3d space graph as well, one where the connected spaces share only one variable. Although movement along edges here indicate 4d transitions, it is not clear how this graph would be derived from the 4d transition graph.

S_0 is the same as $KG(p, 3)$; when $p = 5$, $E_0 = \phi$. If $p \geq 6$, the edges connect disjoint 3d spaces. Movement would represent transitions from one 3d space to another, 6d transitions. One possible visualization connecting ABC to DEF, say, might be a splom of ABC dynamically morphing into a splom of DEF by 4d transitions of scatterplots AB to DE, AC to DF, and BC to EF.

3.4 k-d spaces graph

The decomposition of the complete 3d space graph and its similarity to the Kneser graph suggests a more general graph construction. Let $S(p, k, i)$ denote the k-d space graph whose nodes are the $\binom{p}{k}$ subsets of the p variables and where edges are drawn only between nodes having exactly i variables in common. Clearly $i < k < p$ and if $S(p, k)$ is the complete

graph on this vertex set it can be decomposed as

$$S(p,k) = \sum_{i=0}^{k-1} S(p,k,i).$$

Of course $S(p,k,0) = KG(p,k)$. Such graphs would seem to provide navigational structure between large dimensional spaces. Again one could imagine morphing one splom on k variables into another splom on a disjoint set of k variables by moving along an edge of $S(p,k,0)$.

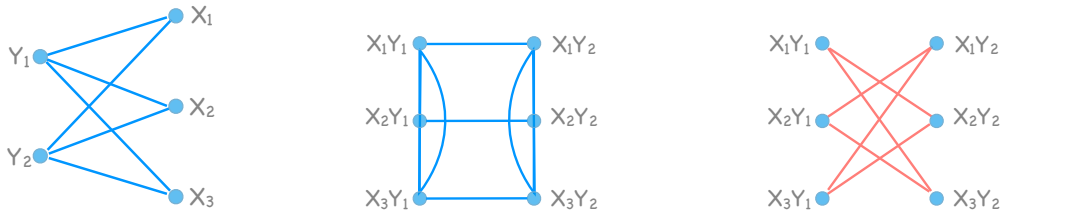
The theoretical properties of these graphs need to be explored. Some mileage might be had from the following observation. Consider the case $p = 5$ illustrated above. The 3d space graph seems to be constructed from the complete variable graph as follows. First construct the 3d transition graph as $L(K_p)$. Now take the line graph of this graph. The result will be a graph whose nodes might be written as $A^a B^b C^c D^d E^e$ where $a, b, c, d, e \in \{0, 1, 2\}$ and $a + b + c + d + e = 4$. Letters with zero exponent need never appear in the label. Suppose $R(\cdot)$ is a graph operator that replaces each power that is 2 or greater by 1 and then replaces all nodes having the same label by the single reduced label, preserving a single edge between any connected and distinct labels. Using this operator, the 3d space graph of Figure 9(a) would be expressible as $R(L^2(K_p))$. Similarly, the 3d space graph of Figure 9(b) is $\overline{R(L^2(K_p))}$. Many interesting graphs seem to be built up from a complete graph through the three operators $L(\cdot)$, $R(\cdot)$ and complement. It would be of interest to know what graph properties, if any, are preserved and/or created by these operations, singly and in composition.

4 Other relevant graph structures

The complete graph on all variables treats all variables symmetrically. In some cases it is more interesting to distinguish variables and/or the pairwise relations we wish to consider in our displays.

For example, a multivariate regression model distinguishes response variables $Y_1, Y_2, \ldots, Y_q$ from explanatory variables $X_1, X_2, \ldots, X_r$. The relationship between Ys and Xs are of primary interest. This interest could be represented by the graph shown for $q = 2$, $r = 3$ in Figure 10(a) where only connections between Y variables and X variables are allowed. That is only the various (explanatory, response) relations are of interest.

