Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 48 (2004) 301–309
Journal of Mathematical Psychology
http://www.elsevier.com/locate/jmp

Graph coloring, minimum-diameter partitioning, and the analysis of confusion matrices

Michael J. Brusco* and J. Dennis Cradit

Marketing, College of Business, Florida State University, Tallahassee, FL 32306-1110, USA

Received 24 May 2003; revised 30 April 2004

Available online 17 June 2004

Abstract

It is well known that minimum-diameter partitioning of symmetric dissimilarity matrices can be framed within the context of coloring the vertices of a graph. Although confusion data are typically represented in the form of asymmetric similarity matrices, they are also amenable to a graph-coloring perspective. In this paper, we propose the integration of the minimum-diameter partitioning method with a neighborhood-based coloring approach for analyzing digraphs corresponding to confusion data. This procedure is capable of producing minimum-diameter partitions with the added desirable property that vertices with the same color have similar in-neighborhoods (i.e., directed edges entering the vertex) and out-neighborhoods (i.e., directed edges exiting the vertex) for the digraph corresponding to the minimum partition diameter.

Keywords: Combinatorial data analysis; Partitioning; Graph coloring; Confusion matrices; Implicit enumeration

1. Introduction

Graph-coloring problems have fascinated mathematicians since at least the mid-1800s when Francis Guthrie suggested the problem to De Morgan (Dailey, 1978; May, 1965). Originally, investigators focused almost exclusively on the "four-color map problem", which concerned the coloring of countries on a two-dimensional map. In the early 1940s, Brooks' (1941) work began a transition from "map coloring" to, more generally, coloring the nodes of a network (or graph coloring). Since that time, research and theory has proliferated into a number of interesting research issues related to the coloring of graphs. An in depth history of graph-coloring problems is provided by Dailey (1978).

Because of their relationship to the grouping of proximity data, graph-coloring methods are particularly relevant to quantitative data analysis problems in the social sciences. Hubert (1974) and Baker and Hubert (1976) developed an important association between graph coloring and complete-link hierarchical clustering, a widely used technique in empirical psychological research (Loomis, 1982; Shepard & Arabie, 1979). The relevance of graph-coloring methods to the closely related nonhierarchical minimum-diameter partitioning problem (MDPP) is also well documented (Brusco, 2003; Hansen & Delattre, 1978; Hubert, 1974; Hubert, Arabie, & Meulman, 2001). Most notably, Hubert (1974, p. 297) observed: "One of the most interesting connections between graph theory and clustering is with respect to the colorability of a graph. … The implications of this relationship between clustering and node labeling are awesome given the phenomenal literature on the coloring problem".

In a somewhat different nonhierarchical application, Everett and Borgatti (1991, 1993) have deployed *role coloring* methods in the social network literature within the context of a formalization of regular equivalence (White & Reitz, 1983). The principles of this type of graph coloring are similar to those of MDPP situations; however, the coloring restrictions and objective criteria are somewhat different. Specifically, the objective of role coloring is to assign colors to actors such that actors of

*Corresponding author. Marketing Department, College of Business, Florida State University, Tallahassee, FL 32306-1110, USA. Fax: +1-850-644-8225.

E-mail address: mbrusco@cob.fsu.edu (M.J. Brusco).

0022-2496/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2004.05.001

similar color have the same type of role relations (links or connections with equivalent actors). Although role coloring (or role assignment) is not directly analogous to the problems we address herein, it does provide an example of a coloring-type problem that has recently received considerable attention in the quantitative social sciences (Roberts, 1998; Roberts & Sheng, 1999, 2001).

Although graph-coloring methods have, indirectly, been applied to the study of confusion data because of their relationship to complete-link hierarchical clustering, it is the premise of this paper that they can play an even larger, more formidable role in the analysis of confusion structure. In particular, we propose an integration of the MDPP with a neighborhood-based coloring algorithm that enables the asymmetric properties of the confusion data to be retained. The neighborhood-based algorithm attempts to produce a partition of vertices of the digraph corresponding to the minimum partition diameter, such that vertices of the same color are similar with respect to their entering and exiting edges. In other words, the neighborhood-based coloring procedure provides a straightforward and powerful tool for selecting a particular coloring of vertices for the digraph corresponding to the minimum partition diameter.

In Section 2 of this paper, we describe the general principles of graph coloring and its relationship to the MDPP. An implicit enumeration solution procedure is also presented and benchmarked in Section 2. Section 3 focuses on the application of coloring procedures to confusion matrices, where we develop the idea of neighborhood-based coloring, and propose an integration of this method with MDPP. A numerical example for synthetic data is used to demonstrate the procedure. The section concludes with an application to a large, well-known empirical confusion matrix. The paper concludes with a brief summary in Section 4.

2. Graph Coloring and the MDPP

2.1. Formulation of the MDPP

To develop the graph coloring problem and its relevance to classification, we define $V = \{v_1, v_2, \ldots, v_n\}$ as a set of n vertices. An edge of the graph is denoted by $e_q = (v_i, v_j)$, and the set of all edges is denoted by E. We also assume that a proximity relationship among the vertices is available in the form of a symmetric $n \times n$ dissimilarity matrix, $\mathbf{A}$, which contains nonnegative off-diagonal entries $a_{ij} = a_{ji}$ and an arbitrary main diagonal (that is ignored in the analyses). We use the term "dissimilarity" to reflect the fact that smaller (larger) matrix elements indicate greater (less) similarity between the pairs of objects. This convention is consistent with most of the previous literature in this area (Baker & Hubert, 1976; Brusco, 2003; Hansen & Delattre, 1978; Hubert, 1973, 1974).

