

# Graph-Theoretic Scagnostics\*

Leland Wilkinson<sup>†</sup>  
SPSS Inc.  
Northwestern University

Anushka Anand<sup>‡</sup>  
University of Illinois at Chicago

Robert Grossman<sup>§</sup>  
University of Illinois at Chicago

## ABSTRACT

We introduce Tukey and Tukey *scagnostics* and develop graph-theoretic methods for implementing their procedure on large datasets.

**CR Categories:** H.5.2 [User Interfaces]: Graphical User Interfaces—Visualization; I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques;

**Keywords:** visualization, statistical graphics

## 1 INTRODUCTION

Around 20 years ago, John and Paul Tukey developed an exploratory visualization method called *scagnostics*. While they briefly mentioned their invention in [42], the specifics of the method were never published. Paul Tukey did offer more detail at an IMA visualization workshop a few years later, but he did not include the talk in the workshop volume he and Andreas Buja edited [7].

Scagnostics was an ingenious idea. Jerome Friedman and Werner Stuetzle, in a paper assessing John Tukey's lifetime contributions to visualization [13], say the following:

Draftman's views (scatterplot matrices) lose their effectiveness when the number of variables is large. Using a projection index similar to that in projection pursuit, the computer could find the most interesting scatterplots to be presented to the user. John had proposals for a wide variety of scagnostic indices to judge the usefulness of scatterplot displays. The widespread use of cognostics and scagnostics has not yet materialized in routine data analysis. These approaches are perhaps among the potentially most useful of John's yet to be explored suggestions.

Scagnostics have yet to be explored by others, despite this encouragement. This may be due to the lack of published details. In any case, this paper summarizes the Tukeys' idea and offers a new approach that we believe follows the spirit of their method. Our approach is based on recent advances in graph-theoretic summaries of high-dimensional scattered point data. We believe our method improves the computational efficiency and extends the scope of the original idea.

We will begin with a brief summary of the Tukeys' approach, based on the first author's recollection of the IMA workshop and subsequent conversations with Paul Tukey. Then we will present our graph-theoretic measures for computing scagnostic indices. Finally, we will illustrate the performance of our methods on real data.

\*John Hartigan, David Hoaglin, and Graham Wills provided valuable suggestions.

<sup>†</sup>e-mail: leland@spss.com

<sup>‡</sup>e-mail: aanand2@uic.edu

<sup>§</sup>e-mail: grossman@cs.uic.edu

## 2 TUKEY AND TUKEY SCAGNOSTICS

A scatterplot matrix, variously called a SPLOM or casement plot or draftman's plot, is a (usually) symmetric matrix of pairwise scatterplots. An easy way to conceptualize a symmetric SPLOM is to think of a covariance matrix of  $p$  variables and imagine that each off-diagonal cell consists of a scatterplot of  $n$  cases rather than a scalar number representing a single covariance. This display was first published by John Hartigan [19] and was popularized by Tukey and his associates at Bell Laboratories [9]. Figure 1 shows a SPLOM of measurements on abalones using data from [27]. Off the diagonal are the pairwise scatterplots of nine variables. The variables are sex (indeterminate, male, female), shell length, shell diameter, shell height, whole weight, shucked weight, viscera weight, shell weight, and number of rings in shell.

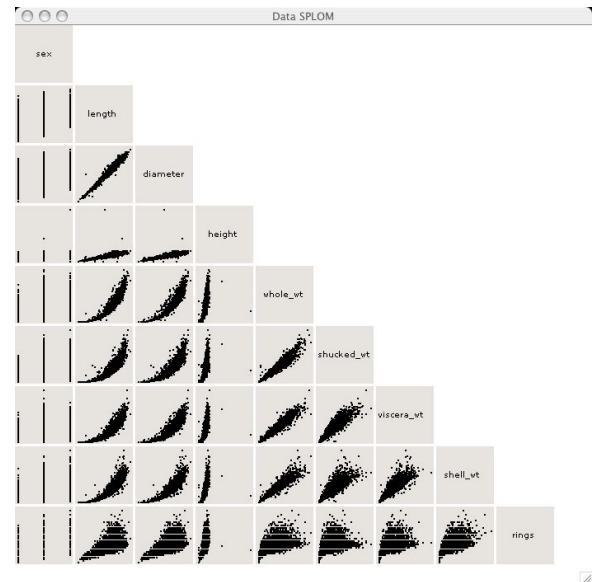


Figure 1: Scatterplot matrix of Abalone measurements

As Friedman and Stuetzle noted, scatterplot matrices become unwieldy when there are many variables. First of all, the visual resolution of the display is limited when there are many cells. This defect can be ameliorated by pan and zoom controls. More critical, however, is the multiplicity problem in visual exploration. Looking for patterns in  $p(p-1)/2$  scatterplots is impractical when there are many variables. This problem prompted the Tukeys' approach.

The Tukeys reduced an  $O(p^2)$  visual task to an  $O(k^2)$  visual task, where  $k$  is a small number of measures of the distribution of a 2D scatter of points. These measures included the area of the peeled convex hull [40] of the 2D point scatters, the perimeter length of this hull, the area of closed 2D kernel density isolevel contours [35], [33], the perimeter length of these contours, the convexity of these contours, a modality measure of the 2D kernel densities, a nonlinearity measure based on principal curves [21] fitted to the 2D scatterplots, and several others. By using these measures,

the Tukeys aimed to detect anomalies in density, shape, trend, and other features in the 2D point scatters.

