Feature network models for proximity data : statistical inference, model selection, network representations and links with related models

Frank, L.E.

Chapter 5

Network Representations of City-Block Models [1]

Abstract

City-block models for similarity always allow network representations that reproduce the same distances as the unique coordinate representation. A rule to construct such networks is given, based on additivity of city-block distances across sequences of intermediate points along monotonic trajectories in space. The paper also defines the concept of internal node, which helps in reducing the complexity of networks and in making them better interpretable. The general graph construction rule and definition of internal nodes also apply to the distinctive features model, the common features model (additive clustering), as well as to hierarchical trees, additive trees, and extended trees. Additivity is the key property that makes the city-block metric so versatile and causes a basic unity of dimensional, hierarchical and featural representations of similarity.

5.1 Network representations of city-block models

The city-block distance rule has been under consideration in psychology as a plausible model for similarity and difference for a long time (Arabie, 1991; Attneave, 1950; MacKay, 2001; Micko & Fischer, 1970; Nosofsky, 1984; Shepard, 1964). It has been used not only for human perception (Borg & Leutner, 1983; Garner, 1974; Shepard, 1987), but also for category learning (Kruschke, 1992; Zaki, Nosofsky, Stanton, & Cohen, 2003), color vision and pattern recognition in honeybees (Backhaus, Menzel, & Kreißl, 1987; Ronacher, 1992), as well as for perception of electric properties of objects by weakly electric fish (Emde & Ronacher, 1994). The model has caused a flux of technical papers concerned with the computational complications that arise when trying to fit city-block distances to error-contaminated (dis)similarity data (Brusco, 2001, 2002; Eisler, 1973; Eisler & Roskam, 1977; Groenen & Heiser, 1996; Groenen,

[1] This chapter has been submitted for publication as: Heiser, W. J. & Frank, L. E. (2005). Network representations of city-block models. *Submitted manuscript.*

Heiser, & Meulman, 1998, 1999; Heiser, 1989, 1991; Hubert & Arabie, 1988; Hubert, Arabie, & Hesson-Mcinnis, 1992; Okada & Imaizumi, 1980). There are other unsolved technical problems; for example, degeneracies in nonmetric multidimensional scaling with all distances tied into only two values are more prevalent in the city-block metric (and in the dominance metric) than in other Minkowski metrics (Shepard, 1974). In this paper, we leave these technical issues aside, and focus primarily on some theoretical properties that lead to equivalent representations of the city-block model.

Substantively, the city-block model has played a major role in the classic distinction between integral and separable stimulus dimensions, which is an essential consideration in most current experimental and theoretical analyses of category learning (Ashby & Maddox, 1990; Goldstone, 1994; Kruschke, 1992; Melara, Marks, & Lesko, 1992; Nosofsky, 1992). Shepard (1964) reported two experiments specifically designed to test if the metric of psychological space depends on the perceptual analyzability of the stimuli, and found that for objects differing in size and angle of orientation the city-block distance gave a better account of subjective judgments of similarity and objective measures of generalization than the Euclidean distance. Together with results on category learning (Shepard & Chang, 1963; Shepard, Hovland, & Jenkins, 1961), these findings also demonstrated a fundamental role of selective attention for analyzable stimuli (Shepard, 1991). This line of research culminated in the generalized context model for category learning and attention allocation (Nosofsky, 1984, 1986, 1987, 1992; Nosofsky & Zaki, 2002; Zaki et al., 2003).

Closely connected to the integrality-separability distinction is the uniqueness of the coordinate system. In the words of Attneave,

> "One possible hypothesis would be that the psychological dimensions are related like physical dimensions in Euclidean space. Another would be that differences along different dimensions combine additively, in which case composite judgments would be predicted by a multiple linear regression equation. Perhaps the most significant psychological difference between these two hypotheses is that the former assumes one frame of reference to be as good as any other, whereas the latter implies a unique set of psychological axes." (Attneave, 1950, p. 555).

One way to distinguish between integral and separable dimensions is to establish whether a stimulus is more readily associated with another stimulus that is close to it in the Euclidean metric or with one that may be farther away but matches it on some pre-determined dimension. Various other converging operations have been used to distinguish between these two types of dimensions (Garner, 1974). The unique coordinate system of the city-block metric has also motivated other utilizations. Buja and Swayne (2002) used dimensional uniqueness to identify an orientation of Euclidean solutions, which are rotationally invariant. Heiser (1989) used dimensional uniqueness as an argument to develop an individual differences city-block model with dimension weighting.

Nevertheless, uniqueness and additivity of city-block dimensions do not tell us what structural relations are valid in the whole space. For example, uniqueness and additivity do not tell us if the stimuli are clustered or not, whether two stimuli are

close neighbors or not, and whether three stimuli have the same order on all dimensions or not. It is remarkable that in applications of the city-block model, there has not been much attention for actual representations. Some authors do not even show or list the coordinates; they only report tests of inter-dimensional additivity, or the relative goodness-of-fit (Melara et al., 1992; Ronacher, 1992; Emde & Ronacher, 1994). One reason for this lack of attention for coordinates might be that the psychological dimensions are supposed to be monotonic with physical dimensions present in the stimuli, so that the order of the coordinates is known by design. However, inter-dimensional additivity does not preclude the possibility that stimulus differences along one dimension change for different levels of the other dimension, i.e., that they show non-linear structural relations (as will become clear in an example of similarity between rectangles that is discussed in the following).

The present study was triggered by the notion that a simple and direct way to describe structural relations between objects is to draw a network, with nodes (or vertices) for the stimuli and with lines (or edges) indicating local connections between neighbors. Distance in a network is the length of the (shortest) path traveled. If the stimuli are clustered, we expect to find fully connected subsets, or cliques. If three stimuli have the same order on all dimensions, we expect to find segmented pathways without sharp turns. If there is interaction between dimensions, we expect a nonlinearly distorted grid, and so on. Would it be possible to use the rectangular grid that is so characteristic for the city-block metric and just connect all pairs of points on the grid whenever there is no other point lying between them, and finish by dropping the rest of the grid? Would it still be possible to reconstruct the distance correctly if we replaced all city-block corners by direct straight lines?

It turns out that it is indeed possible to develop a universal network representation of city-block models that applies regardless of the dimensionality of the coordinate space. This paper first describes the key elements of the network construction method, which are the concepts of *betweenness*, *metric segment*, and *metric-segmental additivity*. Since a network is just a collection of nodes and lines, one needs some embedding to be able to draw it, but the details of this embedding are of secondary importance. While the network is the model, an embedding is one of several possible maps of it. The paper also introduces the possibility of including an additional set of points corresponding to hypothetical stimulus objects, called *internal nodes*. An example of the perception of rectangles will demonstrate their use. It is shown that networks throw new light on a puzzling characteristic of the city-block model, the occurrence of partial isometries. Next, the same theory is applied to the Goodman-Restle symmetric set difference, a special case of the city-block metric, with binary dimensions called distinctive features. This framework contains a rather large class of discrete models for similarity data, including additive similarity trees (Buneman, 1971; Sattath & Tversky, 1977), extended similarity trees (Corter & Tversky, 1986), the additive clustering or common features model (Carroll & Arabie, 1983; Shepard & Arabie, 1979), and a new set-partitioning model with unicities called the *double star tree*. It is shown that the same network construction rule recovers the familiar additive tree graph, and yields new graphical representations for the other models. These are illustrated with several examples of similarity data known from the literature.

5.2 General theory

The discussion starts with the fundamental notion of betweenness in continuous spatial models, and demonstrates how it leads to additivity of distance. Betweenness in more dimensions requires the concept of a metric segment, which is the area between a pair of points that contains all intermediate points for which distance is additive. Then a network representation is formed from a complete network by elimination of lines when additivity applies. Some properties of this representation are discussed with an example of similarity between rectangles. Next, the concept of an internal node is introduced as a supplementary point located in the metric segment of any two objects, or in the intersection of several metric segments. This section concludes with a general characterization of partial isometry, a problematic phenomenon that is specific for the continuous city-block model.

Betweenness of points and additivity of distances

Geometric models like the city-block model consist of points arranged in some continuous space, among which we define distances according to a certain rule (or metric) to account for empirical relations between the experimental objects. An elementary structural property of spatial arrangements is the *betweenness relation*. In some situations, betweenness implies additivity of distance. Taking the simplest case, when we have three ordered points A, B, and C in one dimension, where B is between A and C, the distance between the outer points A and C is the sum of the distances from A to B and from B to C. In other words, when B is between A and C on a line, a condition called *intra-dimensional betweenness*, we have *intra-dimensional additivity*. It is easy to see that more generally, the distance between any two points on a line is equal to the sum of the lengths of the segments that one crosses when going from one to the other through a series of intermediate points.

In more than one dimension, the situation changes because it is a common characteristic of all metrics that distances satisfy the triangle inequality. Denoting the distance between two points A and B by $d(A, B)$, the triangle inequality states that $d(A, C) \leqslant d(A, B) + d(B, C)$. Therefore, going through a third point can only add to the distance. Even if an intermediate point B is between two others on all dimensions, in going from A to C the direct route is generally shorter than going via B. In Euclidean space, the only exception is when three points are located exactly in a one-dimensional subspace, in which special case the triangle inequality reduces to an equality (Torgerson, 1952). By contrast, in city-block space, the triangle equality is much more common, because betweenness in all city-block dimensions (a condition that we will call *metric-segmental betweenness*) always leads to additivity of distance.

We now demonstrate the particular result that under the city-block metric the triangle inequality reduces to an equality for any three points A, B, and C whenever B is between A and C on all dimensions (Busemann, 1955, p. 28). Let A have coordinate values z_{At}, for $t = 1, \cdots, T$ where T denotes the number of dimensions. The city-block distance between A and B is defined as the function

$$d(A, B) \quad = \quad \sum_t |z_{At} - z_{Bt}|. \tag{5.1}$$

The fact that $d(A, B)$ is built up as a sum of dimension-wise differences is called *inter-dimensional additivity* (Suppes, Krantz, Luce, & Tversky, 1989, section 14.4.3). For metric-segmental additivity to hold, the coordinates have to satisfy, for each dimension t, either $z_{At} \leqslant z_{Bt} \leqslant z_{Ct}$ or $z_{At} \geqslant z_{Bt} \geqslant z_{Ct}$ (*monotonicity*: all choices of z_{Bt} within the constrained area lead to monotonically increasing or monotonically decreasing sets of coordinate values). Under monotonicity we must have, for any t,

$$(z_{Ct} - z_{At}) = (z_{Ct} - z_{Bt}) + (z_{Bt} - z_{At}), \tag{5.2}$$

$$|z_{At} - z_{Ct}| = |z_{At} - z_{Bt}| + |z_{Bt} - z_{Ct}|, \tag{5.3}$$

where the three terms in Equation 5.2 are either all positive or all negative, so that we can take absolute values and freely reverse the order of the arguments in Equation 5.3, which expresses intra-dimensional additivity for any dimension. Summing Equation 5.3 over t and using Equation 5.1 we obtain

$$\begin{aligned} \sum_t |z_{At} - z_{Ct}| &= \sum_t |z_{At} - z_{Bt}| + \sum_t |z_{Bt} - z_{Ct}|, \\ d(A, C) &= d(A, B) + d(B, C), \end{aligned} \tag{5.4}$$

that is, *metric-segmental additivity* of distance when we go from A to C via B. This result forms the basis of the network representations that we develop in this paper. Joly and Le Calvé (1994) have defined the general concept of a *metric segment* as the set of points $[AB]_{met} = \{M\colon d(A, B) = d(A, M) + d(B, M)\}$. The metric segment is a generalization of the line segment to multidimensional spaces. In Euclidean space, metric segments are still segments of lines, but in two-dimensional city-block space, they are rectangles with sides parallel to the axes. In three-dimensional city-block space, metric segments are *cuboids* (parallelepipeds with rectangular faces), and in more than three dimensions, *hypercuboids*. When dimension-wise differences are all equal, these structures reduce to squares, cubes and hypercubes.

The prevalence of metric-segmental additivity in city-block space simply expresses the fact that in this type of space, there is a multitude of paths through intermediate points covering exactly the same distance. As every passenger knows, there is a unique shortest route by air from city to city, but within any city where buildings are arranged in rectangular blocks one can reach distant destinations along several different routes that are equally long.