The dissimilarity measures represent the weights of the edges of the complete graph $G(V, E)$, where $E = \{e_q = (v_i, v_j)$ for all $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n\}$. A partial graph of G, is denoted by $G_\delta(V, E_\delta)$, where the edge set $E_\delta = \{e_q = (v_i, v_j) | a_{ij} > \delta\}$. A K-coloring for the partial graph $G_\delta(V, E_\delta)$ consists of an assignment of one of K colors to each vertex such that no two adjacent vertices of the partial graph have the same color. In other words, if $\omega(v_i)$ denotes the color of vertex v_i, then $\omega(v_i) = \omega(v_j) \Rightarrow e_q = (v_i, v_j) \notin E_\delta$. The chromatic number of the partial graph $G_\delta(V, E_\delta)$, which is denoted $\gamma(G_\delta)$, is the minimum value of K that permits a K-coloring of $G_\delta(V, E_\delta)$.

Hansen and Delattre (1978) expanded this line of research by developing the relationship between the graph-coloring problem and the MDPP. The objective of MDPP is to partition the vertex set V into K subsets $(V_1, V_2, \ldots, V_K)$ such that the largest pairwise dissimilarity index across all subsets is minimized. Mathematically, the MDPP can be stated as follows:

$$\min \; Z_1 = \max_{k=1,\ldots,K} \left[\max_{(v_i,v_j) \in V_k} (a_{ij}) \right], \tag{1}$$

$$V_1 \cup V_2 \cup \cdots \cup V_{K-1} \cup V_K = V, \tag{2}$$

$$V_k \cap V_l = \varnothing \quad \forall k = 1, \ldots, K-1, \quad l = K+1, \ldots K, \tag{3}$$

$$|V_k| \geq 1 \quad \forall k = 1, \ldots, K. \tag{4}$$

The objective function (1) represents the partition diameter, and constraints (2)–(4) guarantee that the conditions for a partition of the vertices are met. Constraint (2) requires all vertices in V to belong to one or more subsets. Constraint set (3) ensures that a vertex is not assigned to more than one subset, and constraint set (4) forbids empty subsets by requiring the number of vertices in each subset (denoted $|V_k|$) to be at least one.

It has been observed that applying the complete-link hierarchical clustering algorithm and cutting the resulting tree at K clusters will often provide a good solution to MDPP; however, an optimal solution is not guaranteed (Baker and Hubert, 1976; Hansen and Delattre, 1978). In fact, Hansen and Delattre's results suggested that the complete-link algorithm seldom produced an optimal solution for MDPP for small values of K. An optimal solution for the MDPP can be obtained in polynomial time for the special case of $K = 2$ using a repulsion algorithm developed by Rao (1971); however, Hansen and Delattre demonstrated that MDPP is NP-hard for $K \geqslant 3$. As a plausible method for obtaining an optimal solution when $K \geqslant 3$, Hansen and Delattre proposed a branch-and-bound algorithm that incorporates a graph-coloring procedure developed

by Brown (1972). Alternative optimal solution procedures include mathematical programming (Rao, 1971), dynamic programming (Hubert et al., 2001), and implicit enumeration (Brusco, 2003). In this paper, we use the implicit enumeration approach described in the next subsection.

2.2. An implicit enumeration procedure for the MDPP

Brusco (2003) recently presented a number of enhancements for an implicit enumeration algorithm developed by Klein and Aronson (1991) for minimizing the sum of pairwise dissimilarities within clusters. Brusco also suggested that the algorithm could be modified for the MDPP, but he did not describe the procedure. The general steps of an implicit enumeration algorithm for MDPP are as follows:

Step 0: *Initialize*. Find an initial incumbent solution for the MDPP using a heuristic procedure and store the coloring of each vertex, $\Omega^* = \{\omega^*(v_1), \omega^*(v_2), \ldots, \omega^*(v_n)\}$ and the corresponding diameter, Z_1^*. Initialize the pointer $u = 0$, which specifies the index of the vertex to be colored. Initialize $\phi(k) = 0$ for $k = 1, \ldots, K$ as the number of vertices that are assigned color k. Initialize $\Gamma = \varnothing$ as the set of colors used in the partial coloring, and let $|\Gamma|$ denote the number of colors used.

Step 1: *Vertex reodering*. Reorder the vertices such that those vertices with the most edges in the graph corresponding to the solution, $\Omega*$, are placed first in the reordering.

Step 2: *Advance*. Advance the pointer by setting $u = u + 1$, which specifies the index of the current vertex to be colored.

Step 3: *Initial color*. Set the initial color for the selected vertex to the first color: $\omega(v_u) = 1$. Set $\phi(\omega(v_u)) = \phi(\omega(v_u)) + 1$ and $\Gamma = \Gamma \cup \omega(v_u)$.

Step 4: *Feasibility tests*. Determine whether or not the current partial coloring can produce a feasible solution to the MDPP that has a diameter less than Z_1^*.

Step 4a: Unused colors: the number of yet unused colors must equal or exceed the number of vertices remaining to be assigned a color. If $K - |\Gamma| > n - u$, go to Step 7.

Step 4b: Diameter test: If $a_{ju} \geq Z_1^*$ for any $1 \leqslant j \leqslant u - 1$ and $\omega(v_j) = \omega(v_u)$, then go to Step 7.

Step 4c: Unassigned vertex test: Compute $\xi_j = \min_{k=1,\ldots,K}(\max_{i=1,\ldots,u}(a_{ij}|\omega(v_i) = k))$ for $j = u+1, \ldots, n$. If $\xi_j \geq Z_1^*$ for any $u + 1 \leqslant j \leqslant n$, then go to Step 7.

Step 5: *Complete solution test*. If $u < n$, then go to Step 2.