After calculating these measures, the Tukeys constructed a scatterplot matrix of the measures themselves. This step amounted to a level of indirection, like a pointer, in which each point in the scagnostic SPLOM represented a scatterplot cell in the original data SPLOM. With brushing and linking tools, unusual scatterplots could be identified from outliers in the scagnostic SPLOM.

In his IMA talk, Paul Tukey mentioned an additional step we might take. The scagnostic SPLOM can be thought of as a visualization of a set of pointers. We can therefore construct a set of pointers to pointers. In doing so, we can locate unusual clusters of measures that characterize unusual clusters of raw scatterplots.

### 3 GRAPH-THEORETIC SCAGNOSTICS

There are two aspects of the Tukeys' approach that can be improved. First, some of the Tukeys' measures, particularly those based on kernels, presume an underlying continuous empirical or theoretical probability function. This is appropriate for scatters sampled from continuous distributions, but it can be a problem for other types of data. Second, the computational complexity of some of the Tukey measures is  $O(n^3)$ . Since  $n$  was expected to be small for most statistical applications of this method, such complexity was not expected to be a problem.

We can ameliorate both these problems by using graph-theoretic measures. Indeed, the Tukeys used a few themselves. First, the graph-theoretic measures we will use do not presume a connected plane of support. They can be metric over discrete spaces. Second, the measures we will use are  $O(n \log(n))$  in the number of points because they are based on subsets of the Delaunay triangulation. Third, we employ adaptive hexagon binning [8] before computing our graphs to further reduce the dependence on  $n$ .

There is a price for switching to graph-theoretic measures, however. They are highly influenced by outliers and singletons. Whenever practical, the Tukeys used robust statistical estimators to down-weight the influence of outliers. We will follow their example by working with nonparametric and robust measures. Further, we will remove outlying points in the minimum spanning tree before computing our graphs and the measures based on them.

We next introduce some notation for the projections underlying the scagnostics idea.

#### 3.1 Scagnostic Projections

- Let  $\pi_i$  denote the projection

$$\mathbf{R}^p \longrightarrow \mathbf{R}, \quad x \rightarrow x_i$$

on the  $i$ th coordinate.

- Let  $\pi_i \times \pi_j$  denote the projection

$$\mathbf{R}^p \longrightarrow \mathbf{R}^2, \quad x \rightarrow (x_i, x_j).$$

- Given a set  $X \subset \mathbf{R}^p$ , let  $X_{[i,j]}$  denote the set  $(\pi_i \times \pi_j)(X) \subset \mathbf{R}^2$ .

We will be interested in the  $p(p-1)/2$  different data sets  $X_{[i,j]}$  formed by varying  $i, j = 1, \dots, p$  where  $i < j$ .

To define the scagnostics transform, we first need to fix  $k$  measures  $c_1, \dots, c_k$  defined on sets  $X$  in  $\mathbf{R}^2$ . For example,  $c_1$  can be a measure of the central location of the points,  $c_2$  can be a measure of the dispersion of the points, and so on.

Given a data set  $X$  containing  $n$  points in  $\mathbf{R}^p$ , the scagnostics transform  $\tau(X)$  is the data set of  $p(p-1)/2$  points in  $\mathbf{R}^k$  containing the following points:

$$(c_1(X_{[i,j]}), \dots, c_k(X_{[i,j]})) \in \mathbf{R}^k, \quad i, j = 1, \dots, p, \quad i < j.$$

**Remark 1.** Given a data set  $X \subset \mathbf{R}^p$  containing  $n$  points, we are interested in the scatterplot matrix of  $X$  and the scatterplot matrix of  $\tau(X)$ .

**Remark 2.** Note that the same definition works for  $k$  measures defined on projections in higher dimensional spaces. For example, we can work with projections in three space, with  $X_{[i,j,k]}$  denoting the set  $(\pi_i \times \pi_j \times \pi_k)(X) \subset \mathbf{R}^3$ .

We next introduce and define the graphs we will use as bases for our measures and then we discuss the measures themselves.

#### 3.2 Geometric Graphs

A graph  $G = \{V, E\}$  is a collection of a set of vertices  $V$  and a set of edges  $E$ . An edge  $e(v, w)$ , with  $e \in E$  and  $v, w \in V$ , is an unordered pair of vertices. A geometric graph  $G^* = [f(V), g(E), S]$  is an embedding of a graph in a metric space  $S$  that maps vertices to points and edges to line segments connecting pairs of points. We will omit the asterisk in the rest of this paper and assume all our graphs are geometric. We will also restrict our candidates to geometric graphs that are:

- *undirected* (edges consist of unordered pairs)
- *simple* (no edge pairs a vertex with itself)
- *planar* (there is an embedding in  $\mathbf{R}^2$  with no crossed edges)
- *straight* (embedded edges are straight line segments)
- *finite* ( $V$  and  $E$  are finite sets)

See [12] for alternatives.

Figure 2 shows instances of the geometric graphs on which we will compute our measures. The points are taken from one of the cells in the abalone SPLOM. In this section, we define the geometric graphs that are the bases for our measures.