The 3d transition graph is constructed as before as the line graph of the variables graph, as shown in Figure 10(b). Transitions along here could be displayed as 3d rotations of one scatterplot into another where the horizontal axis is reserved for changes in X variable and the vertical for changes in Y variables. Similarly, the 4d transition graph is shown in Figure 10(c). Again we could imagine the Ys determining the vertical coordinate and the Xs the horizontal coordinate as one scatterplot morphed into the next by interpolation through a 4d space. These sets of transitions are similar to the $2 \times 1d$ tours of `GGobi` with



(a) Initial variable graph: G (b) The 3d transition graph: $L(G)$ (c) The 4d transition graph: $\overline{L(G)}$

Figure 10: The variables are restricted to reflect response (Ys) and explanatory (Xs) variables.

the restriction that each axis only considers either a single variable or linear combinations of a pair of variables in each 1d tour.

Construction of these graphs are the same as before except that we start with a graph, G, structured to reflect the pairwise relationships of interest, instead of K_p. The 3d transition graph is $L(G)$ and the 4d transition $\overline{L(G)}$. The 3d space graph can again be formed as suggested in Section 3.4 as $R(L^2(G))$. In this example, it is the complete graph K_9 corresponding exactly to $L^2(G)$ with nodes relabeled to have only three distinct variables (i.e. here $R(\cdot)$ did not compress any nodes into one). It is even, Eulerian, and possesses many Hamiltonian decompositions (see Hurley and Oldford, 2008a).

Note that there is neither an Eulerian nor a Hamiltonian cycle in G of Figure 10(a). Yet there is a Hamiltonian for its 3d transition plot $L(G)$. Chartrand (1965) gave necessary and sufficient conditions for this to be the case.

Definition 1 *A graph G having q edges is called* sequential *if the edges of G can be ordered as $e_0, e_1, e_2, \ldots, e_{q-1}, e_q = e_0$, so that e_i and e_{i+1}, $i = 0, 1, 2, \ldots, q-1$, are adjacent (share a vertex).*

$L(G)$ contains a Hamiltonian cycle if and only if G is sequential graph (Chartbrand, 1965).

The graph G of Figure 10(a) is a sequential graph and so its 3d transition graph has a Hamiltonian cycle. G is in fact $K_{2,3}$, a complete bipartite graph. The general complete bipartite graph is denoted $K_{q,r}$ and is easily seen to be a sequential graph for all q and r. Consequently, the 3d transition graph $L(K_{q,r})$ will be Hamiltonian for any "multivariate regression problem" as described above.

Chartrand (1965) also gives conditions for a Hamiltonian to exist in repeated (or interated) line graphs, which might be helpful in traversing a 3d-space graph for example. In particular, if G is Hamiltonian, then so is $L^n(G)$ for all $n \geq 1$. And, if G is a nontrivial connected graph of order p, and G is not a path, then $L^n(G)$ is Hamiltonian for all $n \geq p-3$. Later Chartbrand and Wall (1973) showed that it is enough that G is connected and of minimum degree 3 for $L^2(G)$ to be hamiltonian.

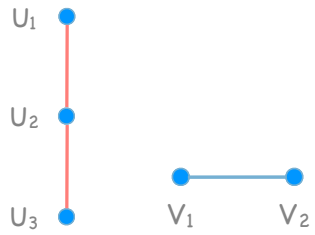
It is also of interest to know when a Hamiltonian decomposition exists for the 3d transition graph $L(G)$. Muthsamy and Paulraja (1995) provide three relevant results. First, if G has a Hamiltonian cycle decomposition into an *even* number of Hamiltonian cycles, then $L(G)$ has a Hamiltonian cycle decomposition. Second, if G has a Hamiltonian cycle decomposition into an *odd* number of Hamiltonian cycles, then the edge set of $L(G)$ can be partitioned into Hamiltonian cycles and a 2-factor. And finally if G is a 2k-regular graph that is Hamiltonian, then the edge set of $L(G)$ can be partitioned into Hamiltonian cycles and a 2-factor. An example of this last result from the "multivariate regression problem" is the symmetric complete bipartite graph $G = K_{q,q}$ when q is even.