Network representation of city-block configurations

The surprising consequence of metric-segmental additivity is that it allows us to construct a model representation of city-block configurations that does not involve coordinate values. This coordinate-free representation consists of a set of *nodes* or *vertices* $\mathcal{V} = \{\nu_1, \cdots, \nu_i, \cdots, \nu_m\}$, representing the objects, and a set of line segments or *edges* $\mathcal{T} = \{\tau_1, \cdots, \tau_l, \cdots, \tau_L\}$, where $L \leqslant \frac{1}{2}m(m-1)$, connecting pairs of nodes. Each edge τ_l has a length q_l, collected in the set $\mathcal{Q} = \{q_1, \cdots, q_l, \cdots, q_L\}$, which indicates the distance between the corresponding pair of nodes. Thus, the triad $\mathcal{N} = \{\mathcal{V}, \mathcal{T}, \mathcal{Q}\}$ forms a valued graph or *network*. In a full, or *complete* network, we have $L = \frac{1}{2}m(m-1)$, that is, al $n = \frac{1}{2}m(m-1)$ pairs of nodes are connected

by an edge. Of course, in applications, we would like simplicity to prevail by aiming at an *incomplete network* with L as small as possible (Klauer & Carroll, 1989). If $L < (m-1)T$, the network is a more parsimonious parametrization than the coordinate space. Moreover, as we shall see shortly, there are other considerations that also can make a graphical representation attractive.

The construction of the network goes as follows. Given a city-block configuration $\mathcal{M} = \{\mathbf{Z}, \mathbf{D}\}$, where $\mathbf{Z}$ is a $m \times T$ matrix of coordinates z_{it}, with T the number of dimensions, and $\mathbf{D}$ a matrix of distances $d(Z_i, Z_j)$ between pairs of points Z_i and Z_j, the set $\mathcal{V}$ is formed by just allocating a separate node ν_i to each distinct Z_i. The set $\mathcal{T}$ is then formed by elimination. We start with a full list of edges, with the distances listed in some fixed order in $\mathcal{Q}$. For all triads of points Z_i, Z_j, and Z_k, we determine the quantity $W_{ik}^j = d(Z_i, Z_j) + d(Z_j, Z_k) - d(Z_i, Z_k)$. Then Z_j belongs to the metric segment $[X_i X_k]_{met}$ if $W_{ik}^j = 0$. When there is at least one Z_j for which $W_{ik}^j = 0$, we can drop the direct edge from ν_i to ν_k from the list $\mathcal{T}$, while keeping the direct edges from ν_i to ν_j and from ν_j to ν_k. In that case, we also omit the corresponding distance from the list $\mathcal{Q}$. Dropping the direct edge is possible since metric-segmental additivity transfers to additivity along the shortest path in the graph. Thus, we are able to use an interesting parallel between the coordinate space and the graphical space. While the city-block distance between a pair of points is equal to a sum of distances through a series of intermediate points in their metric segment, the graphical distance is equal to the sum of edge lengths in a shortest path that connects two nodes. When $W_{ik}^j \neq 0$ for all i, j, k, no edges are dropped and the complete network $\mathcal{N}$ trivially represents $\mathcal{M}$.

There is a caveat for the particular case in which two distinct objects, i and j, have the same location, $Z_i = Z_j$. Then $d(Z_i, Z_j) = 0$ and $d(Z_j, Z_k) = d(Z_i, Z_k)$, from which it follows that $W_{ik}^j = 0$, so that the direct edge between ν_i and ν_k is dropped. The same equalities also give $W_{jk}^i = 0$, with the effect that the direct edge between ν_j and ν_k is dropped. Consequently, two objects with the same location would become two nodes that are disconnected from all other nodes in the graph (isolates). By merging such objects into one node, which has the same distances to the other points, and again determining the relevant metric segments, the graph will generally become connected.

The graph $\mathcal{N}$ by itself is the desired network representation. However, to visualize or interpret $\mathcal{N}$, we must embed it again in some coordinate space. Note that we now have more freedom in choice of embedding, since the primary elements of interpretation are the connectivity and structural order relations between the nodes, while the exact length of the edges is secondary. The embedding may be in the original city-block space if it is two-dimensional, or in some other space with two dimensions. We could use a Euclidean embedding of the graph, obtained, for instance, by a nonmetric MDS method (*cf.* (Buja & Swayne, 2002)), or by a metric MDS with weights to down-weight the large graphical distances (Kamada & Kawai, 1989). Of course, we can reconstruct the original city-block distances $\mathbf{D}$ only if the edges included in the plot of the embedding are precisely those from the list $\mathcal{T}$, labeled with the edge lengths $\mathcal{Q}$. If we would use any other common procedure to draw lines

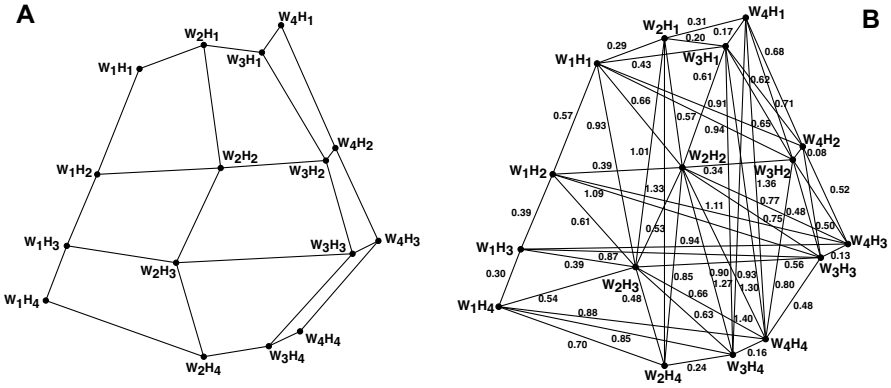

Figure 5.1: City-block solution in two dimensions for the *rectangle* data. The labels $W_1 - W_4$ indicate the width levels, and $H_1 - H_4$ the height levels of the stimulus rectangles.

between pairs of objects in an embedding, for example, by determining a threshold graph or a K-nearest neighbor graph (*cf.* Jain & Dubes, 1988, p. 60), reconstruction of the original distances by their counterparts in the graph is generally inaccurate.

To illustrate the graphical representation of a city-block configuration, we look at some data collected and analyzed by Borg and Leutner (1983). The stimuli used were 16 rectangles of varying width and height, where the two variables each had four levels increasing in equal steps. All 120 possible pairs of stimuli were presented twice, in random order, to 21 subjects, who had to rate each pair on a 10-point scale of dissimilarity. Reliability, calculated per subject as the product-moment correlation over the ratings of stimulus pairs in the two different orders, was 0.75 on average. The data, averaged over all subjects and replications, were analyzed in two dimensions[2] with the smoothing method for city-block multidimensional scaling described in Groenen et al. (1998). This method was specifically designed to avoid being trapped in local minima of the least squares MDS criterion. Figure 5.1 gives two versions of the two-dimensional solution. In both versions, the points are labeled with their width level and their height level. Thus, W_1H_1 (top-left) is the smallest rectangle, and W_4H_4 (bottom-right) the largest. In Figure 5.1A, we have connected the points with their direct neighbors by design; that is, lines connect rectangles differing one level on only one variable (as in Borg & Leutner, 1983, their Figure 3). It shows that the horizontal dimension roughly corresponds to width, the vertical dimension to height, and that the intervals tend to become smaller as the size of the rectangles increases, in both dimensions. Borg and Leutner predicted this nonlinear effect on psychophysical grounds; it was also present in their solution. However, contrary to their solution, the current solution also exhibits interaction: successive width intervals tend to become larger as height levels increase, although not uni-

[2]The fitting criterion used was least squares and metric, since Borg and Leutner (1983) reported that non-metric fitting showed a linear relationship between dissimilarity and distance.

formly.

Figure 5.1B gives the network representation, using the same coordinates. Here, two points are connected if there are no other points between them, that is, if their metric segment is empty. Recall that in a two-dimensional city-block solution, a metric segment has a rectangular shape, with orientation parallel to the horizontal and vertical axes. The two points spanning the metric segment are on a main diagonal of that rectangular area. This main diagonal is shown as a line in Figure 5.1B if the metric segment contains none of the other points. For example, a line connects W_1H_4 and W_2H_3, since they span an empty metric segment. One could also say that the two candidates in the design for being inside their metric segment, W_1H_3 and W_2H_4, are actually located outside of it. Thus, while Figure 5.1A emphasizes conformities of the solution with the design, Figure 5.1B highlights violations as well. Also, note that even though a path like $W_1H_1 - W_1H_2 - W_1H_3 - W_1H_4$ is in conformity with the design, the tilting to the right contradicts that these stimuli are of equal width. However, this contradiction shows up as a property of the spatial city-block solution with the horizontal dimension identified as width, not as a property of the network, which could have been plotted differently. Furthermore, a path like $W_1H_4 - W_2H_4 - W_3H_4 - W_4H_4$ is correctly monotonic in the horizontal direction; yet it has two subadditivities giving direct lines from W_1H_4 to W_3H_4 and W_4H_4. The total number of lines in Figure 5.1B is 56, while the spatial solution has 30 independent coordinates. Therefore, the network representation is not parsimonious, but it does enable a detailed analysis of structural relations in the data.

It might appear that the network representation is unduly complex, compared to the simplicity of the spatial representation, in which we just plot the coordinates. More specifically, the network seems to have the following unfavorable properties:

1. Some relatively long lines appear in Figure 5.1B, e.g. between W_2H_1 and W_2H_4 or between W_1H_3 and W_4H_3. By contrast, the attraction of other network models often is that they have global properties resulting from the action of local connections (short lines). Here, the long lines simply reflect that objects can be opposites on one dimension and direct neighbors on the other dimension.

2. Many nodes have high degree (number of lines that are incident with it, or number of nodes adjacent to it); for example, 11 lines emanate in Figure 5.1B from node W_2H_2, and 10 lines from W_2H_3, while the lowest degree still is 5 (for W_1H_3, W_1H_4, and W_2H_4). As can be seen from Figure 5.1A, the current design predicts nodes with degree 2, 3, or 4.

3. Many crossings of lines occur at locations where there is no intermediate node. For example, the line between W_2H_1 and W_2H_4 in Figure 5.1B crosses 15 other lines without meeting any other node. In the design of Figure 5.1A, these rectangles are connected via the much simpler three-segment path $W_2H_1 - W_2H_2 - W_2H_3 - W_2H_4$.

4. The total number of lines is large, 56. If we would consider each line length as a separate parameter, the network model absorbs many parameters, compared to the number of independent data values (120). However, it should be noted

that we actually fitted only $2 \times (m-1) = 30$ coordinate values, so the line lengths cannot be considered as independent quantities. In the design, we have only 24 lines, and under a model of no interaction, these line lengths would be further constrained to six independent parameters (3 width intervals and 3 height intervals).

Since these properties are also prevalent in other examples that we have analyzed but do not report in detail here, they seem to be a recurring and genuine characteristic of the present network representation. However, as shall become clear in the next section, there is a way to alleviate points 1-3, and to some extent point 4 as well, by introducing an additional set of points, called *internal nodes*, which also play an important role in special cases of the model.

It is of special interest to look at the one-dimensional case. If $\mathbf{D}_1$ is a distance matrix of points along a line, then any point lies between two others except for the endpoints, from which it follows immediately that the graph $\mathcal{N}$ has exactly $m-1$ edges. Assuming the rows and columns of $\mathbf{D}_1$ are ordered in the same way as the order of the points along the line, the edges of the graph correspond to the elements on the subdiagonal of $\mathbf{D}_1$. We find segmental additivity for all pairs of points (i, j) that are not consecutive. Specifically, any other element in the upper-right triangle of the distance matrix $\mathbf{D}_1$ is the sum of consecutive elements in the subdiagonal, starting with the subdiagonal element in the same row, and ending with the subdiagonal element in the same column. Hence, in this case the graph is a chain, which has graphical distances with exactly the same additivity structure as a set of points on a line. The chain can be displayed in many ways (for instance, as a curved, connected sequence of nodes in the plane), all of which give an equivalent reconstruction of the one-dimensional distances, as long as their edge lengths are equal to the subdiagonal elements of $\mathbf{D}_1$.

Internal nodes

It can be useful to add nodes to the network that do not correspond to the original set of points in the city-block configuration $\mathcal{M}$. These additional nodes are called *internal nodes*, and can be chosen in a number of ways. In general, adding one point to a network of m nodes leads to m additional edges in the network. Therefore, the introduction of the internal node should entail the possibility of dropping a number of edges, too. By placing the new point in a metric segment of a pair of existing points, the total number of edges reduces by one. Thus, the internal point could be chosen so that it is in the intersection of as many of the n metric segments as possible.