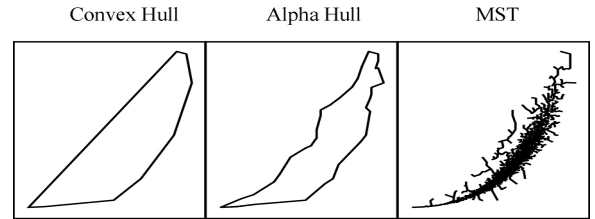


Figure 2: Graphs used as bases for computing scagnostics measures

##### 3.2.1 Convex Hull

A *polygon* is a closed planar region with  $n$  vertices and  $n-1$  faces. The *boundary* of a polygon can be represented by a geometric graph whose vertices are the polygon vertices and whose edges are the polygon faces. A *hull* of a set of points  $X$  in  $\mathbf{R}^2$  is a collection of one or more boundaries of polygons that have a subset of the points in  $X$  for their vertices and that collectively contain all the points in  $X$ . This definition includes entities that range from the boundary of a single polygon to a collection of boundaries of polygons each consisting of a single point.

We now define the convex hull. A set of points  $X$  in  $\mathbf{R}^2$  is *convex* if it contains all the straight line segments connecting any pair of its points. The *convex hull* of  $X$  is the boundary of the intersection of all convex sets containing  $X$ . This definition implies that the convex hull is the boundary of a convex polygon and that its vertices consist of points in  $X$ .

There are several algorithms for computing the convex hull [36]. Since the convex hull consists of the outer edges of the Delaunay triangulation, we can use an algorithm for the Voronoi/Delaunay problem and then pick the outer edges. Its computation thus can be  $O(n \log(n))$ . We will use the convex hull, together with other graphs, to construct measures of convexity.

### 3.2.2 Nonconvex Hull

A *nonconvex hull* is a hull that is not the convex hull. This class includes simple shapes like a *star convex* or *monotone convex* hull [3], but it also includes some space-filling, snaky objects and some that have disjoint parts. In short, we are interested in a general class of nonconvex shapes.

There have been several geometric graphs proposed for representing the “shape” of a set of points  $X$  on the plane. Most of these are proximity graphs [23]. A *proximity graph* (or *neighborhood graph*) is a geometric graph whose edges are determined by an indicator function based on distances between a given set of points in a metric space. To define this indicator function, we use an open disk  $D$ . We say  $D$  *touches* a point if that point is on the boundary of  $D$ . We say  $D$  *contains* a point if that point is in  $D$ . We call the smallest open disk touching two points  $D_2$ ; the radius of this disk is half the distance between the two points and the center of this disk is halfway between the two points. We call an open disk of fixed radius  $D(r)$ . We call an open disk of fixed radius and centered on a point,  $D(p, r)$ .

We considered several candidates for a shape-revealing proximity graph on a set of planar points. Some of them are subsets of the Delaunay triangulation:

- In a *Delaunay graph*, an edge exists between any pair of points that can be touched by an open disk  $D$  containing no points. The external edges of the Delaunay triangulation are the convex hull. This means that the external edges of the Delaunay graph cannot be used to represent non-convex hulls.
- In a *Gabriel graph* [15], an edge exists between any pair of points that have a  $D_2$  containing no points.
- In a *relative neighborhood graph* [39], an edge exists between any pair of points  $p$  and  $q$  for which  $r$  is the distance between  $p$  and  $q$  and the intersection of  $D(p, r)$  and  $D(q, r)$  contains no points. This intersection region is called a *lune*.
- In an *alpha shape graph* [10], an edge exists between any pair of points that can be touched by an open disk  $D(\alpha)$  containing no points.

Others are not subsets of the Delaunay:

- In a *distance graph*, an edge exists between any pair of points that both lie in a  $D(r)$ . The radius  $r$  defines the size of the neighborhood. This graph is not always planar and is therefore not a subset of the Delaunay.
- In a *k-nearest neighbor graph* (KNN), a directed edge exists between a point  $p$  and a point  $q$  if  $d(p, q)$  is among the  $k$  smallest distances in the set  $\{d(p, j) \mid 1 \leq j \leq n, j \neq p\}$ . Most applications restrict KNN to a simple graph by removing self loops and edge weights. If  $k = 1$ , this graph is a subset of the MST. If  $k > 1$ , this graph is not always planar.
- In a *sphere of influence graph* [5], an edge exists between a point  $p$  and a point  $q$  if  $d(p, q) \leq d_{nn}(p) + d_{nn}(q)$ , where  $d_{nn}(\cdot)$  is the nearest-neighbor distance for a point.

- A *beta skeleton graph* [25] is a close relative of the relative neighborhood graph. It uses a lune whose size is determined by a parameter  $\beta$ .

We have chosen the alpha hull for deriving our measures of shape. An *alpha hull* is a nonconvex hull derived from the boundary of an alpha shape graph. There are several reasons behind our choice. First, the alpha shape is relatively efficient to compute because it is a subset of the Delaunay triangulation with a simple inclusion criterion. Second, it is an erosion method. This aspect of the algorithm suits statistical data because we are interested in reducing the influence of outlying points on a shape.

A problem with using the alpha hull for a shape detector is choosing the value of the  $\alpha$  parameter (the radius of the disk that is used to delete edges in the Delaunay triangulation). We choose  $\alpha$  to be the value of the  $\omega$  parameter (discussed below). The  $\omega$  parameter is the cutoff value for identifying outlying edges in the minimum spanning tree. We choose this value because we wish to erase edges in the Delaunay triangulation that are longer than outlying edges in the original MST.