Step 6: *Update incumbent*. $Z_1^* = \max_{i,j}(a_{ij}|\omega(v_i) = \omega(v_j))$ and $\Omega* = \Omega$.

Step 7: *Branching action*. If $\omega(v_u) = K$ or $(\phi(\omega(v_u)) = 1$ and $\phi(\omega(v_u) + 1) = 0$, then go to Step 9.

Step 8: *Recolor vertex*. Set $(\phi(\omega(v_u)) = (\phi(\omega(v_u)) - 1$. If $(\phi(\omega(v_u)) = 0$, then set $\Gamma = \Gamma - \omega(v_u)$. Set $\omega(v_u) = \omega(v_u) + 1$, $(\phi(\omega(v_u)) = (\phi(\omega(v_u)) + 1$, and $\Gamma = \Gamma \cup \omega(v_u)$. Return to Step 4.

Step 9: *Depth retraction*. Set $(\phi(\omega(v_u)) = (\phi(\omega(v_u)) - 1$. If $(\phi(\omega(v_u)) = 0$, then set $\Gamma = \Gamma - \omega(v_u)$. Set $u = u - 1$. If $u > 0$, then go to Step 7; otherwise return the incumbent solution Ω^* as an optimal solution to the MDPP and Stop.

The implicit enumeration algorithm begins with the development of an initial feasible solution from a heuristic procedure. Although Brusco (2003) recommended replications of a biased-sampling version of a complete-link clustering algorithm, we have obtained better initial bounds using the traditional complete-link algorithm followed by an exchange algorithm. The exchange algorithm sequentially evaluates the effect of replacing the current color of a vertex with each of the other possible colors and accepts any recoloring that improves the partition diameter. The exchange algorithm terminates when a complete pass through the vertices occurs without any recolorings. The reordering routine in Step 1 has also been incorporated in the algorithm, and can result in huge computational savings for some problems. By placing vertices with large dissimilarities early in the ordering of vertices, partial solutions are pruned earlier in the algorithm.

The vertex pointer, u, is advanced in Step 2, and this vertex is assigned the first color in Step 3. A series of feasibility tests for partial colorings are performed in Step 4. If the number of unused colors exceeds the number of vertices remaining to be assigned (Step 4a), then the current partial solution cannot ultimately lead to a feasible coloring. If vertex u shares a color assignment with a previously assigned vertex j, and if the dissimilarity between j and u equals or exceeds Z_1^*, then the partial solution cannot ultimately produce a coloring with a better partition diameter (Step 4b). The third feasibility test (Step 4c), which is extremely important for computational efficiency, tests each of the yet unassigned vertices with respect to its best possible coloring. If the best possible coloring for any unassigned vertex would yield a dissimilarity value that equals or exceeds Z_1^*, then the partial solution cannot ultimately yield a coloring with a better partition diameter.

If $u < n$ at Step 5, processing returns to Step 2 for selection of the next vertex. Otherwise, $u = n$, which suggests a complete coloring of the graph, and the

incumbent solution is updated in Step 6. Step 7 determines whether the current vertex is a candidate for recoloring in Step 8, or whether depth retraction (backtracking in the vertex ordering by decrementing the value of u) is required in Step 9. The algorithm terminates in Step 9 when the pointer is decremented to zero.

2.3. Benchmarking the performance of the algorithm

We investigated the efficiency of the implicit enumeration algorithm by applying it to a set of five synthetic dissimilarity matrices. Each of these matrices was constructed by randomly generating $n = 75$ vertices within a plane ring of the unit circle, and subsequently computing the Euclidean distance between each pair of vertices. We selected this generation process because such problems were characterized as 'difficult' by Hansen and Delattre (1978, p. 401). The implicit enumeration algorithm was applied to each of the five test problems using values of K ranging from 2 to 10. The algorithm was coded in Fortran and implemented on a 2.2 GHz Pentium IV PC with 1 GB of RAM. The computational results are displayed in Table 1.

Table 1 reports the average reduction in partition diameter when increasing the number of clusters from K-1 to K. As suggested by Hansen and Delattre (1978), this type of measure can provide the basis for a 'stopping rule' for the MDPP. The CPU times reported in Table 1 suggest that the implicit enumeration algorithm is a computationally feasible strategy, even for reasonably large values of K. Solution times for $K \leqslant 6$ were always less than three seconds, and most of the CPU times for $7 \leqslant K \leqslant 10$ were less than one minute. Clearly, solution-time variability was much greater for larger values of K, with one of the $K = 10$ test problems requiring nearly 45 minutes of CPU time. Nevertheless, the implicit enumeration scheme would meet the need for most MDPP's of practical size, and was used to provide solutions for subsequent experiments in this manuscript. Further, the general implicit enumeration paradigm is readily extensible to the related neighborhood-based coloring problem discussed below.

Table 1
Results for then MDPP implicit enumeration algorithm benchmarking experiments[a]

No. of clusters (K)	% reduction in partition diameter	Mean	Minimum	Maximum
2	—	0.78	0.78	0.80
3	23.48	0.80	0.78	0.81
4	19.30	0.80	0.78	0.83
5	21.76	0.88	0.82	0.98
6	13.86	1.41	0.79	2.55
7	12.59	22.31	1.69	55.77
8	10.59	43.12	1.32	112.76
9	7.86	70.20	3.80	219.53
10	6.97	732.42	33.86	2605.49

[a] For each number of clusters (except $K = 2$), the table reports the mean percentage reduction in partition diameter when increasing the number of clusters from K-1 to K. The mean, minimum, and maximum CPU times (across the five test problems) are also reported for each cluster size.