The case for which the greatest simplification occurs is an equal-distance configuration $\mathbf{Z}_0$, for four points defined as:

$$\mathbf{Z}_0 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{5.5}$$

It is not hard to verify that the city-block distances between all six pairs of points in $\mathbf{Z}_0$ are equal to 2, and that no point is in the metric segment of any other two

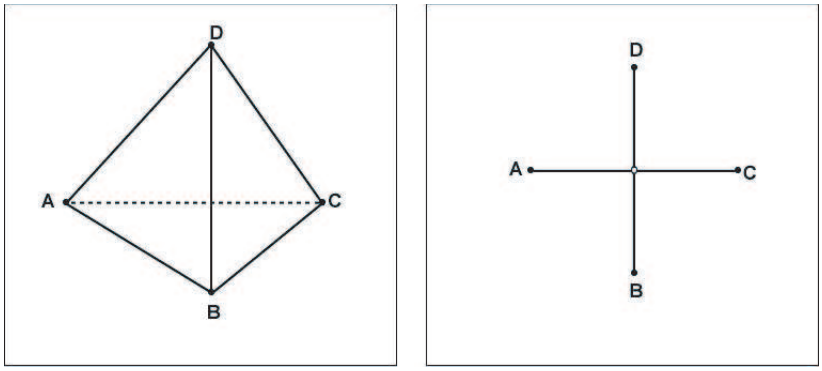

Figure 5.2: Equal city-block distances among four points. Tetrahedron with equal edge lengths (*left panel*) and star graph with equal spokes, which generates the same distances (*right panel*).

points[3]. Hence, the network for $\mathbf{Z}_0$ is a complete graph with equal edge lengths, a structure called a *simplex*. Now, note that the intersection of all metric segments in $\mathbf{Z}_0$ contains exactly one point, the origin $[0\ 0]'$. Introducing the origin as an internal node, four edges are added to the network, all with edge length 1, which brings the total number of edges up to 10. However, since the origin is in all of the six metric segments, the six original edges can all be dropped, bringing the total number of edges down to four. Figure 5.2 shows the simplex and the reduced network. In general, in the equal-distance case the number of edges can always be reduced from $\frac{1}{2}m(m-1)$ to m by the introduction of one internal node. The resulting graph is a special case of a *star graph* (Carroll, 1976) and the resulting metric is called a "center distance" (Le Calvé, 1985).

Let us describe the star graph and the center distance in general terms, as they are a special city-block structure of independent interest; we will encounter this case again later. First, a special four-dimensional configuration yields the same city-block distances as $\mathbf{Z}_0$ in Equation 5.5. It is just the uniform diagonal matrix $\mathbf{Y}_0$ defined as

$$\mathbf{Y_0} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5.6}$$

The uniform diagonal configuration $\mathbf{Y}_0$ in Equation 5.6 can be generalized to arbitrary m and unequal distances, whereas $\mathbf{Z}_0$ in Equation 5.5 cannot. In particular, collecting a set of object-specific, non-negative weights $\mathcal{A} = \{\alpha_1, \cdots, \alpha_m\}$ as diagonal entries in the $m \times m$ diagonal matrix $\mathbf{Y}$, we calculate the city-block distance

[3]Note that a diagonal matrix with all diagonal elements equal to 2 generates the same set of distances.

between any two rows of $\mathbf{Y}$, denoted by A_i and A_j, as

$$d(A_i, A_j) = \sum_t |y_{it} - y_{jt}| = |y_{ii} - 0| + |0 - y_{jj}| = \alpha_i + \alpha_j, \tag{5.7}$$

where the simplification follows from the fact that $y_{ij} = 0$ if $i \neq j$. Thus, diagonality of a city-block configuration leads to an additively decomposable metric. Although the distance function has additive form, note that for $i = j$ we have $d(A_i, A_i) = 0$, and *not* $2\alpha_i$. Therefore, the distance matrix $\mathbf{D}$ is not additive. We can choose between two geometrical representations of Equation 5.7: either as a polytope with m vertices in $m - 1$ dimensions, which follows from the geometry of the rows of $\mathbf{Y}$, or as a star graph with m external nodes or *leaves*, one internal node or *hub* (corresponding to the m-vector of zeros), and m edges or *spokes*. The spokes have the special property that they all coincide in the hub, and are of length α_i. The regular simplex shown in the left panel of Figure 5.2 is a four-point polytope with edges of equal length, while the reduced network in the right panel of Figure 5.2 is a star graph with four leaves, one hub, and four spokes of equal length.

We now return to the general case to demonstrate the use of internal nodes. The two-dimensional city-block solution for the rectangle data of Borg and Leut-

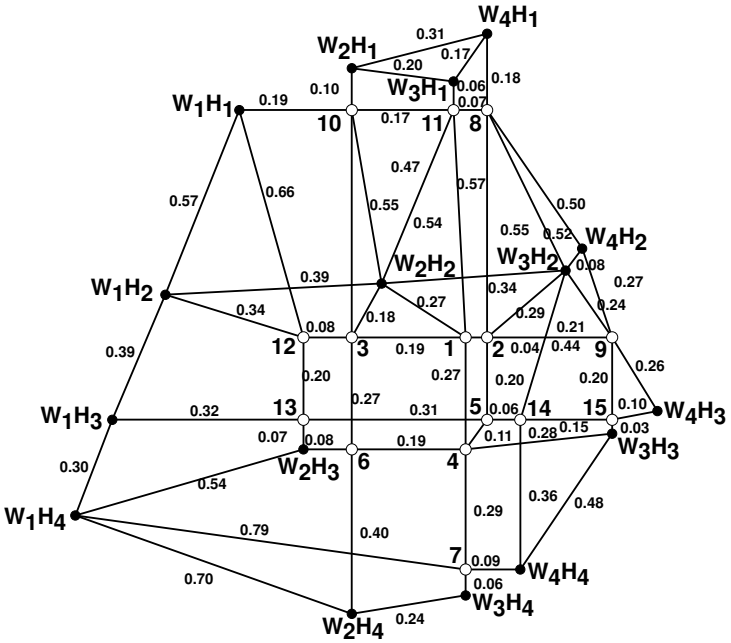

Figure 5.3: Network representation of the two-dimensional city-block solution for the *rectangle* data, including fifteen internal nodes. The labels $W_1 - W_4$ indicate the width levels, and $H_1 - H_4$ the height levels of the stimulus rectangles.

ner (1983), discussed earlier in connection with Figure 5.1, is plotted as a network with internal nodes in Figure 5.3. The internal nodes are indicated with open dots, and labeled according to the order in which they were created, while the external nodes (or leaves) are indicated with solid dots and labeled with their width and height level. The introduction of internal nodes 1 and 2 eliminated all long lines in Figure 5.1B between W_3H_1 and W_4H_1 on the one hand, and W_3H_4 and W_4H_4 on the other hand. Similarly, the introduction of internal node 3 eliminated the long lines between W_2H_1, W_2H_2, W_2H_3 and W_2H_4, all rectangles of width 2. This strategy was continued for all rectangles of height 3 and other subsets until, starting with internal node 7, new nodes were introduced with the additional objective of reducing the degree of the external nodes and the number of crossings. The result in Figure 5.3 more clearly shows the city-block character of the solution than Figure 5.1B, while still accounting for the same distances. It may be verified that if the shortest path between two points includes one or more internal nodes, their direct distance in Figure 5.1B equals the sum of the path lengths in Figure 5.3 (up to rounding error).

After adding 15 new nodes, the total number of lines has a small increase from 56 to 61, of which 20 are among internal nodes only, 28 are between internal and external nodes, and 13 are among external nodes only. The longest lines have been eliminated, and all nodes have lower degree. For example, node W_2H_2 now has degree 6, while it had degree 11 before, and node W_2H_3 now has degree 3, while it had degree 10 before. In addition, the number of crossings at locations without intermediate node has decreased considerably. Thus, internal nodes can indeed simplify several aspects of the network representation, and can make it readily interpretable.

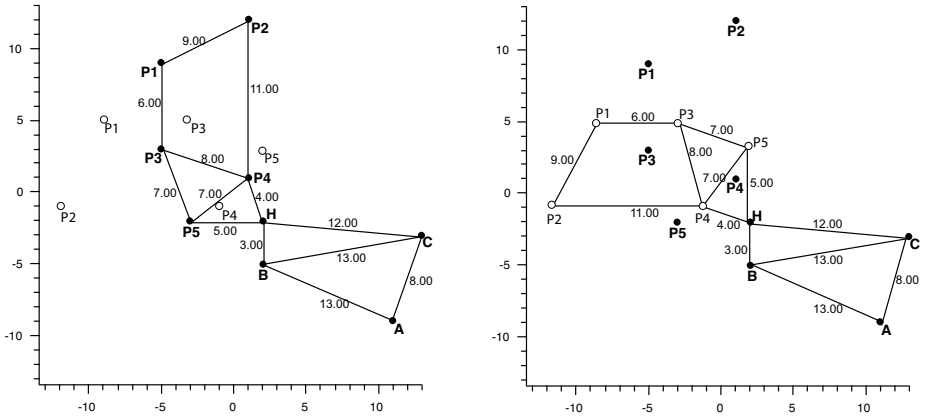

Figure 5.4: Partial isometry: two different configurations with the same city-block distances. *Left panel*: Network representation of A, B, C and the points P1−P5. *Right panel*: Network representation of A, B, C and the points *P1−P5*. The two networks share the internal point H, the hub.

Partial isometries

The network representation helps to explain a puzzling phenomenon that can occur in city-block models. Bortz (1974) has noted that under certain conditions the model coordinates are not unique over and above the usual indeterminacies in distance models, such as invariance of distances under translation (or choice of origin) and reflection of dimensions. Figure 5.4 shows an example (adopted from Bortz, his figure 3), in which two city-block solutions $\mathcal{M}$ and $\mathcal{M}'$ are superimposed: they have the points A, B, and C in common, but $\mathcal{M}$ consists in addition of the points P_1- P_5, while $\mathcal{M}'$ has the points P_1 - P_5. Although these two configurations appear to be quite different, their city-block distances are equal. This effect is high-lighted by including the lines of the network representation of $\mathcal{M}$ (left panel) and $\mathcal{M}'$ (right panel), and an internal node H that lies in the metric segment of all pairs of points for which the first is selected from the set $\{A, B, C\}$, and the second from either P_1- P_5 or $P_1 - P_5$. It is clear that the only difference between the left panel and the right panel of Figure 5.4 is that they are different embeddings of the same network. Only the part above and to the left of H, the internal node called the *hub*, shows a reflection along the 45^o direction, while the part below and to the right of the hub is the same.

A general formulation of this phenomenon is as follows. Partial isometries occur in city-block spaces whenever the set of objects can be partitioned into subsets $\{F_1, F_2, F_3, \cdots\}$ in such a way that the coordinates of objects from different subsets are either monotonically ascending ($z_{At} \leqslant z_{Bt} \leqslant z_{Ct} \leqslant \cdots$) or monotonically descending ($z_{At} \geqslant z_{Bt} \geqslant z_{Ct} \geqslant \cdots$) for any t, with $A \in F_1$, $B \in F_2$, and $C \in F_3$. This condition implies that we can define a hub in the intersection of all metric segments of points selected from any pair of consecutive subsets. In the network representation, all between-subset distances are thus channeled through (one or more) hub(s). In the embedding of the network in city-block coordinates, we can apply reflections within subsets without altering either the within-subset distances (since reflections do not change distance) or the between-subset distances (since distances to the hub remain unaltered). Summarizing, while the coordinate space is not unique under the monotone subset condition, there is only one network, which merely has different embeddings. Both representations allow an interpretation only in terms of the several within-subset constellations and the global order of the subsets.

5.3 Discrete models that are special cases of the city-block model

Some discrete models of similarity are special cases of the city-block model, and therefore we can make network representations by the same token. One may define these discrete models as structures on subsets of objects, but also as city-block models with binary coordinates. We will first discuss a fundamental property of all discrete models in terms of a condition on subsets, called *lattice betweenness*, and show that lattice betweenness is a special case of metric-segmental betweenness when all coordinates are binary. The most general of all discrete models considered is the distinctive features model, a distance model based on the symmetric set difference, well known to be equivalent to the city-block model on binary coordinates. We then discuss the common features (or additive clustering) model, and show how to obtain

a network representation for this model, too. After a discussion of the conditions for obtaining a perfect solution, for both the common and the distinctive features model, we turn to two special cases, the partitioning model with main effects, and finally the additive tree model.