Also related to bipartite graphs is the following result from Pike (1995). If G is bipartite and $(2k+1)$ regular and Hamilton decomposable, then so is $L(G)$. These (and other) results support a general conjecture attributed to Bermond (Alspach et al, 1990), namely that if G has a Hamiltonian cycle decomposition, then so does $L(G)$.

4.1 Graph products

Another way to construct graphs for pairs of variables is via graph products. Suppose the variables separate into two sets $\mathcal{U} = \{U_1, U_2, \ldots, U_m\}$ and $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$ and that there is a graph associated with each of these sets, say G and H respectively. Each graph would separately model the relationships between that variable set according to some semantics.

For example, Figure 11 shows two possible variable graphs G and H. The semantics of



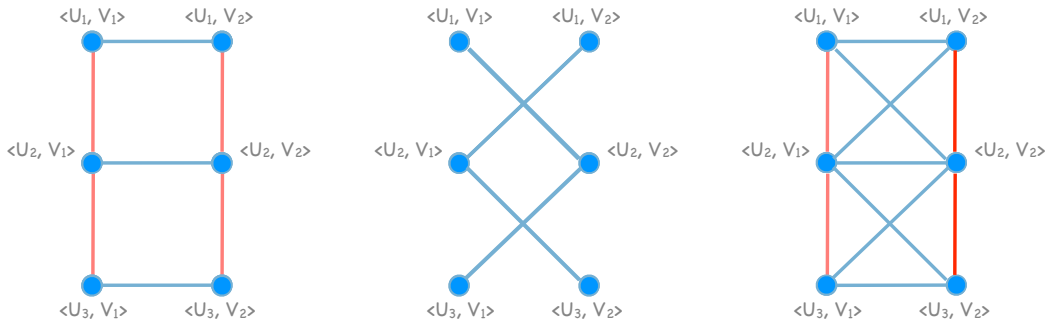
(a) G (b) H

Figure 11: Variable graphs: G for variables from $\mathcal{U}$; H for those from $\mathcal{V}$.

these graphs are such that in G interest lies in how U_1 relates to U_2 and U_2 to U_3 but not U_1 to U_3. (If all three relationships were of interest, then a complete graph K_3 would be in order for G.) For example, U_3 follows U_2 in time, and U_2 follows U_1; having U_1 follow U_3 would not make sense. Another example might be that U_i is the ith principal component (corresponding to the ith largest eigenvalue of some matrix) and interest lies in the effect of adding the principal components in order. Or the data analyst is just declaring interest

in this order of variables for some reason. The graph H would be similarly interpretable for variables in $\mathcal{V}$. $H = K_2$ of Figure 11(b) has only two variables whose relationship is of interest.

Figure 12 shows three different graph products of these graphs. Figure 12(a) shows the



(a) Cartesian product: $G \square H$ (b) Tensor product: $G \times H$ (c) Strong product: $G \boxtimes H$

Figure 12: Graph products of G and H from Figure 11.

Cartesian product, Figure 12(b) the *Tensor product*, and Figure 12(c) the *Strong product*. These products share the same vertex set, $\mathcal{U} \times \mathcal{V} = \{< U_i, V_j >: U_i \in \mathcal{U}, V_j \in \mathcal{V}\}$, and differ only in the edges. Each tries to preserve something of the original variable relationships in G and H. As the figures and notation suggest, $(G \square H) + (G \times H) = (G \boxtimes H)$.

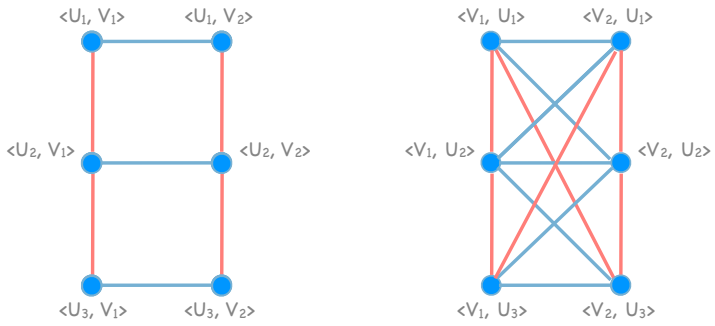
The *Cartesian product*, $G \square H$, has an edge between vertices $< u, v >$ and $< s, t >$ iff either $u = s \in \mathcal{U}$ and v is adjacent to t in H or $v = t \in \mathcal{V}$ and u is adjacent to s in G. This is the 3d transition graph, *restricted to the permitted transitions between variables.*