3. Graph coloring and MDPP procedures for confusion matrices

3.1. Applying MDPP procedures to confusion matrices

An $n \times n$ confusion matrix, **C**, provides a common representation of data for recognition tasks. An element of the matrix, c_{ij}, represents the number of times (or proportion of times) that response j was offered for the presentation of stimulus i. Confusion matrices have a similarity (rather than dissimilarity) interpretation because more frequent confusion of two objects suggests that they are more similar. More importantly, confusion matrices are typically asymmetric because, at least for some objects, j is more frequently a response to i than i is a response to j. Nevertheless, it is not uncommon in the empirical literature to transform a confusion matrix to a symmetric matrix and apply hierarchical clustering procedures (Hubert & Baker, 1977; Loomis, 1982; Morgan, Chambers & Morton, 1973). Hubert (1973) provides an excellent discussion of various approaches for applying single-link and complete-link hierarchical clustering methods to asymmetric proximity matrices.

To illustrate the transformation of an asymmetric confusion matrix (**C**) to an asymmetric dissimilarity matrix (**B**), and, subsequently, to a symmetric dissimilarity matrix (**A**), we use the synthetic data displayed in Table 2. The asymmetric dissimilarity matrix in Table 2 was obtained by subtracting each element of the original confusion matrix by the largest off-diagonal element in that confusion matrix ($c_{14} = 10$). In other words, $\mathbf{B} = [c_{\max}]_{6\times 6} - \mathbf{C}$, where $c_{\max} = \max_{i \neq j}(c_{ij})$ and $[c_{max}]_{6\times 6}$ is a 6×6 matrix with $c_{\max}$ as the off-diagonal entries (again, the main diagonal is ignored). Such a transformation is common practice in this type of analysis (Cho, Yang, & Hallett, 2000; Hubert, Arabie, & Meulman, 1997). The elements of the symmetric dissimilarity matrix (**A**) are obtained using $a_{ij} = a_{ji} = max(b_{ij}, b_{ji})$, which is consistent with Hubert's (1973) "strong completeness" transformation of asymmetric dissimilarity data.

A minimum-diameter partition for $K = 3$ clusters was obtained for the symmetric dissimilarity matrix in Table 2. The resulting partition of $(\{v_1, v_2, v_3\}\{v_4, v_5\}\{v_6\})$ yields a minimum diameter of $Z_1^* = 5$. As is frequently the case for empirical proximity matrices, this partition is not a unique optimum. Indeed, for many empirical

Table 2
Synthetic matrices for numerical demonstration

		Response					
	Stimulus	1	2	3	4	5	6
A synthetic 6×6 confusion matrix, **C**	1	—	6	5	10	7	1
	2	5	—	8	6	8	6
	3	9	5	—	9	8	3
	4	8	8	7	—	9	4
	5	7	9	2	5	—	7
	6	7	4	8	6	2	—
		1	2	3	4	5	6
An asymmetric dissimilarity matrix, **B**, formed by subtracting the elements in **C** from 10.	1	—	4	5	0	3	9
	2	5	—	2	4	2	4
	3	1	5	—	1	2	7
	4	2	2	3	—	1	6
	5	3	1	8	5	—	3
	6	3	6	2	4	8	—
		1	2	3	4	5	6
A symmetric dissimilarity matrix, **A**, formed based on "strong completeness" of matrix **B**.	1	—	5	5	2	3	9
	2	5	—	5	4	2	6
	3	5	5	—	3	8	7
	4	2	4	3	—	5	6
	5	3	2	8	5	—	8
	6	9	6	7	6	8	—

matrices, there can be a large number of partitions that produce the same minimum diameter. These alternative optima can vary widely in their configurations, and a systematic procedure for selecting a solution from the set of alternative optima would be especially valuable. The method developed in the next subsection is a plausible means for this selection process.

3.2. Neighborhood-based coloring of confusion matrices

The primary disadvantage of the traditional application of complete-link clustering or the MDPP to confusion matrices is that they do not directly capture the asymmetric information in the analysis. One possible remedy for this problem is to apply graph-coloring methods to a digraph corresponding to the asymmetric dissimilarity matrix, **B**. We define D as a digraph corresponding to the asymmetric dissimilarity matrix **B**. A partial graph of D, is denoted by $D_\delta(V, E_\delta)$, where the *directed* edge set $E_\delta = \{e_q = (v_i, v_j) | b_i > \delta\}$. We further define, for each vertex v_i in the digraph, the "in-neighborhood" as $N_I(v_i) = \{v_j | e_q = (v_j, v_i) \in E_\delta\}$, and the "out-neighborhood" as $N_O(v_i) = \{v_j | e_q = (v_i, v_j) \in E_\delta\}$. The goal is then to develop a coloring such that vertices with the same color have similar in-neighborhoods and out-neighborhoods. Although many criteria for achieving this objective are possible, we employ the following criterion in this paper:

$$\min : Z_2 = \sum_{[(i<j)|\omega(v_i)=\omega(v_j)]} \begin{pmatrix} \left|(N_I(v_i) \cup N_I(v_j)) \backslash (N_I(v_i) \cap N_I(v_j))\right| \\ +\left|(N_O(v_i) \cup N_O(v_j)) \backslash (N_O(v_i) \cap N_O(v_j))\right| \end{pmatrix}. \quad (5)$$

This objective function represents the total sum of inconsistencies in the in-neighborhoods and out-neighborhoods for vertex pairs of the same color. For each vertex pair, $(v_i, v_j) \ni \omega(v_i) = \omega(v_j)$, the index collects the number of vertices that are in $N_I(v_i)$ but not $N_I(v_j)$ and vice versa, as well as the number of vertices that are in $N_O(v_i)$ but not $N_O(v_j)$ and vice versa. The minimization of (5) can be accomplished using the same implicit enumeration scheme described in Section 2 with only minor modifications for the objective criterion. We incorporate the constraint that $[\{e_q = (v_i, v_j)\} \in E_\delta$ and/or $\{e_q = (v_j, v_i)\} \in E_\delta] \Rightarrow \omega(v_i) \neq \omega(v_j)$. Specifically, we propose integration of the MDPP with the neighborhood-based coloring of the digraph via the following two-stage procedure:

> Stage 1: Obtain a minimum diameter partition using for matrix **A** using the algorithm described in Section 2.2 and let $\delta = Z_1^*$.
>
> Stage 2: Form the digraph, $D_\delta(V, E_\delta)$, for the asymmetric dissimilarity matrix **B** and obtain a neighborhood coloring of the digraph by minimizing (5).