Lattice betweenness of feature sets

Restle (1959) tried to justify a metric analysis of psychological similarity from set-theoretic considerations, by using the concept of betweenness of sets of qualitative elements and the symmetric set difference as a distance measure[4]. We first discuss the nature of betweenness in this context, returning to the set-theoretic distance in the next section. The common definition in logic for betweenness of sets is to say that if $\mathcal{S} = \{S_1, \cdots, S_i, \cdots, S_m\}$ is a family of subsets of some set of arbitrary or qualitative elements, S_j is between S_i and S_k if the following condition holds:

$$(S_i \cap S_k) \subseteq S_j \subseteq (S_i \cup S_k) \tag{5.8}$$

Thus, to be between S_i and S_k, subset S_j has to share at least all elements common to them, while it cannot have elements not present in either of them. The set of all subsets ordered by the inclusion operator $\subseteq$ is a complete lattice (*cf.* Davey & Priestley, 2002). Therefore, we refer to Equation 5.8 as *lattice betweenness*. To clarify the relation of lattice betweenness and metric-segmental betweenness, we have to make explicit the reliance of the subsets in $\mathcal{S}$ on the base set of qualitative elements. Let this base set be $\mathcal{F} = \{F_1, \cdots, F_t, \cdots, F_T\}$, where the T elements are called *features*[5]. We define the feature matrix $\mathbf{E} = \{e_{it}\}$ as an $m \times T$ binary incidence matrix, where $e_{it} = 1$ if S_i has feature F_t, and $e_{it} = 0$ if not. Thus, the rows of $\mathbf{E}$ characterize an object in terms of a subset of features, while the columns of $\mathbf{E}$ characterize a feature in terms of a subset of objects.

We now show that lattice betweenness is a special case of metric-segmental betweenness. For metric-segmental betweenness between A, B, and C to hold, the coordinates of a city-block configuration have to be either monotonically ascending ($z_{At} \leqslant z_{Bt} \leqslant z_{Ct}$) or monotonically descending ($z_{At} \geqslant z_{Bt} \geqslant z_{Ct}$) for all t (if B is between A and C). Transferring this condition to the binary coordinates in $\mathbf{E}$, we must have either $e_{it} \leqslant e_{jt} \leqslant e_{kt}$ or $e_{it} \geqslant e_{jt} \geqslant e_{kt}$ for all t (if S_j between S_i and S_k). To get from here to Equation 5.8, consider all eight possible (0, 1)-patterns of e_{it}, e_{jt}, and e_{kt}. One may easily verify that six of them satisfy monotonicity, while two of them indicate violation of monotonicity. In particular, violation occurs if

$$(1 - e_{it})e_{jt}(1 - e_{kt}) = 1 \quad or \quad e_{it}(1 - e_{jt})e_{kt} = 1 \tag{5.9}$$

for any t. Interpreting Equation 5.9 in terms of features, we see that metric-segmental betweenness implies that the center subset S_j cannot possess any feature F_t that the

[4]The logician Nelson Goodman already studied the order and topology of qualities in his 1951 book *The Structure of Appearance*, a revised version of his 1940 doctoral thesis *A Study of Qualities* (Harvard University). Galanter (1956) introduced Goodmans ideas in psychology and put them to work with some preliminary experimental findings on color vision.

[5]The index t and parameter T were used earlier for the dimensions of the city-block space, but there is no danger of confusion, as it will turn out that features have exactly the same role as dimensions.

two outer subsets S_i and S_k fail to have, and that S_j also cannot lack any feature F_t that the two outer subsets S_i and S_k both possess. Thus, Equation 5.9 is equivalent to $\overline{S_i} \bigcap S_j \bigcap \overline{S_k} = \emptyset$ and $S_i \bigcap \overline{S_j} \bigcap S_k = \emptyset$ holding at the same time (this is the formulation in Definition 2 used by Restle, 1959), which is in turn equivalent to Equation 5.8. Therefore, a single notion of betweenness for a finite set of points applies equally well in continuous space as in feature space.

Distinctive features model

What metric can we use in a representation of objects as subsets of features? Natural candidates for a distance between subsets are functions of the symmetric set difference,

$$d(S_i, S_j) = \mu\left[(S_i \cup S_j) - (S_i \cap S_j)\right] = \mu\left[(S_i - S_j) \cup (S_j - S_i)\right], \tag{5.10}$$

where $\mu[\cdot]$ is some measure function, usually just a count of the features in the subset[6]. The first part of Equation 5.10 expresses the symmetric set difference in terms of the subset of relevant features (i.e., features in the union) that is not common to the two objects (i.e., features in the intersection). The second part of Equation 5.10 expresses the same notion in terms of the total number of features that belong to S_i but not to S_j (distinctive features for S_i with respect to S_j) and those that belong to S_j but not to S_i (distinctive features for S_j with respect to S_i). Because of the latter formulation, Tversky (1977) has called a model based on equation Equation 5.10 a *distinctive features* model. Note that we do not interpret the term "distinctive" as a qualification of the features (as do Navarro and Lee (2004) in their Modified Contrast Model), but as a qualification of what contributes to the similarity or difference in *pairs of objects*.

One of Restle's (1959) results was that lattice betweenness is equivalent to additivity of the distinctive feature distance, i.e., $d(S_i, S_k) = d(S_i, S_j) + d(S_j, S_k)$. Although in the present context this result readily follows from the equivalence of lattice betweenness and metric-segmental betweenness, it is instructive to derive it explicitly here (via the feature coordinates in $\mathbf{E}$). Suppose $\mu[\cdot]$ is a weighted count measure with weight η_t for feature F_t. As a preliminary step, note that introduction of the feature coordinates allows us to write Equation 5.10 as

$$d(S_i, S_j) = \sum_t \eta_t \left[(1 - e_{jt})e_{it} + (1 - e_{it})e_{jt}\right], \tag{5.11}$$

from which it follows that

$$d(S_i, S_j) = \sum_t \eta_t (e_{it} - e_{jt})^2 = \sum_t |z_{it} - z_{jt}|, \tag{5.12}$$

with $z_{it} = \eta_t e_{it}$. Due to the binary nature of e_{it}, we can replace the squares in Equation 5.12 with absolute values. Thus, the distinctive feature distance is a city-block

[6] Restle (1959) mentions that Hays (1958) used the same distance concept, calling it the "implicational difference", and that he used multidimensional scaling to embed feature distances in Euclidean space

distance, where the points are constrained to lie on the corners of a rectangular (hyper-) block, and where the coordinates on any dimension are limited to two values, zero or η_t. Each feature splits the objects into two classes, and η_t measures how far these classes are apart; for this reason, Heiser (1998) called the feature weight η_t a *discriminability* parameter.

Next, consider three points S_i, S_j, and S_k, and assume betweenness in that order; then we may rewrite the distance between S_i and S_j in Equation 5.11 as

$$d(S_i, S_j) = \sum_t \eta_t \left[e_{it}(1 - e_{jt})(1 - e_{kt}) + (1 - e_{it})e_{jt}e_{kt} \right], \tag{5.13}$$

since if S_i has F_t and S_j has not, then S_k cannot have that feature either, so that $e_{it}(1 - e_{jt}) = e_{it}(1 - e_{jt})(1 - e_{kt})$, while if S_i does not have F_t but S_j does, then S_k must have it too, so that $(1 - e_{it})e_{jt} = (1 - e_{it})e_{jt}e_{kt}$. In other words, Equation 5.13 follows from Equation 5.9. With an analogous expression for $d(S_j, S_k)$, we find

$$\begin{aligned} d(S_i, S_j) + d(S_j, S_k) &= \sum_t \eta_t \left[e_{it}(1 - e_{jt})(1 - e_{kt}) + (1 - e_{it})e_{jt}e_{kt} \right] \\ &+ \sum_t \eta_t \left[e_{it}e_{jt}(1 - e_{kt}) + (1 - e_{it})(1 - e_{jt})e_{kt} \right] \\ &= \sum_t \eta_t \left[e_{it}(1 - e_{kt}) + (1 - e_{it})e_{kt} \right] = d(S_i, S_k). \end{aligned}$$

This equality establishes the result. The implication is that we have metric segments in feature space that are paths along the corners of a (hyper-) block, or equivalently (Flament, 1963, p. 17) as paths in the lattice spanned by the feature sets. Hence, the distinctive features model can be represented as a weighted graph or network, using the same graph construction strategy as the one used for the general city-block model; for discrete models, Heiser (1998) called these representations *feature graphs*. We can also construct internal nodes in the same way. Recall that internal nodes correspond to additional points that are located in one or more metric segments generated by the original (external) points. From condition Equation 5.8, it follows that this rule is equivalent to choosing internal nodes as intersections of feature sets.

Corter and Tversky (1986) provided the first method to fit the distinctive features model, by constructing a so-called extended similarity tree. They used a three-stage procedure: in the first stage, their procedure fits the best additive tree to the data, which limits the features to be either nested or disjoint; in the second stage, it selects additional features to be included in the model, and the third stage the feature weights are estimated for the total set of features. Heiser (1998) used a two-stage alternating least squares method, which just cycles between improvement of the feature structure and improvement of the weight estimates, without the backbone of the additive tree. A third method was recently proposed by Navarro and Lee (2004) as a special case of a more general approach, in which they used maximum likelihood estimation assuming that the similarities are normally distributed with common variance, and employing a greedy heuristic to find the feature sets. These methods were all developed independently, and what their relative merits are, is an open question. There are only a few applications without a priori known feature structure. Parault and Schwanenflugel (2000) used extended similarity trees

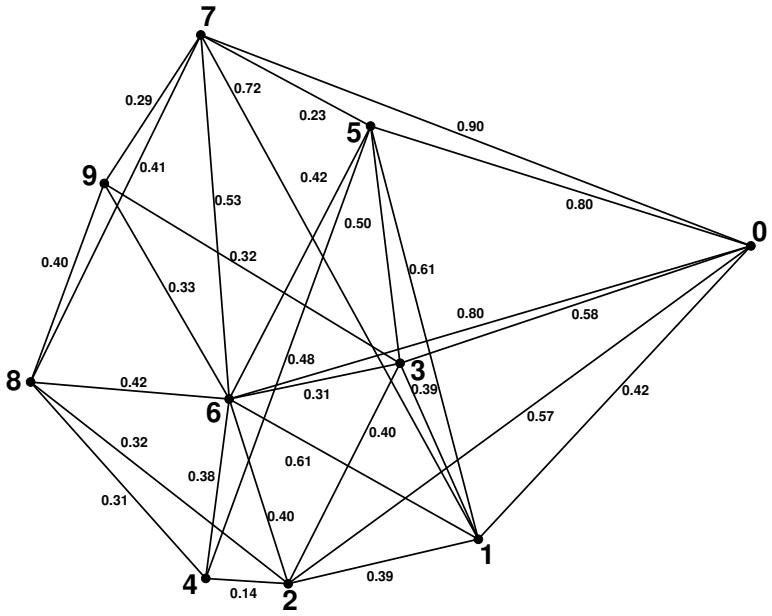

Figure 5.5: Network representation of distinctive features model for the *number* data, without internal nodes. Nodes labeled by stimulus value.

to study the development of childrens categorical knowledge of attention. Heiser and Meulman (1997) used the distinctive features model to cluster profiles of binary multivariate data.

We now demonstrate the construction of a feature network with an example of data collected by Shepard, Kilpatric, and Cunningham (1975), who obtained ratings of similarity between all pairs of integers from zero to nine, considered as abstract concepts. For ease of comparison, we use the same twelve features as Corter and Tversky (1986) found with their EXTREE method. Features in models like these are undefined qualitative elements, but are interpretable by listing the stimuli that have them. The first three form exclusive subsets: the additive and multiplicative identities F_1 = {0, 1}, powers of two $F_2 = \{2,4,8\}$, and a heterogeneous subset of remaining integers $F_3 = \{3,5,6,7,9\}$. Next, we have the nested features primes larger than three $F_4 = \{5,7\}$, multiples of three $F_5 = \{3,6,9\}$, powers of three $F_6 = \{3,9\}$, and the first two powers of two $F_7 = \{2,4\}$. There are five more features that form overlapping subsets: sets of consecutive integers $F_8 = \{0,1,2,3\}$, $F_9 = \{7,8,9\}$, $F_{10} = \{0,1,2,3,4\}$, and $F_{11} = \{4,5\}$, and the multiples of two, or even numbers $F_{12} = \{2,4,6,8\}$. Finally, we included two unique features $F_{13} = \{0\}$ and $F_{14} = \{1\}$, since otherwise zero and one would have identical feature sets, so that they would not be distinguished in the model, obtaining mutual distance of zero, and would become disconnected from the network.