The *tensor product* (or direct product or weak product or conjunction), $G \times H$, has an edge between vertices $< u, v >$ and $< s, t >$ iff u and s are adjacent in G and v and t are adjacent in H. This is a 4d transition graph, restricted to the permitted transitions between variables.

The *strong product* (or strong direct product or normal product), $G \boxtimes H$, has an edge between vertices $< u, v >$ and $< s, t >$ iff either u and s are adjacent in G and $v = t$, or, $u = s$ and v and t are adjacent in H. As the figures and notation suggest, $(G \square H) + (G \times H) = (G \boxtimes H)$. So the *restricted* 3d transition graph is the complement of the *restricted* 4d transition graph in $G \boxtimes H$ (i.e. not in the complete graph).

These products are but three of a potential twenty graph products which can be formed having vertex set $V(G) \times V(H)$ and edge set determined only by the edge sets in G and H (Imrich and Izbicki, 1975). These three graph products are both associative and commutative (in that the resulting graphs are isomorphic to one another). An example of an associative graph product which is not symmetric is the *lexicographic product* which is perhaps better named the *composition* of two graphs.

The *composition* (or lexicographic product or sometimes the wreath product) $G[H]$ of graphs G and H has an edge between vertex $< u, v >$ and vertex $< s, t >$ if and only if either u is adjacent to s in G or $u = s$ and v is adjacent to t in H. Note the asymmetry of G and H in the definition. Figure 13 shows the compositions $G[H]$ and $H[G]$ respectively.



(a) $G[H]$ (b) $H[G]$

Figure 13: Graph compositions of G and H from Figure 11.

It is straightforward to see that the composition of two complete graphs is symmetric. Indeed, the composition of two graphs commutes if and only if either both graphs are complete, or both are edgeless, or both are "powers" (with respect to composition) of the same graph G (p. 489 of Gross and Yellen, 2004). The composition is also the only one of the above products that is self complementary, in the sense that $\overline{G}[\overline{H}] \cong \overline{G[H]}$.

Much is known about these products, in particular about their Hamiltonian decompositions. For example, some results are summarized in Table 1 (dashed entries indicate absence of information). Most of these can be found in either Bosák (1990) or Gould (2004). Clearly, it helps to have G and H both be Hamiltonian decomposable to start, then all four products are Hamiltonian decomposable, the tensor product being the only one with an extra condition.

Recall that $G \square H$ is the relevant 3d transition graph and $G \times H$ the relevant 4d transition graph. Complete graphs are of special interest as the products correspond to moving about rectangles of a scatterplot matrix (splom). This means that we can focus on some region in the scatterplot matrix and construct Hamiltonians and/or Eulerians in most cases. That case where we cannot (i.e. $m + n$ odd), there will still be a Hamiltonian decomposition but there will be redundant moves in connecting the 1-paths of the 1-factor. The theory and algorithms of Hurley and Oldford (2008ab) are relevant to all these cases. When $G = H = K_p$, the splom is square and has a univariate display on its diagonal (e.g. that produced by `GGobi` which has a univariate density estimate on its diagonal cells).