This two-stage process will yield a minimum-diameter partition that also provides a coloring such that vertices of the same color have similar neighborhoods. The need for Stage 2 arises because of the myriad of minimum-diameter partitions that are typically available for empirical matrices. It is also possible to repeat Stage 2 after incremental increases in δ to determine whether or not large decreases in Z_2 can be achieved at the expense of small increases in Z_1.

To illustrate the two-stage procedure for coloring confusion matrices, we return to the data from Table 2. The colored digraph for the minimum-diameter partition obtained in Stage 1 is shown in Fig. 1 (to improve readability, we use digraphs with "double-arrowed" edges instead of two separate edges). This coloring yields values of $Z_1^* = 5$ and $Z_2 = 7$. Stage 2 provides the colored digraph shown in Fig. 2, which corresponds to the partition $(\{v_1, v_3, v_4\}\{v_2, v_5\}\{v_6\})$. Although this coloring also produces a minimum diameter of $Z_1^* = 5$, it has a smaller value for the neighborhood index ($Z_2 = 3$). For illustrative purposes, the computation of Z_2 is shown in Table 3.

Stage 2 of the solution procedure resulted in a modification of the Stage 1 solution by swapping the colors of v_2 and v_4. These recolorings occurred because,

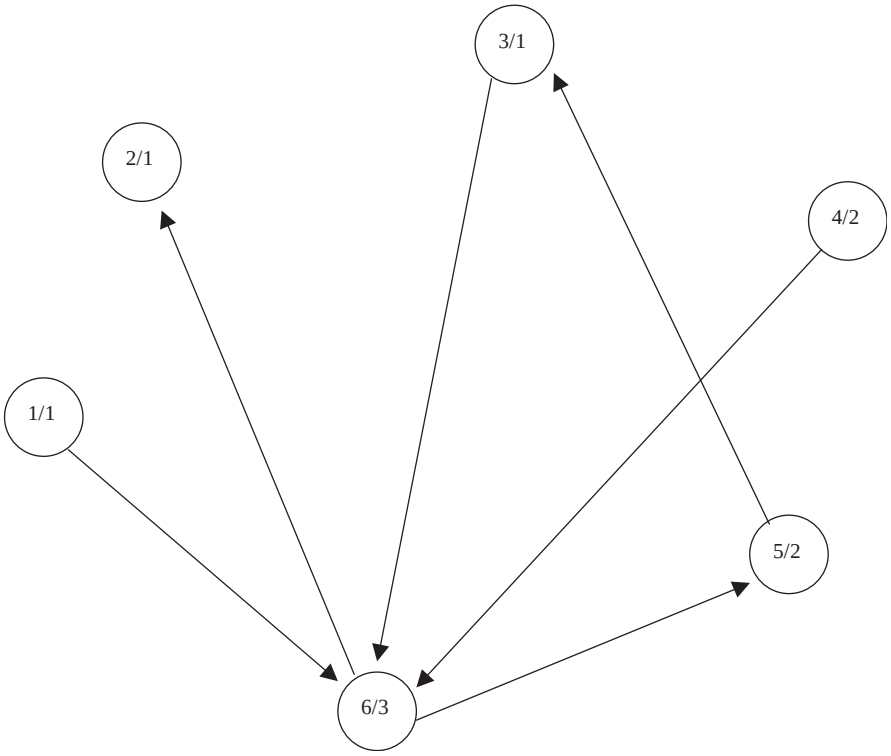

Fig. 1. 3-coloring of the digraph, D_5, for matrix **B** of Table 2 resulting from Stage 1. The convention within each vertex is: vertex number/color of the vertex.