After estimating the weights using nonnegative least squares (Frank & Heiser, in press a; Heiser, 1998), and applying our basic edge deletion method of dropping the direct edge between two points whenever an intermediate point exist in their metric segment, one gets the network displayed in Figure 5.5. This solution accounts for 98.36% of the dispersion (raw sum of squares) of the data, using 29 edges and 14 parameters. The network itself is a discrete structure in fourteen-dimensional space, but it was embedded in the Euclidean plane by multidimensional scaling of the estimated city-block (feature) distances with the program PROXSCAL (Heiser & Busing, 2004), using a simplex start and allowing a ratio transformation. All edge lengths are included in Figure 5.5 since the Euclidean distances in the plot only approximate them. In the embedded network, the three major features F_1, F_2, and F_3 differentiate well, as do the nested features F_4, F_5, F_6, and F_7. With his large number of connections, stimulus 6 clearly exhibits its overlapping position as a member of the even numbers on the bottom-left and the multiples of three in the center. At the (bottom-)right side of the plot, we have the overlapping features F_8 and F_{10}, and on the top the primes F_4 and top-left the large numbers F_9. Therefore, it appears that the embedded network successfully displays the major characteristics of the distinctive features model in an accessible way.

Nevertheless, the introduction of internal nodes can make the structure even more transparent. The natural choice of internal nodes in the distinctive features model is to identify a cluster in the high-dimensional feature space with a new point that is located in the intersection of the features shared by the objects in that cluster. For example, the internal node corresponding to the cluster $C_1 = \{0, 1\}$ has the features F_1, F_8, and F_{10}, since zero and one share exactly these features. This way of defining internal nodes ensures that the distance between the two members S_i and S_j of cluster C_ℓ splits: $d(S_i, S_j) = d(S_i, C_\ell) + d(C_\ell, S_j)$, because for an additive measure μ we have $d(S_i, S_j) = \mu(S_i - S_j) + \mu(S_j - S_i) = \mu(S_i - C_\ell) + \mu(S_j - C_\ell)$ when $C_\ell = S_i \cap S_j$, and $d(S_i, C_\ell) = \mu(S_i - C_\ell)$ since $\mu(C_\ell - S_i) = 0$. Similarly, we have $d(S_i, C_\ell) = d(S_i, C_k) + d(C_k, C_\ell)$ for members of two nested clusters with $S_i \subset C_k \subset C_\ell$. These additivities lead to better interpretable paths in the network and a low degree for the external nodes, especially if the features are nested or disjoint. Nested features lead to nested clusters, represented as a chain of internal nodes. Let us see how this representation works for the digit data.

The introduction of internal nodes for all clusters corresponding to the 12 features, as well as five extra internal nodes associated to the objects 2, 5, 7, 8, and 9, for which we fitted additional unique features, lead to a network of 27 nodes in 19 binary dimensions in city-block space. Including unique features for the other objects did not improve the fit. We obtained a PROXSCAL embedding with the same options as before; Figure 5.6 displays the result. The seventeen internal nodes are plotted as open dots, while the ten object nodes (external nodes or leaves) are plotted as solid dots. Every object node with a unique feature is connected to the rest of the network via an (unlabeled) internal node with a spike of length equal to the unique feature weight. Note that all paths from 0 and 1 go through {0, 1} and then through {0, 1, 2, 3}, all paths from 2 go through {2, 4} and {0, 1, 2, 3}, all paths from 3 go through {3, 9} and {0, 1, 2, 3}, all paths from 4 go through {2, 4} and {4, 5}, and so on. In other words, in this example all objects have only two direct neighbors, which are always

internal nodes (clusters), except for 0 and 1, which have only one direct neighbor because they differ only in their unique features, and 7 and 9, which have a mutual link in addition to their two cluster connections.

Figure 5.6 also shows how nesting of features leads to nested clusters in a chain of internal nodes. The most important chains are: the small numbers {(0, 1), (0, 1, 2, 3), (0, 1, 2, 3, 4)}, the even numbers {(2, 4), (2, 4, 8), (2, 4, 6, 8)}, and the prime numbers plus powers and multiples of three {(3, 9), (3, 6, 9), (3, 5, 6, 7, 9)}. The three chains are connected in a triangle in the center of the display, which forms a complete sub-network of internal nodes together with the medium-sized numbers (4, 5) and the large numbers (7, 8, 9). All object nodes are connected via one or two paths to this basic complete sub-network. The path can be linked either directly, like from 6 to (2, 4, 6, 8) and from 8 to (7, 8, 9), or go through the closest node in one of the chains, like from 2 to (2, 4) in the chain of even numbers, or from 2 to (0, 1, 2, 3) in the chain of small numbers. The global structure of the embedding appears to consist of a basic plane with the powers of two (2, 4, 8) on one side and the powers of three (3 and 9) on the other side, with small numbers at the right and large numbers at the left. It appears that objects 5, 6, and 7 do not fit well into this plane, either because they have a large unicity (5, 7) or because they share only partly features from both sides (6 is a multiple of both two and three, but not a power of them). The identities 0 and 1 have an eccentric position with large unicity, but as a cluster, they are close to the small numbers.

Additive clustering or the common features model

Shepard and Arabie (1979) proposed an additive clustering model, which builds up the similarity $s(S_i, S_j)$ between S_i and S_j from unrestricted binary features, according to the rule

$$s_{ij} = s(S_i, S_j) = \mu\left[S_i \cap S_j\right] = \sum_t \theta_t e_{it} e_{jt}, \tag{5.14}$$

where the θ_t are again nonnegative weight parameters. This model thus uses a weighted count of the features in the intersection of the feature sets of each object in a pair. Since the model only takes features into account that the pair of objects have in common, Tversky (1977) has called it a *common features model*. Mirkin (1987) developed the model independently under the name *qualitative factor analysis*, around the same time as Shepard and Arabie, and adjusted it to the analysis of contingency tables (the two-mode case) in Mirkin (1996). Arabie and Carroll (1980) and Carroll and Arabie (1983) developed algorithms for finding the feature sets and the feature parameters of the additive clustering model and its three-way generalization. Soli, Arabie, and Carroll (1986) reported an application of the three-way additive clustering model. More recent algorithmic strategies are given in Mirkin (1990, 1998), Chaturvedi and Carroll (1994), and Ten Berge and Kiers (2005), among others.

In the additive clustering model, each feature defines a cluster of objects. The unrestricted nature of the features implies that the clusters need not be exclusive and may overlap. As noted by Carroll and Corter (1995), graphical representations of non-nested overlapping clustering are usually complex and difficult to in-

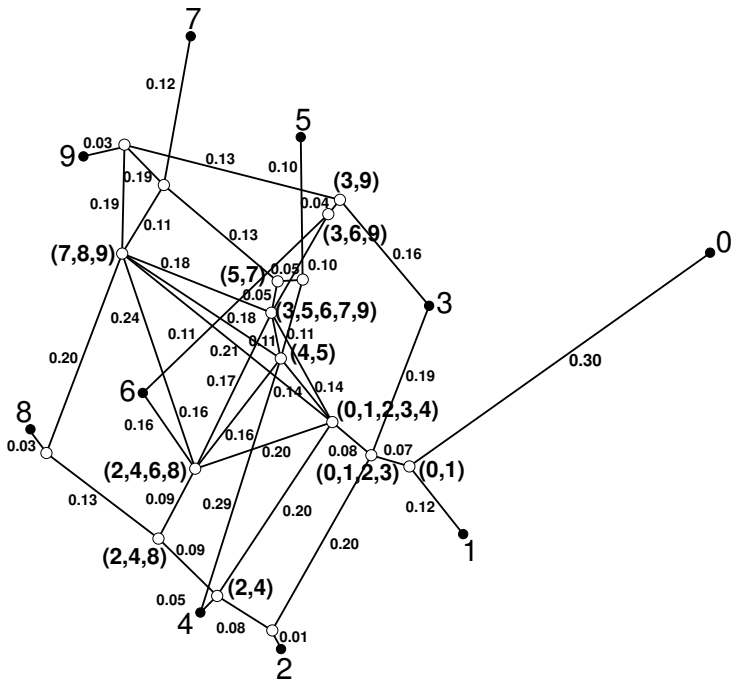

Figure 5.6: Network representation of distinctive features model for the number data, with internal nodes. Solid dots are stimuli labeled by stimulus value, open dots are internal nodes labeled by subset.

terpret. Shepard and Arabie (1979) used a two-dimensional projection of a three-dimensional city-block embedding of the original data, and then added contour lines around sets of points that correspond to clusters in the additive clustering solution. Carroll and Pruzansky (1980) proposed representing non-nested clustering by multiple trees, and (Corter & Tversky, 1986) by extended trees. What we want to show now is that the clusters derived from the additive clustering model have a natural representation as a feature network. To demonstrate this possibility, we express the common features model as a special case of the distinctive features model. This relationship was first established by Sattath and Tversky (1987).

Suppose that we have a feature set $\mathcal{F}$, coded in a feature matrix $\mathbf{E}$, and weight parameters $(\hat{\theta}_1, \cdots, \hat{\theta}_t, \cdots, \hat{\theta}_T)$, that approximate some similarity ς_{ij} according to the common features model (Equation 5.14), where we denote the approximation by $\hat{s}_{ij} = \sum_t \hat{\theta}_t e_{ij} e_{jt}$. We want to demonstrate that it is possible to form a specific linear transformation $\hat{d}_{ij} = 2K - 2\hat{s}_{ij}$ that follows exactly a distinctive features model, where K is some constant that we will specify later. We can use the same feature set $\mathcal{F}$, but we have to append to it a set of m unique features. A *unique feature* is a feature with only one object associated to it, with non-negative weight. To distinguish the

features in **E** from the unique features, we call the former *shared features*, since there are always two or more objects sharing a non-unique feature. The feature matrix of a set of unique features is diagonal, so that by themselves they form an additively decomposable metric associated with a star graph, as we saw in the discussion of Figure 5.2. We will use the notation e_{it^*} for the unique features, with $t^* = 1, \cdots, m$, and with the understanding that $e_{it^*} = 1$ if $i = t^*$ and $e_{it^*} = 0$ otherwise. Without danger of confusion, we use $\alpha_i = \sum_{t^*} \alpha_{t^*} e_{it^*}$ for the unique weight of object i.

To let the switch from common features model to distinctive features model work, it suffices to take identical weights for the shared features, while the weights for the unique features are a simple function of the shared feature weights, specified as follows:

$$\begin{aligned} \hat{\eta}_t &= \hat{\theta}_t \\ \hat{\alpha}_i &= K - \sum_t \hat{\theta}_t e_{it}. \end{aligned} \tag{5.15}$$

The constant K can be chosen freely as long as it does not make the weights α_i negative, i.e., as long as it satisfies $K \geqslant \max_i \sum_t \hat{\eta}_t e_{it}$. From Equations 5.15 we have $K = \hat{\alpha}_i + \sum_t \hat{\eta} e_{it}$, so that we may write

$$\begin{aligned} \hat{d}_{ij} &= 2K - 2\hat{s}_{ij} = K + K - 2\sum_t \hat{\theta}_t e_{it} e_{jt} \\ &= \hat{\alpha}_i + \sum_t \hat{\eta}_t e_{it} + \hat{\alpha}_j + \sum_t \hat{\eta}_t e_{jt} - 2\sum_t \hat{\eta}_t e_{it} e_{jt} \\ &= \left[\sum_t \hat{\eta}_t e_{it} + \sum_t \hat{\eta}_t e_{jt} - 2\sum_t \hat{\eta}_t e_{it} e_{jt}\right] + \left[\hat{\alpha}_i + \hat{\alpha}_j\right] \\ &= \sum_t \hat{\eta}_t |e_{it} - e_{jt}| + \sum_{t^*} \hat{\alpha}_{t^*} |e_{it^*} - e_{jt^*}|. \end{aligned} \tag{5.16}$$

Note that whenever we have the approximation $\hat{d}_{ij}$, we also recover $\hat{s}_{ij} = K - \frac{1}{2}\hat{d}_{ij}$. Also, note that $(\eta_1, \cdots, \eta_t, \cdots, \eta_T)$ and $(\alpha_1, \cdots, \alpha_{t^*}, \cdots, \alpha_m)$ should not be seen as a set of $T + m$ independent parameters, because both are functions of the T parameters $(\theta_1, \cdots, \theta_t, \cdots, \theta_T)$. Clearly, $\hat{d}_{ij}$ in Equation 5.16 has the desired form of a distinctive feature distance, since the sum of two feature distances is again a feature distance with dimensionality equal to the sum of the two original dimensionalities. Therefore, we can check all triads of points for lattice betweenness to see which edges of the network we can delete, as usual.