It would be nice to extend this table to other graphs whose structure is of interest. For example paths are useful to model time order or data from specified links in a chain

Table 1: Summary of Hamiltonian decompositions of Graph Products

G	H	$G \square H$	$G \times H$	$G \boxtimes H$	$G[H]$
Ham-decomp	Ham-decomp	Ham-decomp	—	Ham-decomp	Ham-decomp
and order(G) is odd		—	Ham-decomp	—	—
K_m	K_n	Ham-decomp as:	(m+n-2)-reg.	$K_{m \times n}$	$G \boxtimes H$
(m+n) even		$(m+n-2)/2$ cycles	– Eulerian		
(m+n) odd		$(m+n-3)/2$ cycles and one 1-factor			
$G = K_m$	$H = K_n$	all rectangular moves in a splom rectangle	all diagonal moves	both	both
$G = H = K_p$		(+ univariate)	(+ univariate)	(+ univariate)	(+ univariate)

(e.g. Markov or causal). Bipartite graphs are useful to model "multivariate regression" situations. It is easy to imagine situations where exploring the product of these (in any combination path times bipartite, etc) could be of statistical interest.

5 Concluding remarks

The above theory may be directly applied to the layout of statistical graphics. The temporal morphing of one scatterplot into another via 3d or 4d transitions could be easily implemented in any system that supports fast drawing and erasure of scatterplots. These could, for example, be specialized tours in `GGobi`. Similarly, a whole square block of a scatterplot matrix could be simultaneously morphed into another target block of the same size to show a transition from a many dimensional space to another of the same size. There are no doubt many new graphics that could be designed around traversal of these graphs.

For small p the graphs could even provide a user interface to drive the transitions. For example, moving a ball around a Petersen graph could be an effective control for the user. For large graphs, it might be more useful to use as a map to show where we are going as well as where we have been, perhaps marking interesting paths as we proceed.

Weights can be placed on the edges of any of these graphs. In Hurley and Oldford (2008a) we used various scagnostic indices (Wilkinson et al, 2005) on the edges of the complete variable graph to produce interesting arrangements of parallel coordinate axes. In the case of the 3d and 4d transition graphs, the edges represent transitions between two

dimensional spaces. Some of the scagnostics naturally port to higher dimensions, others do not. There is an opportunity to develop fundamentally new cognostics for transitions between these and higher dimensional spaces.

With weighted graphs, Hamiltonians of minimum weight (a travelling salesman problem), greedy Eulerian ordering, and total weight of each cycle of a Hamiltonian decomposition become interesting.

Most of the discussion has used the real time morphing of one scatterplot into another as illustration, but this need not be the case. The displays at each node could be anything on those variables and the layout might be spatial rather than temporal. Moreover the semantics of the edge transitions might be different. For example, in a 4-space graph $S(p, 4, 2)$ ($p \geq 6$) nodes $ABCD$ and $CDEF$ would be connected and the transition from one node to another might mean conditioning on the shared variables. We might temporally morph Cleveland's (1993) conditional plot of $(A, B)|(C, D)$ into that of $(E, F)|(C, D)$.

A large number of variables need not be intimidating. As Section 4 pointed out, in many cases there may not be interest in connections between some variables. For example, the original variable graph might be bipartite. Or it might decompose into two (or more) subgraphs for which some graph product is of interest. In these cases, there are still graph theoretic results available vis-à-vis hamiltonians, Eulerians, and meaningful decompositions.

In Hurley and Oldford (2008ab) we used graph theory to good effect for the spatial ordering of pairwise displays such as parallel coordinate plots, glyphs, and multiple comparison plots. In the present paper, we have shown that graph theoretic structure is of interest more generally for the layout (spatially or temporally) of any displays of high dimensional information (variables or otherwise). The structure is navigable and can often be decomposed and/or simplified. We anticipate that a broader graph theoretic approach to layout will lead to new tools and methods for visualizing high dimensional data.

Appendix

The generative formula referred to at the end of Section 3.2 is simply derived.

Fact 1 *For $p > 5$, the 4d transition graph (or $KG(p, 2)$) is even whenever*

$$p = \frac{5 + \sqrt{1 + 4a(a+1)}}{2}$$

with integer $a > 0$ such that $a \ (mod\ 4) \in \{0, 3\}$.