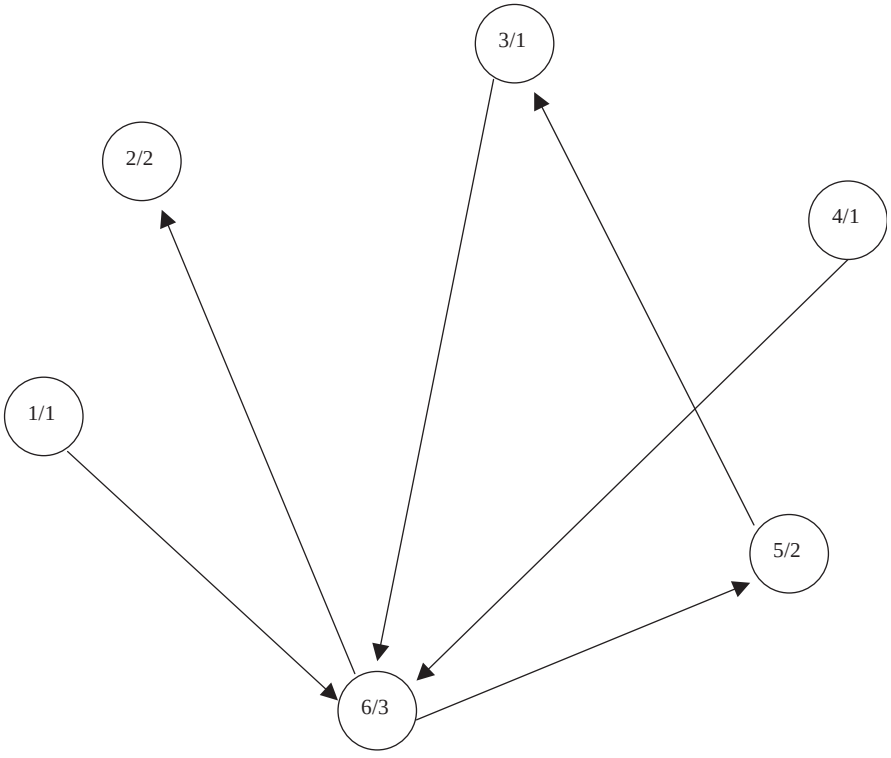

Fig. 2. 3-coloring of the digraph, D_5, for matrix **B** of Table 2 resulting from Stage 2. The convention within each vertex is: vertex number/color of the vertex.

relative to v_4, the neighborhood of v_2 is more comparable to the neighborhood of v_5. Similarly, relative to v_2, the neighborhood of v_4 is more comparable to the neighborhoods of v_1 and v_3. As shown in Table 3, v_1 and v_4 have identical in-neighborhoods and out-neighborhoods, and both of these vertices have the same

out-neighborhood as v_3. Because $\{v_5\}$ is in the in-neighborhood of v_3, but not v_1 and v_4, there is an inconsistency contribution of $+2$ associated with the vertices of color $k=1$. The vertices of color $k=2(\mathrm{v}_2\mathrm{and}\mathrm{v}_5\}$ have identical in-neighborhoods and only slightly different out-neighborhoods (with an inconsistency contribution of $+1$ because $\{v_3\}$ is in the out-neighborhood of v_5 but not v_2).

To provide an interpretation of the Stage 2 partition with respect to the original confusion matrix, **C**, it is helpful to consider the complement, $\bar{D}_5$, of the 3-colored digraph in Fig. 2, which is depicted in Fig. 3. The vertices v_2 and v_5 in Fig. 3, which are each assigned color $k=2$, are connected by a double-arrowed edge. This indicates that the dissimilarities (in both directions) for this pair of vertices with color $k=2$ are less than or equal to 5. Alternatively, this means that the confusion entries (both directions) for this vertex pair are greater than or equal to 5. We observe from Table 2 that $c_{25}=8$ and $c_{52}=9$, revealing high confusion among objects 2 and 5.

The digraph in Fig. 3 also shows that objects 2 and 5 are similar in their confusion structures with other objects. Vertices v_2 and v_5 are each connected to v_1 and v_4 by double-arrowed edges, which reveals that stimuli 2 and 5 are both highly confused with stimuli 1 and 4 in both directions. Fig. 3 also displays a directed edge from v_2 to v_6 and from v_5 to v_6, which indicates that stimulus 6 is often provided as an incorrect response for stimuli 2 and 5 ($c_{26}=6$ and $c_{56}=7$), but stimuli 2 and 5 are less frequently offered as incorrect responses to stimulus 6 ($c_{62}=4$ and $c_{65}=2$). The only inconsistency for vertices v_2 and v_5 is that v_2 is connected to v_3 by a double-arrowed edge, but v_5 is only connected by a directed edge from v_3. This indicates that although stimuli 2 and 5 are both frequently incorrect responses for stimulus 3 ($c_{32}=5$ and $c_{35}=8$), stimulus 3 is frequently an incorrect response for stimulus 2 ($c_{23}=8$) but not for stimulus 5 ($c_{53}=2$).

Table 3
Computation of Z_2 for the 3-coloring of the digraph in Fig. 2.

Vertex (v_i)	$N_I(v_i)$	$N_O(v_i)$	Color, $\omega(v_i)$
v_1	$\{\varnothing\}$	$\{v_6\}$	1
v_2	$\{v_6\}$	$\{\varnothing\}$	2
v_3	$\{v_5\}$	$\{v_6\}$	1
v_4	$\{\varnothing\}$	$\{v_6\}$	1
v_5	$\{v_6\}$	$\{v_3\}$	2
v_6	$\{v_1, v_3, v_4\}$	$\{v_2, v_5\}$	3

Computation of index

Pair	In	Out
$\{v_1, v_3\}$	+1	+0
$\{v_1, v_4\}$	+0	+0
$\{v_3, v_4\}$	+1	+0
$\{v_2, v_5\}$	+0	+1
Sum =	2	1
Total sum =		3