It is often useful in this approach to the graphical representation of the common features model to define m internal nodes, one for each object, with shared features that are the same, but without unique features. The effect will be that the feature graph displays the structure of the shared features in its internal nodes, each of which corresponds to (and can be labeled with) exactly one object. In addition, each internal node has one unique edge (a spoke toward one external node) attached to it, the length of which indicates the relative distance of an object towards all others. This spoke (and its length) is analogous to a unique factor (and its variance) in factor analysis. Hence, Mirkin (1987) name qualitative factor analysis for the common features model is well chosen.

In factor analysis, the diagonal of the correlation matrix to which we fit the model is constant, and the unique factors are necessary to account for the variance left unexplained by the common factors. In the common features model, the shared features produce diagonal terms equal to $\hat{s}_{ii} = \sum_t \hat{\theta}_t e_{it}$, the sum of the weights that an object possesses, and these will generally not be constant either. Hence, if one would like to account for the diagonal elements of the similarity matrix, one would need to append a set of m unique features to the common features model as well, that is, write the model as

$$\underline{s}(S_i, S_j) = \sum_t \theta_t e_{it} e_{jt} + \sum_{t^*} \beta_{t^*} e_{it^*} e_{jt^*},$$

with e_{it^*} denoting the unique feature of object i and $\beta_{t^*} = \beta_i \geqslant 0$ its non-negative weight. The diagonal elements are equal to the sum of all weights relevant for object i:

$$\underline{s}(S_i, S_i) = \sum_t \theta_t e_{it} + \sum_{t^*} \beta_{t^*} e_{it^*} = \sum_t \theta_t e_{it} + \beta_i,$$

while the off-diagonal elements remain the same as before, since $\sum_{t^*} \beta_{t^*} e_{it^*} e_{jt^*} = 0$ if $i \neq j$. Therefore, if we want diagonal elements equal to some constant value, that is, $\underline{s}(S_i, S_i) = K$, the unique weights for the extended common features model should be chosen as

$$\hat{\beta}_i = K - \sum_t \hat{\theta}_t e_{it},$$

that is, identical to the unique weights under the extended distinctive features model in Equation 5.15, because $\hat{\eta}_t = \hat{\theta}_t$. Since in practice one usually does not model the diagonal elements of the similarity matrix, the issue never seems to arise. However, to make the model complete, the common features model needs the same unique features as the distinctive features model.

In the distinctive features model, the effect of the unique features with weights $\hat{\alpha}_i$ defined in Equation 5.15 also is to make the sum of the weights for each object constant. This property can be expressed geometrically by calculating the feature distance of object S_i with respect to the origin O (internal node with all-zero profile), and inserting Equation 5.15:

$$d(S_i, O) = \sum_t \hat{\eta}_t e_{it} + \hat{\alpha}_i = \sum_t \hat{\eta}_t e_{it} + K - \sum_t \hat{\eta}_t e_{it} = K.$$

Hence, the common features model for similarity matrices with equal self-similarities is a special case of the distinctive features model: if the unique feature weights satisfy Equation 5.15, then reversing the argument in Equation 5.16 shows that the distances satisfy a common features model. In practical terms, for any fitted common features model we can find an equally well fitting distinctive features model, with object nodes at constant distance from the origin, with the same shared features and feature weights, and with the same number of independent parameters.

As an example of the network representation of the common features model, consider the body parts data collected by Miller (1969), which was reanalyzed by

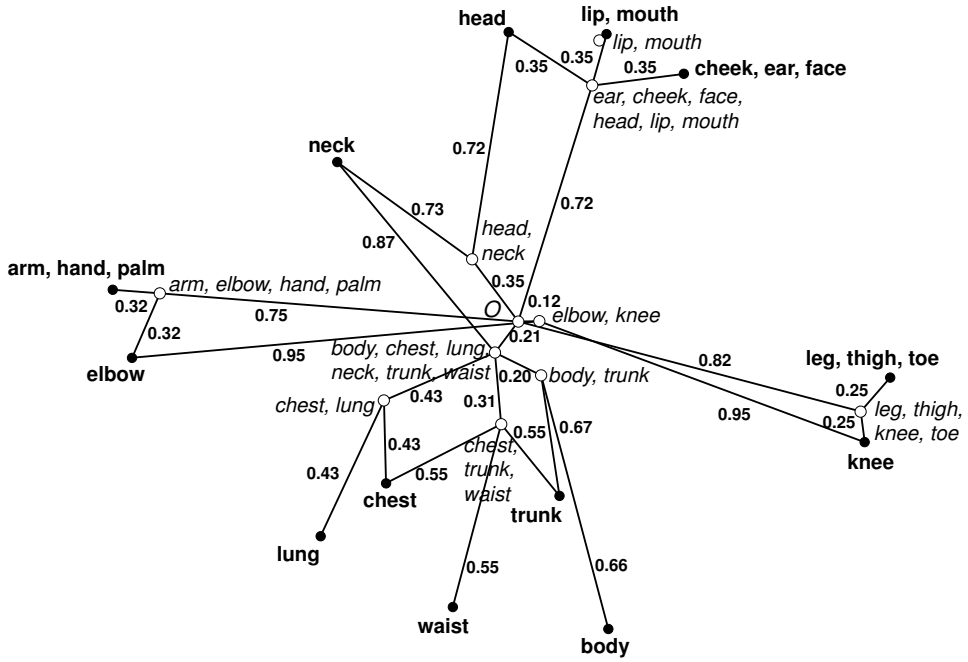

Figure 5.7: Network representation of common features model for *body-parts* data, with internal nodes.

Carroll and Chang (1973), and Shepard and Arabie (1979). Miller's data on the perceived similarity of 20 body parts are counts of the number of times in which 50 subjects, in a free sorting task, put a pair of stimuli into the same group. As noted by Shepard and Arabie, the body parts had been chosen based on a rather clear hierarchy of anatomical inclusion, but with some ambiguities. We have used the same 10 common features and weights found by their ADCLUS procedure, which accounted for 95.6% of the variance in the similarities. Using Equation 5.15 to calculate weights for the shared and unique features under the distinctive features model and applying the general graph construction rule we obtain the network representation in Figure 5.7. In this representation, eleven internal nodes have been included to reduce the degree of some of the nodes. One of them is labeled by O, and can be interpreted as the root or the origin of the network, since it is defined by a profile of zeros on all features. The other internal nodes are labeled by the subsets found by Shepard and Arabie. They are defined by the intersection of the features of the objects in the subset that they represent. The network clearly shows that there are four major clusters: a trunk cluster (consisting of body, chest, lung, neck, trunk, and waist), a leg cluster (knee, leg, thigh, toe), arm cluster (arm, elbow, hand, palm), and a head cluster (ear, cheek, face, head, lip, mouth), which is consonant with previous

analyses. In this solution, we do not find evidence that trunk, leg, arm, and head are especially close to their closest cluster point, to warrant the higher-order status that they had in the Carroll and Chang (1973) solution.

A strong point of the current representation is that violations of hierarchical structure are recognizable as cycles in the network. One major cycle is between the origin, the trunk cluster and the head cluster (via head, neck), and another one is between the origin, the arm cluster, and the leg cluster (via elbow, knee). These cycles arise from the presence of feature 5, which connects two physically neighboring parts, head and neck, and feature 10, which connects two functionally analogous parts, elbow and knee. Thus, in addition to its simple portrayal of the additive clustering solution, in which the clusters themselves can be included in a natural way, the network representation of a common features model allows an immediate diagnosis of departures from purely hierarchical structure. Note the following regularities in Figure 5.7. The sum of the line lengths from each leave node (solid dot) to the origin is constant and equal to 1.07 (up to rounding error). One can read off the dissimilarity between two leaves as the length of their shortest connecting path. At the same time, one can read off their similarity as the sum of the line lengths of a path from the origin to the smallest cluster that they share.

Exact fit of feature models

Is there a feature set that always yields an exact fit to an arbitrary (dis)similarity matrix under these feature models? There is no exact answer in the literature to this question, but the previous section shows how to obtain one. It is clear that under the common features model a basis consisting of all size-two clusters corresponding to all pairs of objects would be sufficient to fit any similarity matrix exactly. Let us denote the features of this basis by $e_{i(k,l)}$, where k, l varies over all ordered pairs, and we have the property $e_{i(k,l)} = 1$ if $i = k$ or $i = l$, and $e_{i(k,l)} = 0$ otherwise. Since $e_{i(k,l)}e_{j(k,l)} = 0$ for all k, l except if $i = k$ and $j = l$, there is exactly one feature for each similarity, so that we can choose $\hat{\theta}_{(k,l)} = \varsigma_{kl}$, obtaining an exact fit $\varsigma_{ij} = \hat{s}_{ij}$.

Under the distinctive features model, we can use the same basis of all size-two clusters, but we need again to include unique features to reproduce any dissimilarity matrix up to a known additive constant. It is not hard to show that in this feature structure no object is between any other object, so that the feature network is a complete graph[7]. The specification of the parameters is

$$\begin{aligned} \hat{\eta}_{(k,l)} &= L - \frac{1}{2}\delta_{kl}, \\ \hat{\alpha}_i &= \frac{1}{2}\sum_{j \neq i}\delta_{ij} - (m-2)L, \end{aligned} \tag{5.17}$$

[7]For three objects A, B and C, the relevant features are AB, AC, and BC. Thus, it suffices to consider $A = \{AB, AC\}, B = \{AB, BC\}$, and $C = \{AC, BC\}$. Whatever object is chosen as the middle one, it violates the requirement defined in Equation 5.9 that it should not lack any feature that the two outer objects possess. For instance, A and C share AC, but B lacks it.

where L is some positive constant. With these weights, the feature distance becomes

$$\begin{aligned} d(S_i, S_j) &= \sum_{k,l} \hat{\eta}_{(k,l)} |e_{i(k,l)} - e_{j(k,l)}| + \hat{\alpha}_i + \hat{\alpha}_j \\ &= 2(m-1)L - \frac{1}{2}\sum_{j \neq i} \delta_{ij} - \frac{1}{2}\sum_{i \neq j} \delta_{ij} + \delta_{ij} - 2L \\ &\quad + \frac{1}{2}\sum_{j \neq i} \delta_{ij} + \frac{1}{2}\sum_{i \neq j} \delta_{ij} - 2(m-2)L \\ &= \delta_{ij}. \end{aligned} \tag{5.18}$$

Thus, for any choice of L in Equation 5.17, we have perfect reconstruction of the dissimilarities. It turns out that adding a constant to the weights of the shared features can be compensated by subtracting (another) constant from the unique features. This indeterminacy is caused by the fact that all pairs of objects differ on the same number of shared features $(m-1)$, and on the same number of unique features (two). We can identify a solution by selecting L so that the smallest unicity becomes zero, for example. However, there is another consideration. When choosing L too small we obtain one or more negative weights $\hat{\eta}_{(k,l)}$ for the shared features, and when choosing L too large we obtain one or more negative weights $\hat{\alpha}_i$ for the unique features, both of which are violations of the model assumptions. Requiring nonnegativity of the two sets of weights in Equation 5.17 gives the following bounds for L:

$$\max_{(i,j)} \delta_{ij} \leqslant L \leqslant \min_i \frac{1}{m-2} \sum_{j \neq i} \delta_{ij}. \tag{5.19}$$

Therefore, we can identify a solution whenever the dissimilarities allow finding an L in the interval Equation 5.19. If no such L exists, we can add the smallest positive constant to the dissimilarities ensuring that Equation 5.19 becomes satisfied. Finding such an additive constant is possible, because the lower bound involves only one dissimilarity, while the upper bound involves the sum of $m-1$ dissimilarities divided by $m-2$, so that the upper bound grows faster than the lower bound. In conclusion, a feature network based on size-two clusters and singletons can always reproduce an arbitrary dissimilarity matrix.