### 3.2.3 Minimum Spanning Tree

A *tree* is an acyclic, connected, simple graph. A *spanning tree* is an undirected graph whose edges are structured as a tree. A *minimum spanning tree* (MST) is a spanning tree whose total length (sum of edge weights) is least of all spanning trees on a given set of points [24]. The edge weights of a geometric MST are computed from distances between its vertices.

The MST is a subgraph of the Delaunay triangulation. There are several efficient algorithms for computing an MST for a set of points in the plane [28], [31].

## 3.3 Measures

We are interested in assessing five aspects of scattered points: *outliers*, *shape*, *trend*, *density*, and *coherence*. Our measures are derived from several features of geometric graphs:

- The length of an edge,  $length(e)$ , is the Euclidean distance between its vertices.
- The length of a graph,  $length(G)$ , is the sum of the lengths of its edges.
- A *path* is a list of vertices such that all pairs of adjacent vertices in the list are edges.
- A path is *closed* if its first and last vertex are the same.
- A closed path is the boundary of a *polygon*.
- The perimeter of a polygon,  $perimeter(P)$ , is the length of its boundary.
- The area of a polygon,  $area(P)$  is the area of its interior.
- The diameter of a graph,  $diameter(G)$ , is the longest shortest path in  $G$ .

All our measures are defined to be in the closed unit interval. To compute them, we assume our variables are scaled to the closed unit interval as well.

### 3.3.1 Outliers

Tukey [40] introduced the use of the peeled convex hull as a measure of the *depth* of a level set imposed on scattered points. For points on the 1D line, this amounts to successive symmetric trimming of extreme observations. Tukey's idea can be used as an outlier identification procedure. We compute the convex hull, delete points on the hull, compute the convex hull on the remaining points, and continue until (one hopes) the contours of successive hulls do not substantially differ.

We have taken a different approach. Because we do not assume that our point sets are convex (that is, comparably dense in all sub-regions of the convex hull), we cannot expect outliers will be on the edges of a convex hull. They may be located in interior, relatively empty regions. Consequently, we have chosen to peel the MST instead of the hull. We consider an outlier to be a vertex with degree 1 and associated edge weight greater than  $\omega$ .

There are theoretical results on the distribution of the largest edge for an MST on normally distributed data [30], but we decided to work with a nonparametric criterion for simplicity. Following Tukey [41], we choose

$$\omega = q_{75} + 1.5(q_{75} - q_{25})$$

where  $q_{75}$  is the 75th percentile of the MST edge lengths and the expression in the parentheses is the *interquartile range* of the edge lengths.

- Outlying

This is a measure of the proportion of the total edge length due to extremely long edges connected to points of single degree.

$$c_{outlying} = \text{length}(T_{outliers}) / \text{length}(T)$$

### 3.3.2 Shape

The shape of a set of scattered points is our next consideration. We are interested in both topological and geometric aspects of shape. We want to know, for example, whether a set of scattered points on the plane appears to be connected, convex, inflated, and so forth. Of course, scattered points are by definition *not* these things, so we are going to need additional machinery (based on our graphs that we fit to these points) to allow us to make such inferences. The measures that we propose will be based on these graphs.

In the formulas below, we use  $H$  for the convex hull,  $A$  for the alpha hull, and  $T$  for the minimum spanning tree. In our shape calculations, we ignore outliers.

- Convex

This is the ratio of the area of the alpha hull and the area of the convex hull. This ratio will be 1 if the nonconvex hull and the convex hull have identical areas.

$$c_{convex} = \text{area}(A) / \text{area}(H)$$

- Skinny

The ratio of perimeter to area of a polygon measures, roughly, how skinny it is. We use a corrected and normalized ratio so that a circle yields a value of 0, a square yields 0.12 and a skinny polygon yields a value near one.

$$c_{skinny} = 1 - \sqrt{4\pi \text{area}(A) / \text{perimeter}(A)}$$

- Stringy

A stringy shape is a skinny shape with no branches. The stringy measure is based on the  $\pi$  index, which is the ratio of width to length of a network. If the longest shortest path through a minimum spanning tree is almost as long as the sum of all the edges in the tree itself, then the tree is path-like or stringy.

$$c_{stringy} = \text{diameter}(T) / \text{length}(T)$$

- Straight

The *spanning ratio* or *dilation* of a geometric graph is the maximum ratio (over all pairs of vertices) of the shortest path (geodesic) between two vertices and the Euclidean distance between same vertices. Because it jumps to nearest neighbors, the MST has a dilation of  $O(n)$  [11]. For a perfectly linear geometric graph (all points in *general position*), the MST has a spanning ratio of 1.

We invert this ratio so that a straight graph has a value of 1 and other graphs have a smaller value. We modify this ratio further, however. Our ratio is the Euclidean distance between the points at the ends of the longest shortest path of the MST divided by the longest shortest path itself (the diameter of the MST). Even if the Pearson correlation is zero, our straightness index will be large for a set of points lying near a straight line. We will leave it to our monotonicity index to pick up scatters that are straight *and* highly correlated. Thus, our measure is

$$c_{straight} = \text{dist}(t_j, t_k) / \text{diameter}(T)$$

where  $t_j$  and  $t_k$  are the vertices in  $T$  on which the diameter is defined.