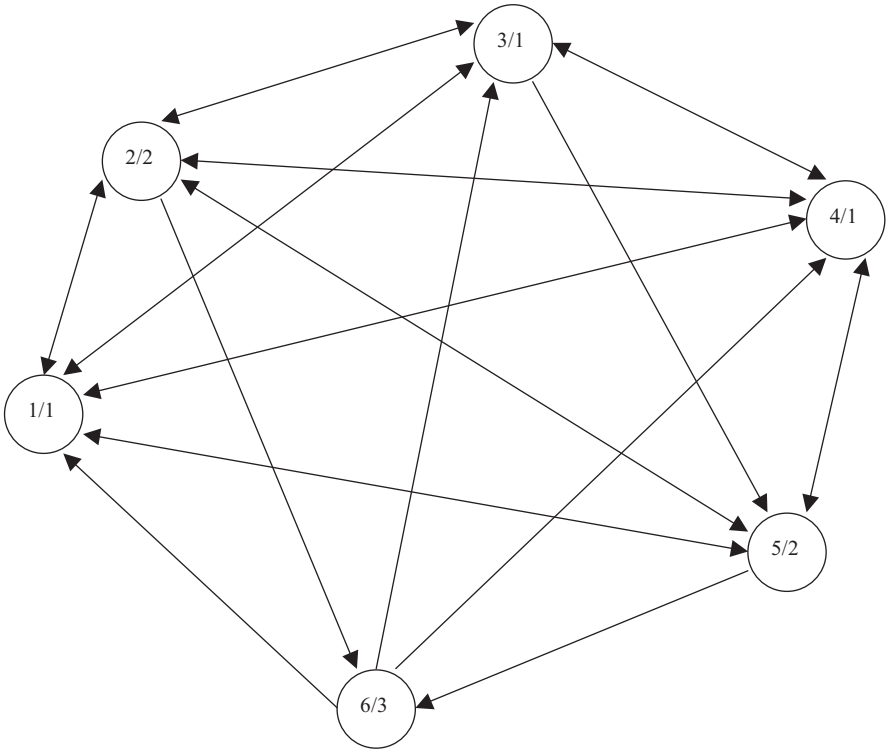

Fig. 3. The complement, $\bar{D}_5$, of the digraph in Fig. 2. The convention within each vertex is: vertex number/color of the vertex.

3.3. An application to an empirical confusion matrix

Although illustrative, the synthetic confusion matrix used in previous subsections cannot demonstrate the significant benefits associated with the two-stage coloring procedure. Therefore, we applied the procedure to Morse code confusion data among the 26 letters and 10 digits $0, 1, 2, \ldots, 9 (n = 36)$, as reported by Hubert (1987, pp. 164-165) based on data originally collected by Rothkopf (1957). The resulting 36×36 confusion matrix is one of the largest available in the literature, and was particularly useful for demonstrating the computational plausibility of the approach presented herein.

We applied the two-stage procedure to the Morse code data for $2 \leqslant K \leqslant 12$ clusters, and the results are summarized in Table 4. Partition diameter monotonically decreased as the number of clusters increased, and a particularly large percentage decrease was observed when moving from 9 to 10 clusters. The neighborhood index was smallest at two clusters, reached a maximum at five clusters, and monotonically decreased across the range from 6 to 12 clusters. Stage 2 of the procedure provided a reduction in Z_2 (without penalty of Z_1) for all values of K, except $K = 2$. In some cases, the improvement in Z_2 was considerable (e.g., a 19.1% reduction in Z_2 for $K = 6$). It is important to recognize, however, that the reduction in Z_2 should not be evaluated independently from the reduction in Z_1. For example, when increasing K from 9 to 10, the value of Z_1 decreases substantially from 68 to 62. Although the decrease in Z_2 from 949 to 881 might seem small, it must be recognized that the definition of neighborhood is now much tighter because of the significant reduction in the diameter.

Because of the particularly small increase in partition diameter when decreasing K from 11 to 10, as well as the large increase in diameter when decreasing K from 10 to 9, we selected the 10-cluster solution for evaluation. As observed by one of the reviewers of this manuscript, good performance in the diameter criterion for K clusters is often associated with somewhat poor performance at $K - 1$ clusters. Table 5 presents the 10-cluster solutions from Stages 1 and 2, which demonstrate the reassignments that occur. The first seven clusters of the two partitions are identical; however, clusters 8–10 are slightly different. Cluster 9 from the Stage 1 solution was split apart in the Stage 2 solution, with the letter "L" becoming a component of cluster 8 in the Stage 2 partition, and the letter "R" moving to cluster 10 of the Stage 2 partition. One advantage of this change from an interpretability standpoint is that the Morse code symbol for the letter "L" has four characters (•-••), as do the symbols for the letters "B", "C", and "X" in cluster 8. In the Stage 1 solution, the letter "L" had been grouped with two letters, "D" and "R", which have Morse code symbols that consist of only three characters. The letters "H" and "S" from cluster 10 of the Stage 1 solution were joined with "D" to form cluster 9 of the Stage 2 solution, and the letter "U" remained in cluster 10 of the Stage 2 solution but was grouped with "R" instead of "D" and "S".

A couple of additional points regarding the two-stage procedure are necessary. First, the improvement realized from Stage 2 is dependent upon the ordering of the vertices, which affects the first minimum-diameter partition found (and, therefore, stored) in Stage 1. For example, there are alternative minimum-diameter partitions for the 10-cluster solution that would have revealed more significant refinement in the Stage 2 solution. Thus, the benefit of Stage 2 is that it does not leave the identified minimum-diameter partition to chance but, instead, provides a rational basis for choosing among minimum-diameter partitions. Second, although we have restricted our analysis in Stage 2 to requiring minimum-diameter partitions, the quantitative analyst could possibly relax this constraint to determine whether a small sacrifice in diameter would permit a large improvement in the neighborhood index.