Proof: The 4d transition graph is a regular graph with vertex degree $\binom{p-2}{2}$ and is even iff this equals $2k$ for some integer $k > 0$. And this is true

$$\begin{aligned} &\iff \quad (p-2)(p-3) = 4k \\ &\iff \quad p = \frac{5 \pm \sqrt{1 + 16k}}{2}. \end{aligned}$$

This last quantity is rational iff $\sqrt{1+16k}$ is an odd integer. Or for some integer $a \geq 0$ we must have

$$\begin{aligned} & \sqrt{1+16k} = 2a+1 \\ \iff \quad & k = \frac{a(a+1)}{4}. \end{aligned}$$

k is integer and because one of a or $(a+1)$ must be odd, 4 must divide the even one – either $(a+1)$ or a. So we have either $a \equiv 0 (\text{mod } 4)$ or $a \equiv 3 (\text{mod } 4)$. □

References

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.

Anderson, E. (1935). "The irises of the Gaspé Peninsula". *Bulletin of the American Iris Society*, 59, pp 2-5.

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.

Bosák, J. (1990), *Decompositions of Graphs*, Kluwer, Dordrecht.

Buja, A., D. Asimov, C. Hurley, and J. A. McDonald (1988), "Elements of a Viewing Pipeline for Data Analysis" in *Dynamic Graphics for Statistics*, W. Cleveland and M. McGill (eds.), Wadsworth, pp. 277-308.

Buja, A. J.A. McDonald, J. Michalak, and W. Stuetzle (1991), "Interactive data visualization using focusing and linking", *Proc. 2nd IEEE Conference on Visualization '91*, pp. 156-163.

Chartrand, G. (1965), "The existence of complete cycles in repeated line-graphs", *Bulletin of the American Mathematical Society*, 71, pp. 668-670.

Chartrand, G. and C.E. Wall (1973), "On the Hamiltonian index of a graph", *Studia Sci. Math. Hungar.*, 8, pp. 43-48.

Chen, Y-C. (2000), "Kneser graphs are Hamiltonian for $n \geq 3k$" *Journal of Combinatorial Theory, Series B*, 80, pp. 69-79.

Cleveland, W.J. (1993), *Visualizing Data*, Hobart Press.

Cook, D., A. Buja, J. Cabrera and C. Hurley (1995) Grand Tour and Projection Pursuit, *Journal for Computational and Graphical Statistics*, 4, 155-172.

Cox, B.A. and C.A. Rodger (1996), "Cycle Systems of the Line Graph of the Complete Graph", *Journal of Graph Theory*, 21, pp. 173-182.

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.

Fisher, R.A. (1936), "The use of multiple measurements in taxonomic problems" *Annals of Eugenics*, 7, pp. 179-188.

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

Gould, R.J. "Hamiltonian graphs", Chapter 4.5 of *Handbook of Graph Theory* (J. Gross and J. Yellen, eds.), CRC Press, pp.261-278.

Gross, J.L. and J. Yellen (eds.) (2004), *Handbook of Graph Theory*, CRC Press.

Heinrich, K. and H. Verrall (1997), "A Construction of a Perfect Set of Euler Tours of K_{2k+1}", *Journal of Combinatorial Designs*, 5, pp. 215-230.

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.

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

Hurley, C. and R.W. Oldford (2008a), "Pairwise Display of high-dimensional information via Eulerian tours and Hamiltonian decompositions", (submitted), 30 pages

Hurley, C. and R.W. Oldford (2008b), "Eulerian tour algorithms for data visualization and the PairViz package", (submitted), 15 pages

Imrich. W. and H. Izbicki (1975), "Associative products of graphs" *Monatschefte für Mathematik*, 80, pp. 277-281.

Pike, D.A. (1995), "Hamilton Decompositions of Line Graphs", *Journal of Graph Theory*, 20, pp. 473-479.

Verrall, H. (1998), "A Construction of a Perfect Set of Euler Tours of $K_{2k} + I$", *Journal of Combinatorial Designs*, 6, pp. 183-211.

Swayne, D.F., D. Cook and A. Buja (1998), "XGobi: Interactive Dynamic Data Visualization in the X Window System", *Journal for Computational and Graphical Statistics*, 7, pp. 113-130.

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