### 3.3.3 Trend

The following index helps reveal whether a given scatter is monotonic.

- Monotonic

If a set of scattered points is functional on  $x$  (plus error), a monotonicity coefficient should assess whether that function is monotonic. We have chosen the squared Spearman correlation coefficient, which is a Pearson correlation on the ranks of  $x$  and  $y$  (corrected for ties). We square the coefficient to accentuate the large values and remove the distinction between negative and positive coefficients. We assume investigators are most interested in strong relationships, whether negative or positive.

$$c_{monotonic} = r_{spearman}^2$$

This is our only coefficient not based on a subset of the Delaunay graph. Because it requires a sort, its computation is  $O(n \log(n))$ .

### 3.3.4 Density

The following indices detect different distributions of points.

- Skewed

The distribution of edge lengths of a minimum spanning tree gives us information about the relative density of points in a scattered configuration. Some have used the sample

mean, variance, and skewness statistics to summarize this edge length distribution, *e.g.*, [1]. However, theoretical results [37], [30] show that the MST edge-length distribution for many types of point scatters can be approximated by an extreme value distribution with fewer parameters. Other theoretical results [17] suggest that the mean MST edge length for a geometric graph on the unit square depends more on the number of points than on the distribution of points. Our Monte Carlo simulations using the distributions in Figure 3 found little variation in mean MST edge length for fixed  $n$ . By contrast, the skewness of the histograms of the MST edge length distributions for these points varied considerably. Consequently, we use a simple measure of skewness based on a ratio of quantiles of the edge lengths.

$$c_{skew} = (q_{90} - q_{50}) / (q_{90} - q_{10})$$

This statistic is relatively robust to outliers.

- **Clumpy**

An extremely skewed distribution of MST edge lengths does not necessarily indicate clustering of points. For this, we turn to another measure based on the MST: the Hartigan and Mohanty RUNT statistic [20]. This statistic is most easily understood in terms of the single-linkage hierarchical clustering tree called a *dendrogram*. The runt size of a dendrogram node is the smaller of the number of leaves of each of the two subtrees joined at that node. Since there is an isomorphism between a single-linkage dendrogram and the MST [18], we can associate a runt size ( $r_j$ ) with each edge ( $e_j$ ) in the MST, as described by Stuetzle [38]. The runt graph ( $R_j$ ) corresponding to each edge is the smaller of the two subsets of edges that are still connected to each of the two vertices in  $e_j$  after deleting edges in the MST with lengths less than  $length(e_j)$ .

Our runt-based measure emphasizes clusters with small intra-cluster distances relative to the length of their connecting edge and ignores runt clusters with relatively small runt size.

$$c_{clumpy}(T) = \max_j \left[ 1 - \max_k [length(e_k)] / length(e_j) \right]$$

where  $j$  indexes edges in the MST and  $k$  indexes edges in each runt set derived from an edge indexed by  $j$ .

### 3.3.5 Coherence

We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree. Smooth algebraic functions, time series, and curves (*e.g.*, spirals) fit this definition. So do points arranged in flows or vector fields.

There are numerous diagnostic measures for time series. Most of these involve conditioning on a model and examining residuals for lack of independence. For stochastic series, diagnostics usually assume stationarity in order for the tests to be valid. Stationarity implies that the mean, variance and autocorrelation structure of a process are defined and do not change over time. Verifying this can be problematic and time consuming.

We have chosen an alternative approach by devising a robust measure of coherence. This measure is not intended to be a time series detector. Not every time series is smooth. For example, moving average processes such as MA(1) may oscillate at frequencies that are higher than the resolution of the

mesh on  $x$ . Nor will it detect large negative autocorrelation structures, for similar reasons. Instead, we expect this measure to have large values when a series is relatively smooth. Furthermore, our measure does not assume a path is single-valued on  $x$ .

- **Striated**

The bottom scatterplot in Figure 3 shows points on parallel lines. We could recognize this pattern with a Hough transform [22]. Other configurations of points that represent vector flows or striated textures might not follow linear paths, however. We are interested in a more general measure. Recognizing that almost all of the adjacent edges in the MST on these points are collinear. Graham Wills (personal communication) proposed the following measure, which sums angles over all adjacent edges. Let  $V^{(2)} \subseteq V$  be the set of all vertices of degree 2 in  $V$ . Then

$$c_{striate}(T) = \frac{1}{|V^{(2)}|} \sum_{v \in V^{(2)}} |\cos \theta_{e(v,a)e(v,b)}|$$

## 3.4 Binning

We use hexagon binning [8] to improve performance. We begin with a 40 by 40 hexagon grid for each scatterplot. If there are more than 250 nonempty cells, we reduce the bin size by half and rebin. We continue this process until there are no more than 250 nonempty cells.

We examined by Monte Carlo simulation the effect of binning on our measures. Because we designed the measures to minimize this effect, the only one that showed a substantial trend with bin size was the *stringy* index. Not surprisingly, substantially larger bin sizes produce a higher index. Small edges are lost in larger bins, so the denominator of the index decreases with increase in bin size. We tried several adjustments to the index, including estimating the within-bin size of edges and adding this estimate to the denominator. All of the adjustments we tried substantially reduced the sensitivity of the measure, so we decided to leave the *stringy* index unadjusted.