Even though perfect reproduction involves as many as $\frac{1}{2}m(m+1)$ features, while there are merely $\frac{1}{2}m(m-1)$ independent data values, it should be noted that each α-weight can be written as a linear function of the data values, so that we actually do rely on exactly $\frac{1}{2}m(m-1)$ independent quantities. A calculation similar to Equation 5.18 shows that the feature distance of any object to the origin is constant; in particular, we have $d(S_i, O) = L$. Since the square root of the feature distance is Euclidean (see Equation 5.12), it follows that the vertices of the complete graph that perfectly reproduces an arbitrary dissimilarity matrix are located on a hypersphere of dimension $\frac{1}{2}m(m+1)$ with radius $\sqrt{L}$.

Partitioning in clusters with unicities: the double star tree

Consider the situation in which the model consists of a set of unique features and a set of shared features, where the latter has the special property of forming a partition of the set of objects. Thus, each shared feature is disjoint from (or non-overlapping with) any other shared feature, and no object lacks a shared feature (in addition to its unique feature). Without the presence of unique features, this case would be a standard clustering task for which several methods have been developed (*cf.* Hubert, Arabie, & Meulman, 2001). As we saw earlier, unique features can be represented as a star graph. A partitioning in T subsets can be represented as a star graph, too, in which each subset is a vertex and the center of the star is again the origin (an internal node with zero on all features). Fitting a model that is the sum of two star graphs is a simple special case of Carrolls (1976) multiple tree structure approach, but surprisingly no one has considered it in any detail[8].

The graphical representation obtained for the sum of the distances in the star graph of the unique features and the distances in the star graph of the partitioning has a particularly simple form. We need one internal node for the origin, and T other internal nodes (where T is the number of clusters), each having a single non-zero value for one of the features defining the partitioning. With our usual graph construction procedure of eliminating direct lines when two nodes are reachable through another node in their metric segment, we obtain a graph in which each internal cluster node connects only with the origin and with the leaves that constitute the cluster. Thus, the origin node has degree T, the cluster nodes have degree $n_t + 1$, where n_t is the number of objects in cluster t, and the object nodes are leaves with degree one. Any distance between two objects in different clusters equals the sum of four line lengths along the unique path connecting them. Starting with S_i, we have the line from the leave of S_i to the node of the cluster where S_i belongs to, the line from that cluster node to the origin, the line from the origin to the cluster node of S_j, and finally the line from the cluster node of S_j to the leave of S_j. The distance between two objects in the same cluster is just the sum of two line lengths. The graph of the double star tree is simple because it contains no cycles and has only $T + m$ lines. There is also a one-to-one relation between line lengths and feature weights.

For the Shepard et al. (1975) number data, analyzed earlier with the general distinctive features model and displayed in Figures 5.5 and 5.6, Hubert et al. (2001) repeatedly found the optimal partition {(0, 1), (2, 4, 8), (3, 6, 9), (5, 7)}, with different clustering criteria. Therefore, we adopted this partitioning and estimated weight parameters for the shared and unique features with nonnegative least squares. Figure 5.8 displays the resulting network. We see that the graph has all the properties described in the previous paragraph. It has no cycles, and since $m = 10$ and $T = 4$ in this case, it contains 4 + 10 = 14 lines. The origin only connects with the four cluster nodes, and each object only with the cluster node of its own cluster. The distance between 1 and 5, which belong to different clusters, is the sum of the four line

[8]The closest example of a partitioning model with unicities that we could find in the literature is one of the hierarchical tree structure models proposed by Carroll and Chang (1973), which they call the "branches only" model. The partitioning occurs incidentally in their example of the body-parts data, because the fitted tree is not fully resolved

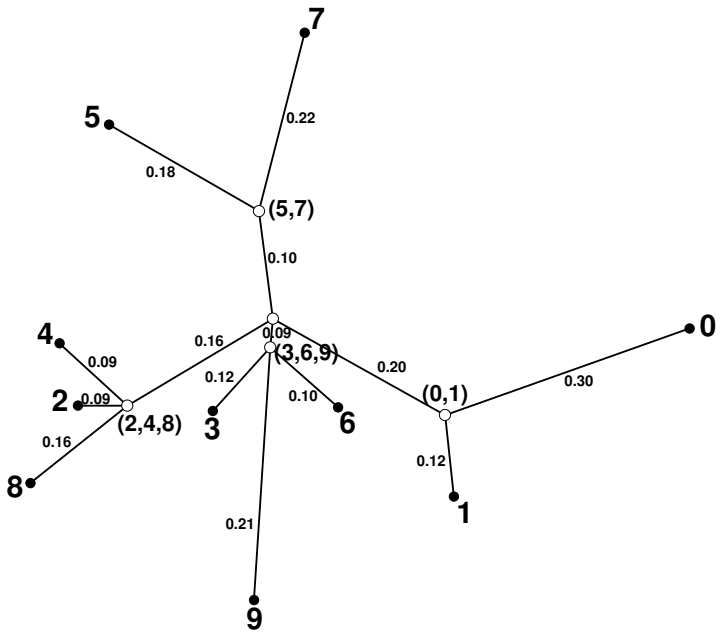

Figure 5.8: Network representation of double star tree for the *number* data.

lengths along the path $(1) - (0,1) - (O) - (5,7) - (5)$, amounting to 0.12 + 0.20 + 0.10 + 0.18 = 0.60 (compared to 0.61 in Figure 5.5). The distance between 2 and 8, which belong to the same cluster, is the sum of the two line lengths along the path $(2) - (2,4,8) - (8)$, amounting to 0.09 + 0.16 = 0.25 (compared to 0.32 in Figure 5.5). The double star tree accounts for 94.84% of the dispersion (compared to 98.36% for the more general model), so it still has a good fit. It is easy to check that none of the within-cluster distances is larger than any between-cluster distance, which is a sign of the quality of the partitioning; for example, compare the largest distance of 0.42 within cluster (0, 1) with a smallest distance of 0.44 between 4 and 6 in the powers of two and the multiples of three clusters. It is a strong point of the double star tree that it models within-cluster distances in addition to between-cluster distances. In contrast, other partitioning methods usually assume that the within-cluster distances are random or zero.

Additive tree model

Consider building up a model with one feature defining a partitioning in two clusters and a full set of unicities, and introduce an extension with features that are limited to be proper subsets of previous clusters (excluding singletons, since there is no need to duplicate the unicities). This construction leads to at most $m - 3$ shared

features that are either nested or disjoint. In terms of feature sets, the implication is that any three objects can be labeled so that $S_i \cap S_j = S_i \cap S_k \subset S_j \cap S_k$. Now the feature distance satisfies a special property that is characteristic for an *additive* or *weighted tree*, called the additive inequality or the four-point condition (Buneman, 1971; 1974). A tree is a connected graph without cycles, and the qualifier additive underlines the property that the distance between any two nodes in a weighted tree is the sum of the weights (line lengths) along the shortest path connecting the nodes. Tversky (1977) was the first to give an interpretation of an additive tree in terms of the distinctive features model, and advocated its use as a practical simplification of his more general Contrast Model. Colonius and Schulze (1981) gave a measurement-theoretical characterization of the tree structure in terms of topological relations between pairs of objects and described corresponding sorting tasks for data collection.

Cunningham (1974, 1978), Carroll (1976) and Sattath and Tversky (1977) motivated their work on the additive tree by pointing out limitations of the more common hierarchical tree and multidimensional scaling representations as models for similarity data. Given two disjoint clusters in a hierarchical tree, for example, all within-cluster distances are smaller than all between-cluster distances, which are all equal. Such severe constraints do not necessarily hold in an additive tree. Pruzansky, Tversky, and Carroll (1982) offered guidelines for deciding between spatial and tree representations on the basis of data properties such as skewness of the (dis)similarity distribution (under an additive tree model the distance distribution is skewed to the left, and under a spatial model distances are skewed to the right). Carroll, Clark, and DeSarbo (1984) proposed extensions of additive tree model to three-way data. Despite its elegance and flexibility, applications of additive trees in psychology are sparse, except perhaps in categorization research. An example is the study of contrast categories in predicting typicality ratings by Verbeemen, Vanoverberghe, Storms, and Ruts (2001).

Several algorithms are available for fitting an additive tree (see Barthélemy & Guénoche, 1991). The major ones are ADDTREE (Sattath & Tversky, 1977), ADDTREE/P (Corter, 1982), an improved implementation of the ADDTREE algorithm because it allows for using metric information, the closely related and widely used neighbor-joining (NJ) method (Saitou & Nei, 1987), and a least squares method due to De Soete (1983). GTREE (Corter, 1998) uses only metric information to select the nearest neighbor for each object and therefore represents an entirely distinct algorithm from ADDTREE and ADDTREE/P. Viewed as a distinctive features model, the tree is characterized by at most $m - 3$ shared features that are either nested or disjoint, and m unique features. Given the tree structure, we can find anyone of the features by cutting any branch of the tree, causing the objects to fall apart in two exclusive subsets. Repeated cutting of all $2m - 3$ branches gives the complete set of features. Given the feature structure, the tree can be found by the present graph construction method, where each of the $m - 3$ shared features is included as an additional internal node (defined as the intersection of the profiles of the objects sharing the feature). The origin should be included as well; this internal node corresponds to the complement of the subset defined by the first feature. There is a one-to-one relation between line lengths and feature weights. An interesting special case arises if we constrain each internal node to be equal to one of the objects (the "branches only"

model of Carroll & Chang, 1973), which amounts to setting the weight of some of the unique features equal to zero. This constrained model has only $m - 1$ parameters.

Corter and Tversky (1986) found an additive tree for the Shepard et al. (1975) number data with ADDTREE. Using the procedure just outlined, we recovered seven shared features. The weight parameters for the shared and unique features have been re-estimated with nonnegative least squares. Our usual graph construction method yielded the network displayed in Figure 5.9, which is comparable with our earlier results for the more general distinctive features model in Figure 5.6 and the more restricted double star tree model in Figure 5.8. The %DAF of this solution of 95.37 is between those of the other two. There are clear common elements between the three solutions, in particular the fact that they share the clusters (0, 1), (5, 7), (2, 4, 8), and (3, 6, 9). The additive tree refines (2, 4, 8) into (2, 4) versus (8), and (3, 6, 9) into (3, 9) versus (6), while it introduces the super-ordinate class (3, 5, 6, 7, 9) by joining (3, 6, 9) and (5, 7). Remarkable differences between the three solutions are the following. In the general distinctive features model, 9 is close to 7, but not in the two other models; this is due to the cluster of large numbers (7, 8, 9), which joins elements from three of the four major clusters apparent in the other two models. Similarly, the cluster of small numbers (0, 1, 2, 3 ,4) in the general feature model forms a major violation of the hierarchical structure, since it also combines elements from three of the four main clusters in the other two models. Both the tree and the general model join (5, 7) with (3, 6, 9) to form (3, 5, 6, 7, 9), but this cluster does not occur in the partitioning model. In the tree, (2, 4) joins with 8 into (2, 4, 8), but does not continue with (2, 4, 6, 8) like in the general model, since 6 is located in another branch of the tree. All differences are understandable from the structural properties of the three models.

5.4 Discussion

Additivity across dimensions and uniqueness of coordinate system have always been the two most appealing properties of the city-block distance, ever since Landahl (1945) started thinking of models for similarity and difference, and Attneave (1950) started experimenting with them (Arabie, 1991). Undoubtedly, the simplest rule for the combination of psychological differences on different dimensions is to add them up with equal weights (Cross, 1965). This paper has shown that the city-block distance is not only additive across its component dimensions, but also across sequences of intermediate points along certain trajectories in space. As an unexpected consequence, the extra additivity of distance allows dropping the whole coordinate system. If we can embed dissimilarities in a city-block coordinate system, we can equally well embed them in a network.