Table 4
Experimental results for the Morse code data

	Stage 1 solution		Stage 1 solution	
No. clusters	Z_1	Z_2	Z_2	% reduction
2	82	392	392	0.0
3	81	1186	1070	9.8
4	79	2134	1944	8.9
5	76	2131	2064	3.3
6	74	1995	1613	19.1
7	72	1629	1507	7.5
8	69	1346	1240	7.9
9	68	1123	977	13.0
10	62	949	881	7.2
11	61	719	646	10.2
12	60	648	531	18.1

Table 5
10-cluster partitions for the Morse code data from Stage 1 and Stage 2

Cluster number	Stage 1 solution	Stage 2 solution
1	{A, I}	{A, I}
2	{E, T}	{E, T}
3	{M, N}	{M, N}
4	{F, V, 4, 5}	{F, V, 4, 5}
5	{G, K, O, W}	{G, K, O, W}
6	{J, P, 2, 3}	{J, P, 2, 3}
7	{Q, 1, 8, 9, 0}	{Q, 1, 8, 9, 0}
8	{B, C, X, 6, 7}	{B, C, L, X, 6, 7}
9	{D, L, R}	{D, H, S}
10	{H, S, U}	{R, U}

4. Conclusions

4.1. Summary

In this paper, we have developed some possible applications for graph coloring and minimum-diameter partitioning methods with respect to the analysis of confusion matrices. Specifically, we propose that minimum-diameter partitioning methods can be integrated with neighborhood-based coloring methods for digraphs of asymmetric dissimilarity matrices. The advantage of our two-stage procedure is that it provides a systematic means for finding minimum-diameter partitions that also provide consistent edge neighborhoods for vertices within the same cluster (i.e., the same color). We have presented an implicit enumeration scheme for obtaining the minimum-diameter partition in Stage 1 of the procedure, and this scheme was easily adapted to color the minimum-diameter digraph in Stage 2 so as to minimize the neighborhood index. The implicit enumeration scheme efficiently provides guaranteed optimal solutions for both stages of the procedure.

4.2. Modeling alternatives and extensions

This paper has focused on a graph-coloring perspective for the analysis of confusion matrices. We have implemented our algorithms using compact-clustering criteria based on partition diameter; however, it is possible to employ objective function criteria other than those used in Stages 1 and 2 of the proposed method. Such criteria could encompass a wide range of indices corresponding to homogeneity within clusters and separation among clusters. For example, the quantitative analyst could replace the neighborhood-based clustering procedure in Stage 2 with a weighted-edge version of the neighborhood index, or a within-cluster sum of dissimilarities criterion (either with or without adjustment for cluster size). Fortunately, the branch-and-bound algorithm described herein can be easily be adapted for these and many other objective functions.

References

Baker, F. B., & Hubert, L. J. (1976). A graph–theoretic approach to goodness of fit in complete-link hierarchical clustering. *Journal of the American Statistical Association*, *71*, 870–878.

Brooks, R. L. (1941). On colouring the nodes of a network. *Proceedings of the Cambridge Philosophical Society*, *37*, 194–197.

Brown, J. R. (1972). Chromatic scheduling and the chromatic number problem. *Management Science*, *19*, 456–463.

Brusco, M. J. (2003). An enhanced branch-and-bound algorithm for a partitioning problem. *British Journal of Mathematical and Statistical Psychology*, *56*, 83–92.

Cho, R. Y., Yang, V., & Hallett, P. E. (2000). Reliability and dimensionality of judgments of visually textured materials. *Perception & Psychophysics*, *62*, 735–752.

Dailey, D. P. (1978). *Graph coloring by humans and machines: A polynomial complete problem solving task*. Doctoral dissertation, University of Colorado.

Everett, M. G., & Borgatti, S. P. (1991). Role colouring a graph. *Mathematical Social Sciences*, *21*, 183–188.

Everett, M. G., & Borgatti, S. P. (1993). An extension of regular colouring of graphs to digraphs, networks and hypergraphs. *Social Networks*, *15*, 237–254.

Hansen, P., & Delattre, M. (1978). Complete-link cluster analysis by graph coloring. *Journal of the American Statistical Association*, *73*, 397–403.

Hubert, L. (1973). Min and max hierarchical clustering using asymmetric proximity measures. *Psychometrika*, *38*, 63–72.

Hubert, L. J. (1974). Some applications of graph theory to clustering. *Psychometrika*, *39*, 283–309.

Hubert, L. J. (1987). *Assignment methods in combinatorial data analysis*. New York: Marcel Dekker.

Hubert, L., Arabie, P., & Meulman, J. (1997). Linear and circular unidimensional scaling for symmetric proximity matrices. *British Journal of Mathematical and Statistical Psychology*, *50*, 253–284.

Hubert, L., Arabie, P., & Meulman, J. (2001). *Combinatorial data analysis*: *Optimization by dynamic programming*. Philadelphia: Society for Industrial and Applied Mathematics.

Hubert, L. J., & Baker, F. B. (1977). The comparison and fitting of given classification schemes. *Journal of Mathematical Psychology*, *16*, 233–253.

Klein, G., & Aronson, J. E. (1991). Optimal clustering: A model and method. *Naval Research Logistics*, *38*, 447–461.

Loomis, J. M. (1982). Analysis of tactile and visual confusion matrices. *Perception & Psychophysics*, *31*, 41–52.

May, K. O. (1965). The origin of the four-color conjecture. *Isis*, *56*, 346–348.

Morgan, B. J. T., Chambers, S. M., & Morton, J. (1973). Acoustic confusion of digits in memory and recognition. *Perception & Psychophysics*, *14*, 375–383.

Rao, M. R. (1971). Cluster analysis and mathematical programming. *Journal of the American Statistical Association*, *66*, 622–626.

Roberts, F. S. (1998). Role assignments and indifference graphs. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology* (pp. 33–46). Mahwah, NJ: Lawrence Erlbaum Associates.

Roberts, F. S., & Sheng, L. (1999). Role assignments. In Y. Alavi, D. Lick, & A. Schwenk (Eds.), *Combinatorics, graph theory, and algorithms* (pp. 729–745). Kalamazoo, MI: New Issues Press.

Roberts, F. S., & Sheng, L. (2001). How hard is it to determine if a graph has a 2-role assignment? *Networks*, *37*, 67–73.

Rothkopf, E. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*, 94–101.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87–123.

White, D. R., & Reitz, K. P. (1983). Graph and semigroup homomorphism on networks of relations. *Social Networks*, *5*, 193–235.