## 4 PERFORMANCE

Figure 3 shows the results of our graph-theoretic measures applied to 11 different scatter patterns. We notice that the red or orange rectangles align with patterns the respective measures are designed to flag.

We can get a graphical idea of the sensitivity and specificity of the measures by examining the first two principal components of these measures applied to these patterns. Figure 4 shows a biplot [16] of the measures and patterns. We find that the measures fill the space and are near the patterns they target. The *clumpy* measure has a short vector in the plot because it loads heavily on a third component. The scree (plot of eigenvalues against factor indices) is smoothly descending with no clear elbow, so the dimensionality of these measures on the selected point sets is not low.

### 4.1 Time

Because we use binning and efficient triangulation, computation time is roughly  $O(np^2)$ . On a Macintosh G4 PowerBook running Java 1.4.2, computing the measures on 100,000 random cases distributed uniformly on 10 variables requires approximately 10 seconds. Computing the measures on 100,000 cases and 25 variables requires approximately 75 seconds. Large, uniformly distributed

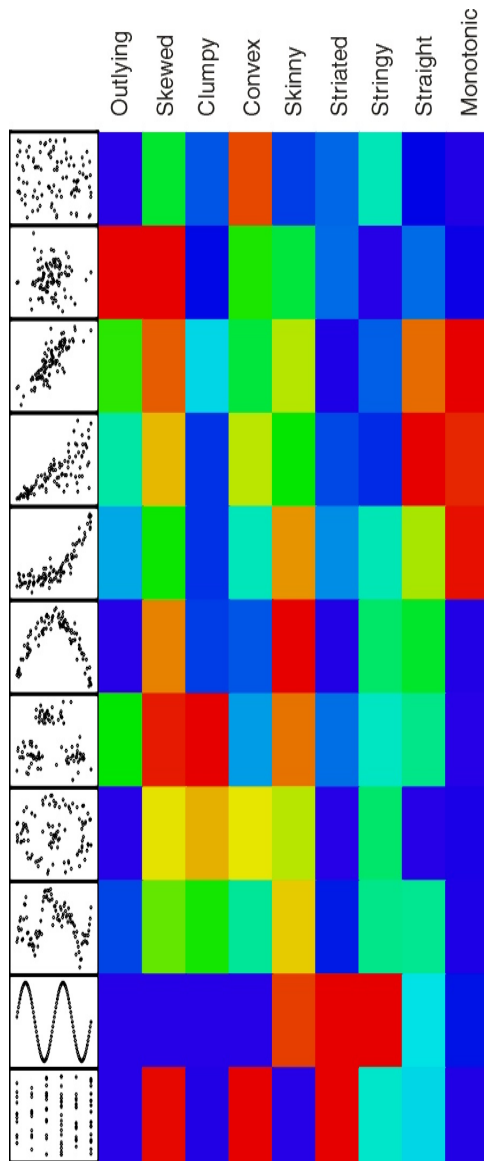


Figure 3: Scaled graph-theoretic measures (blue=low, red=high) for eleven scatter patterns

point sets tend to increase computation time because of the need for rebinning. Datasets with compact, nonuniform distributions are considerably faster to compute.

## 5 EXAMPLES

Figure 5 shows a scagnostics SPLOM for the Abalone dataset. We have highlighted a point in the SPLOM that represents one unusual scatterplot. The linked plot is shown in the upper right of the figure. From what we see in the SPLOM, we can characterize this scatterplot as relatively high in outliers, skewed in edge lengths, clumpy, nonconvex, striated, and stringy.

Figure 6 shows a scagnostics SPLOM of 17 variables from a dataset based on statistics for selected countries compiled by the World Health Organization and the United Nations [43]. We have highlighted a point in the graph-theoretic SPLOM to give an example of the type of anomalies that scagnostics can uncover. The red point is clearly an outlier in most of the panels. The scatterplot

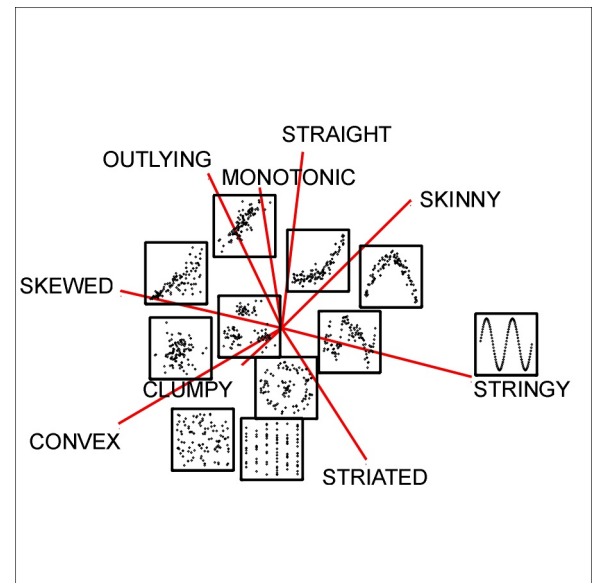


Figure 4: Biplot of measures and scatters

corresponding to this point is shown in the upper right of the figure. We were surprised to find that we had included two versions of a categorical variable measuring the type of government – one recoded to have fewer categories. Because these two artifactually-related variables are not perfectly correlated, they would have not have induced a singularity in a statistical analysis and might have been undetected in a thoughtless automated data mining.