Our construction of the network representation rested upon the notion of the metric segment between any pair of points in space. A metric segment is the area of all intermediary points for which additivity of distance applies. City-block space has metric segments that are rectangles in two dimensions, cuboids in three dimensions, or hyper-cuboids in more than three dimensions. Since these areas are large enough to accommodate a considerable number of intermediate points in any finite set of

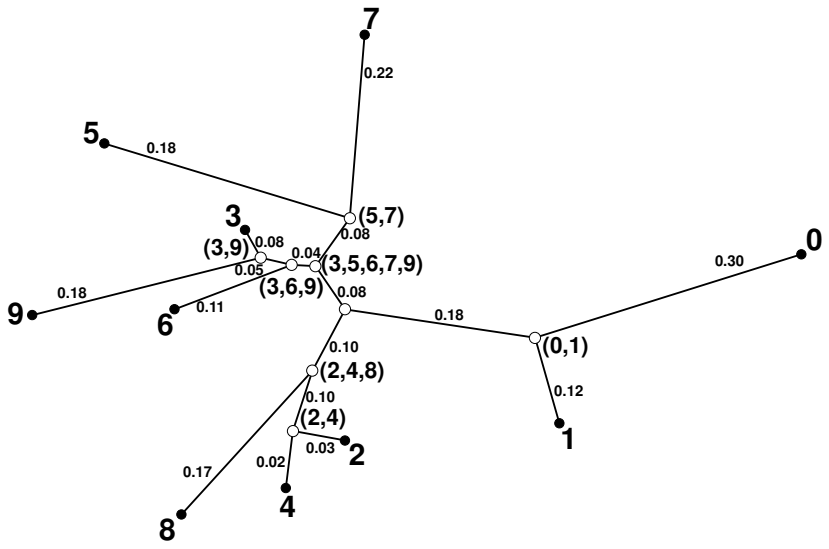

Figure 5.9: Network representation of additive tree for the *number* data.

objects, the possibility of network construction is realistic for city-block models. By contrast, in Euclidean space metric segments are always line segments, and chances of finding intermediate points on line segments are negligible with fallible data on a finite set of objects. We also introduced the general concept of an internal node, which is a supplementary point in the intersection of several metric segments. Internal nodes can be helpful in reducing the complexity of the network, and in making the representation more transparent and better susceptible for interpretation.

A network is coordinate-free, that is, it is entirely determined by the presence or absence of edges between the nodes, and the lengths of these edges; in other words, it exists independently from an embedding in some coordinate system. In some applications, such as the example of the Borg and Leutner (1983) data, one could consider that property undesirable, since coordinates of objects are essential: they are the psychological part of the psychophysical function. Nevertheless, when fitting the city-block model without restrictions enforcing that the dimensions are indeed simple functions of the independent variables, a procedure often used, there is no guarantee whatsoever that the coordinates satisfy the expectations. Indeed, they often do not correspond exactly with the predicted dimensions, as was also clearly the case in our analysis of the Borg and Leutner data in Figure 5.1. In those situations, the coordinate-free representation with internal nodes can be useful in that it offers suggestions of the type of violations that occurred, as we have seen in the discussion of Figure 5.3. For a real test of inter-dimensional additivity, it might still be the best to follow simply Attneave (1950), who predicted observed differences between stimuli varying on two dimensions from observed differences

between stimuli varying only within dimensions and fitted a regression equation.

Coordinate-free models rely merely on distance and local relations. One may argue that these two elements are enough to navigate mentally through cognitive space. There is growing evidence that human navigation in physical space has two distinct means of keeping track of position and orientation during travel: landmark-based navigation and path integration (Klatzky, Beall, Loomis, Golledge, & Philbeck, 1999). While landmark-based navigation depends on some coordinate system - be it Cartesian or with polar coordinates - path integration is a mechanism that builds up a mental image of the trajectory traversed by encoding distances and turns, on the basis of sensed self-velocity, self-acceleration, and self-rotation. Thus, in some circumstances a network representation might have more psychological reality than a coordinate representation, which also often assumes more continuity in psychological space than is warranted by the data.

An important difference between networks for continuous city-block models (most often of low dimensionality) and networks for discrete city-block models (most often of high dimensionality) is the type of embedding needed to achieve an interpretable display. In the first case, the coordinates of the continuous solution often suffice, and no extra embedding is necessary (except for high-dimensional solutions). In the second case, we do need a form of multidimensional scaling for visualizing the nodes and the edges, which adds some arbitrariness to the final display, since several variations in analysis options are possible (type of fit function used, type of possible distance transformation specified, type of start configuration used, and so on). Nevertheless, the linking structure and the edge lengths are invariant. Therefore, when reporting a network, either the edge lengths or the feature parameters themselves should always be included. In addition, the goodness-of-fit between data and reconstructed network distance (network fit) is a more important consideration than the goodness-of-fit between reconstructed network distance and the distances in the visual display (embedding fit).

Network representation of feature structures offers a fruitful framework for theoretical comparison and practical use of a whole range of scaling and clustering methods. For example, our derivation of the common features model as a special case of the distinctive features model is a new result, owing to a more transparant notation than the one used in Sattath and Tversky (1987) and Carroll and Corter (1995). Since our network construction rule applies to continuous and discrete models alike, it turns out to be a unifying factor for understanding the relations between them. Figure 10 gives an overview of these relations. From top to bottom, Figure 10 has six levels, each adding some extra restriction to the model. One-step down from the most general continuous case, the distinctive features model arises from the restriction that coordinate values be binary (where the distance between the two values is not necessarily equal for all dimensions). At the same level of generality, we have Corter and Tversky (1986) extended similarity tree, which is an equivalent form, provided that we allow the tree being unresolved (for instance, if all features overlap without nesting, we can only have an extended similarity tree representation by reducing the tree to a bipartition). Continuing further down in Figure 10, we have:

- *Third level* (*common features model, additive tree*). The additive tree arises from the distinctive features model by the restriction that all features are either nested or disjoint and from the extended similarity tree by the exclusion of marked segments. As shown in this paper, the common features model arises from a restriction on the weights of the unique features, so that the total sum of all

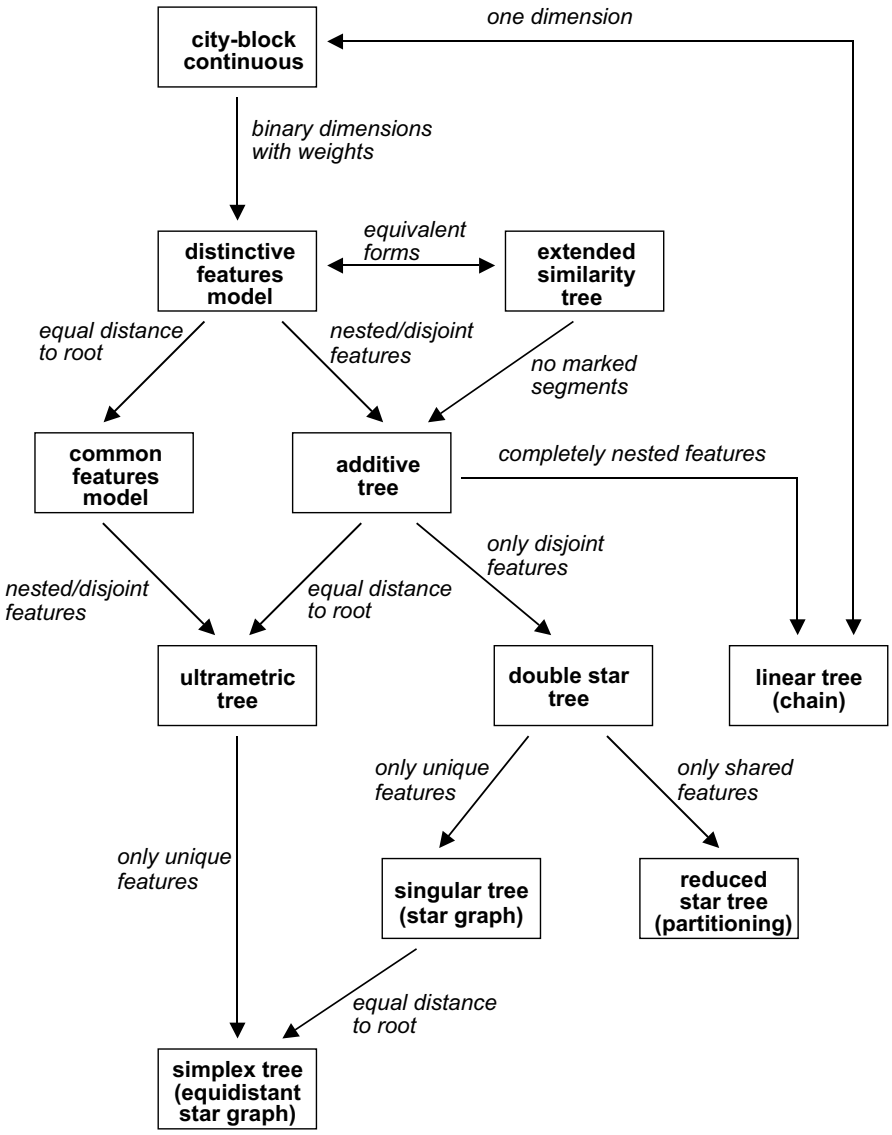

Figure 5.10: Relationships between city-block models.

feature weights is constant. Although Sattath and Tversky (1987) have stated that the distinctive features model and the common features model have the same level of generality, we believe that their argument runs into a contradiction with respect to the diagonal entries of the similarity matrix (see Heiser and Frank (2005), for a more detailed argumentation).

- *Fourth level* (*ultrametric tree, double star tree, linear tree*). The ultrametric tree arises from the additive tree by the restriction that all nodes are equidistant from the root (Carroll, 1976), but it is also a special case of the common features model in which all features are restricted to be either disjoint of nested (Carroll & Corter, 1995). If all features are completely nested and unique feature weights are zero except for one, we have a chain or linear tree (Sattath & Tversky, 1977), which is equivalent to a one-dimensional continuous distance model. As shown in this paper, if all shared features are disjoint, we have a double star tree.

- *Fifth level* (singular tree, reduced star tree). If all unique features in the double star tree are restricted to have zero weights, we obtain a reduced star tree, or simply a partitioning. If all shared features in the double star tree are restricted to have zero weights, we obtain a star graph (Carroll, 1976), also called a singular tree (Sattath & Tversky, 1977).

- *Last level* (*simplex tree*). If the leaves of a star graph are equidistant to the root, that is, if all unique feature weights are equal, we obtain the equidistant star graph, or simplex tree. Equal distances also arise if an ultrametric tree is completely unresolved (that is, the weights of all shared features reduce to zero).

It appears that all known discrete models of similarity fit well into this scheme. They are all special cases of the distinctive features model, and the general rule proposed in this paper produces their usual graphical representations, thanks to the introduction of internal nodes.

One model not mentioned in Figure 5.10, Tversky's (1977) Contrast Model, is decomposable into a symmetric and a skew-symmetric component, which are uncorrelated; the skew-symmetric component is linear and depends only on the sum of the feature weights (Zielman & Heiser, 1996),. As already noted by Tversky (1977), the symmetric version of the Contrast Model is equivalent to a distinctive features model. Therefore, the symmetric component of the Contrast Model fits in the scheme of Figure 5.10, and has a network representation. The model recently proposed by Navarro and Lee (2004), like the Contrast Model, is a linear combination of common and distinctive features, with the specification that each feature enters either into a common features combination rule or into a distinctive features combination rule. Converting the common component into a distinctive component with the specifications in Equation 5.15 in this paper, we have an additive combination of two distinctive features models, which again is a distinctive features model in the total feature space. In fact, this hybrid type of model is an example of Carroll's (1976) general strategy of decomposing a (dis)similarity matrix into the sum of multiple trees or other graphical structures. Although the sum of two additive trees is not a tree, it

still is a distinctive features model (albeit perhaps not a parsimonious one). Finally, it is of interest to mention the possibility to combine these discrete structures with generalized context models and geometric prototype models (M. D. Lee & Navarro, 2002; Nosofsky & Zaki, 2002; Verbeemen, Storms, & Verguts, 2004; Zaki et al., 2003).

Several aspects of network representations allow statistical refinement. Given the feature structure, estimation of the feature weights is a rather standard statistical problem. Frank and Heiser (in press a) have shown how to determine standard errors and confidence intervals for the feature weights in the distinctive features model. When the data can be split up in a training and a testing sample, it is also possible to calculate statistical accuracy of parameter estimates, do model tests and find a well-balanced compromise between model fit and model complexity when the features are unknown (Frank & Heiser, in press b). Similar work has been done by M. D. Lee (2001) for additive clustering, Navarro and Lee (2001) for the Contrast Model, and Frank and Heiser (2005) for additive trees. The emergence of a full-fledged methodology for city-block models owes much to their additivity, the very same property that makes them such attractive models for psychological similarity and difference.