Figure 7 shows a scagnostics SPLOM of 62 variables from a microarray dataset [2]. The SPLOM shows the high degree of homogeneity in the 2D marginal distributions of the 62 cell lines in the array. The linked scatterplot shows a typical 2D distribution, evidently not bivariate normal. The authors identified clusters of non-cancerous and cancerous cells in these data. We must remember that lack of evidence for clusters in the 2D scatters is not evidence for lack of clusters in higher dimensions. Instead, the scagnostic SPLOM lends some support to the authors' findings by downweighting the possibility that clusters might be due to measurement artifacts (*e.g.*, mixing discrete and continuous variables in the analysis). This application illustrates the appropriate focus of scagnostics as a preliminary screening method.

Figure 8 shows a SPLOM of weather data. The data comprise hourly meteorological measurements over a year at the Greenland Humboldt automatic weather station operated by NASA and NSF. These measurements are part of the Greenland Climate Network (GC-Net) sponsored by these federal agencies. We have sorted the variables in the SPLOM using the size of the loadings on the first principal component of the scagnostic measures. The sorting clearly segregates the discrete and continuous variables and clusters similar marginal 2D distributions.

## 6 CONCLUSION

As we have seen, scagnostics offers the possibility of detecting anomalies in large scatterplot matrices. There is more to do, however. First, we can construct multivariate models from these and other graph-theoretic measures and apply them to the more general pattern detection problem in high-dimensional space. This approach follows those taken by the manifold learning community [32], [4], [6]. Second, we can use these methods to sort scatterplot matrices in order to make them more accessible to lensing and

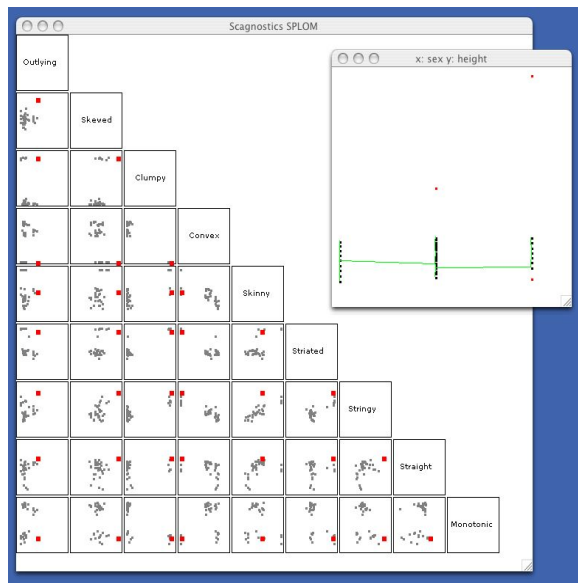


Figure 5: Scagnostics SPLOM of Abalone measurements

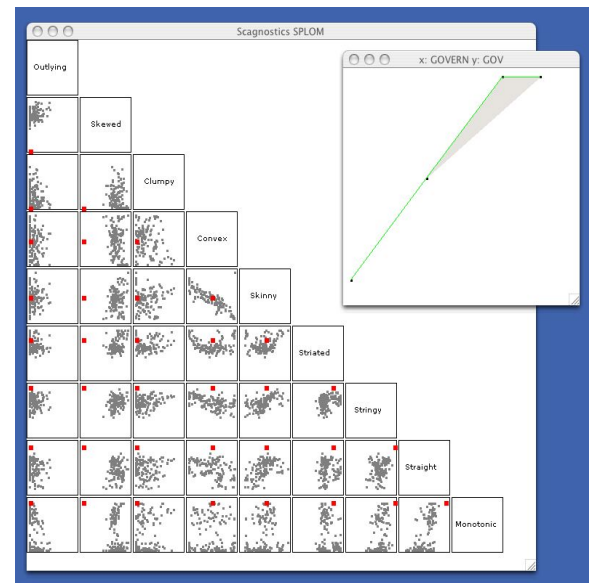


Figure 6: Scagnostics SPLOM of world countries data

other focus methods [14], [26], [29], [34].

## REFERENCES

- [1] C. Adami and A. Mazure. The use of minimal spanning tree to characterize the 2d cluster galaxy distribution. *Astronomy & Astrophysics Supplement Series*, 134:393–400, 1999.
- [2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96:6745–6750, 1999.
- [3] E. M. Arkin, Y-J Chiang, M. Held, J. S. B. Mitchell, V. Sacristan, S. Skiena, and T-H Yang. On minimum-area hulls. *Algorithmica*, 21(1):119–136, 1998.
- [4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.
- [5] E. Boyer, L. Lister, and B.L. Shader. Sphere-of-influence graphs using the sup-norm. *Mathematical and Computer Modelling*, 32:1071–1082, 2000.
- [6] M. Brand. Nonlinear dimensionality reduction by kernel eigenmaps. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 547–552, 2003.
- [7] A. Buja and P. Tukey (Eds.). *Computing and Graphics in Statistics*. Springer-Verlag, New York, 1993.
- [8] D. B. Car, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436, 1987.
- [9] W. S. Cleveland. *The Elements of Graphing Data*. Hobart Press, Summit, NJ, 1985.
- [10] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29:551–559, 1983.
- [11] D. Eppstein. Spanning trees and spanners. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 425–461. North-Holland, Amsterdam, 2000.
- [12] J. H. Friedman and L. C. Rafsky. Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 11:377–391, 1983.
- [13] J. H. Friedman and W. Stuetzle. John W. Tukey’s work on interactive graphics. *The Annals of Statistics*, 30:1629–1639, 2002.
- [14] M. Friendly and E. Kwan. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4):509–539, 2003.
- [15] K. R. Gabriel and R. R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18:259–278, 1969.
- [16] K.R. Gabriel. The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58:453–467, 1971.
- [17] J. Gao and J. M. Steele. Sums of squares of edge lengths and spacefilling curve heuristics for the traveling salesman problem. *Siam Journal on Discrete Mathematics*, 7:314–324, 1994.
- [18] J. C. Gower and G. J. S. Ross. Minimal spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.
- [19] J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4:187–213, 1975.
- [20] J. A. Hartigan and S. Mohanty. The runt test for multimodality. *Journal of Classification*, 9:63–70, 1992.
- [21] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.
- [22] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.
- [23] J. Jaromczyk and G. Toussaint. Relative neighborhood graphs and their relatives, 1992.
- [24] J.B. Kruskal Jr. On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.
- [25] D. G. Kirkpatrick and J. D. Radke. A framework for computational morphology. In G. Toussaint, editor, *Computational Geometry*, pages 217–248. North-Holland, Amsterdam, 1985.
- [26] A. MacEachren, X. Dai, F. Hardisty, D. Guo, and G. Lengerich. Exploring high-d spaces with multiform matrices and small multiples. In *Proceedings of the IEEE Information Visualization 2003*, pages 31–38, 2003.
- [27] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. The population biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the north coast and islands of Bass Strait. Technical report, Sea Fisheries Division, 1994.
- [28] J. O’Rourke. *Computational Geometry in C (2nd ed.)*. Cambridge University Press, Cambridge, 1998.
- [29] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Information Visualization 2004*, pages 89–96, 2004.
- [30] M. D. Penrose. Extremes for the minimal spanning tree on normally distributed points. *Advances in Applied Probability*, 30:628–



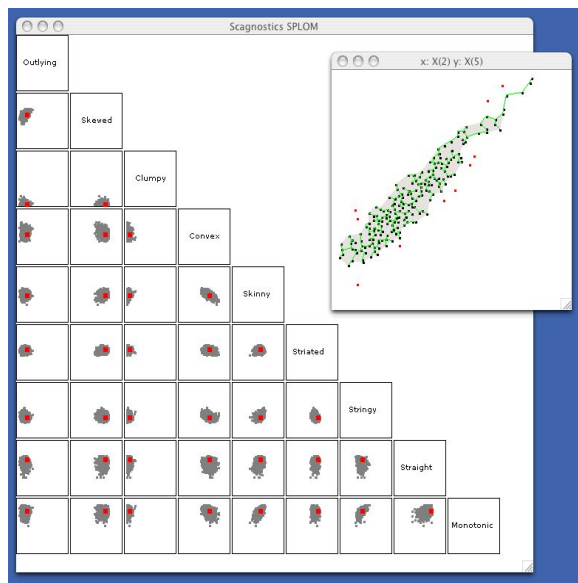


Figure 7: Scagnostics SPLOM of microarray data

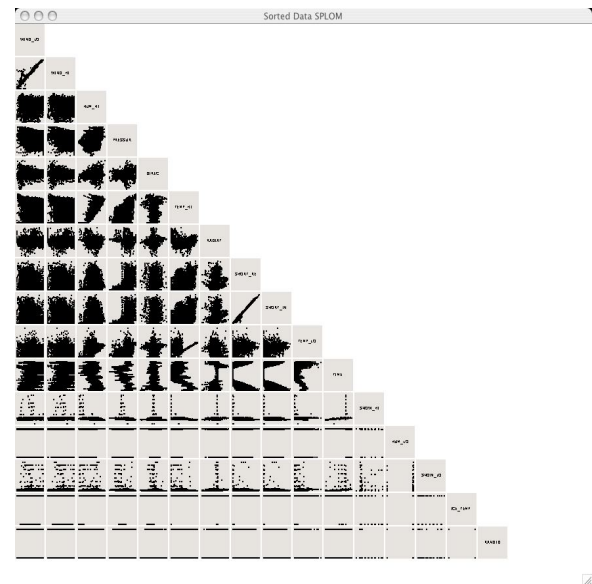


Figure 8: Sorted SPLOM of wind data

- 639, 1998.
- [31] F. P. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, New York, 1985.
  - [32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
  - [33] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, 1992.
  - [34] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proceedings of the IEEE Information Visualization 2004*, pages 65–72, 2004.
  - [35] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York, 1986.
  - [36] Steven S. Skiena. *The Algorithm Design Manual*. Springer-Verlag, New York, 1998.
  - [37] J. M. Steele. Growth rates of Euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, 16:1767–1787, 1988.
  - [38] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.
  - [39] G. Toussaint. The relative neighborhood graph of a finite planar set. *Pattern Recognition*, 12:261–268, 1980.
  - [40] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, pages 523–531, Vancouver, Canada, 1974.
  - [41] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1977.
  - [42] J. W. Tukey and P.A. Tukey. Computer graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85*, Fairfax, VA, United States, 1985. National Computer Graphics Association.
  - [43] L. Wilkinson. *The Grammar of Graphics*. Springer-Verlag, New York, 1999.