

Cluster Spectroscopy

Leonid Andreev, Equicom, Inc., Scottsdale, AZ 85255, U.S.A.

We propose a cluster spectroscopy method that provides visualization of the process of hierarchical clustering according to the earlier proposed method for evolutionary transformation of similarity matrices. The examples discussed in this paper demonstrate that aggregations of objects into subclusters are accompanied by the emergence of signals described by Gaussian curves whose peaks correspond to maximums in the rates of respective subcluster aggregation processes. Heights of peaks that correspond to formation of higher hierarchy clusters represent the totals of heights of peaks reflecting the formation of lower hierarchy subclusters. Thus registered cluster spectra consisting of multiple Gaussian curves reflect the process of hierarchical clustering of objects described by a similarity matrix under unsupervised automated evolutionary transformation.

In the course of the automated unsupervised evolutionary transformation of similarity matrices (see Methods below), a set of objects in a high-dimensional space of parameters is transformed into two points with coordinates of 0 and 1 on a similarity scale (the so-called bifocal compression of information). Then, each of the two groups of objects corresponding to the points of the initial bifocal compression individually undergoes a new cycle of bifocal compression within that group, and so on. This process of iterative and successive cycles of bifocal compression of each newly emerging subgroup of objects results in a complete hierarchical clustering of an entire set of objects under analysis, and the analysis output is presented in the form of a hierarchical tree and a dendrogram (see MeaningFinder 2.2 User's Guide on this web site).

In this paper, we present a brief explanation of the principle of cluster spectroscopy – a new method that enables a data analyst to

watch the formation and convergence of subclusters in the course of one cycle of evolutionary transformation of a similarity matrix. In other words, while the entire clustering process consists of evolutionary transformation cycles alternating with division of objects into two groups, the cluster spectroscopy provides the visualization of the clustering process within a span of one transformation cycle resulting in bifocal compression. Each object in a system under analysis has its individual cluster spectrum which can be computed according to our modification of the Shannon entropy equation:

$$e_i(T) = \frac{2}{n} \times \sum_{i=1}^n [S_{i,k}]_{T,C} \times \log_2 [S_{i,k}]_{T,C} \quad (1),$$

where $e_i(T)$ is a signal of the individual spectrum of object i at T number of transformations, and $[S_{i,k}]_{T,C}$ is a coefficient of similarity between objects i and k at T number of transformations and C value of Contrast (see Methods).

We will demonstrate the basic mechanism of cluster spectroscopy by the example of analysis of a 20-vertex graph described in Table 1 which is constructed by the type of a similarity matrix wherein vertices that have a common edge are represented as “1”, and those that do not share an edge, as “0” (this table can also be constructed by the type of a distance matrix, i.e. a dissimilarity matrix, by replacing 0's by 1's and vice versa, and the analysis result will be the same). By using the XR-metric (see Methods), we can compute a similarity matrix of the objects (vertices) presented in this table. The

Table 1. Description of a 20-vertex graph. See explanation above.

Vertex No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
2	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0	1	0	0	0	0
3	0	1	1	0	1	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0
4	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1
5	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	1	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
8	1	0	0	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	1
9	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	1
10	0	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	1	0
11	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	0	1	0	1
12	0	1	0	0	0	0	1	1	0	1	0	1	1	1	1	0	1	0	0	1
13	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1
14	0	0	1	0	0	0	0	0	1	1	1	1	0	1	0	1	0	1	1	0
15	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	1	1	1	0	1
16	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0

17	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0
18	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1
19	1	0	0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	0	1	0
20	0	0	0	1	0	0	0	1	1	0	0	1	1	0	1	0	0	1	0	1

similarity matrix was computed by hybridization of monomer similarity matrices (see **Methods**) Based on the resulting similarity matrix, partial cluster spectra can be obtained by applying the “Egoentropy” function of MeaningFinder 2.2 (a free trial copy of the program is available for downloading from this web site). Cluster spectra can be obtained for any set of objects described by any number of parameters. Fig. 1 shows the total spectra of all 20 vertices of the graph under analysis, and Fig. 2 demonstrates each

object’s individual spectrum in the $e_i(T) - \ln T$ coordinates. As is seen, the spectra are represented by sets of Gaussian curves whose peak maximums correspond to various e_i values. The $e_i(T)$ values for each peak’s maximum emerging in the course of analysis of the entire graph are shown in Table 2 and demonstrate the dynamics of the hierarchical clustering of this system of data points. At T (number of

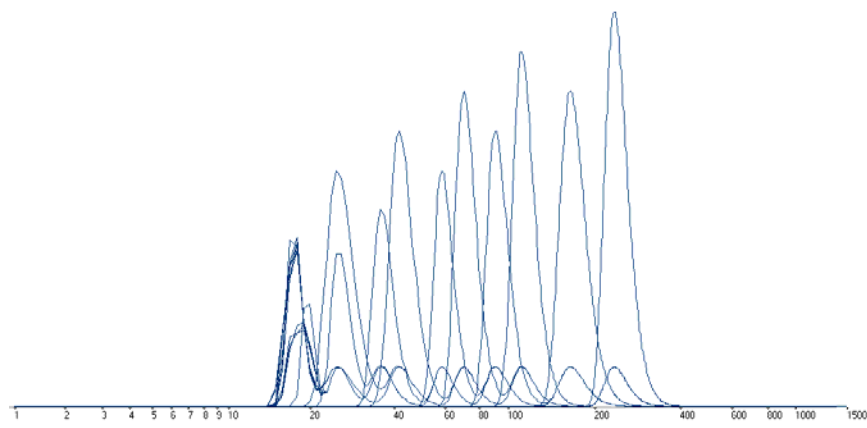


FIG. 1. Total cluster spectra of the graph described in Table.1.

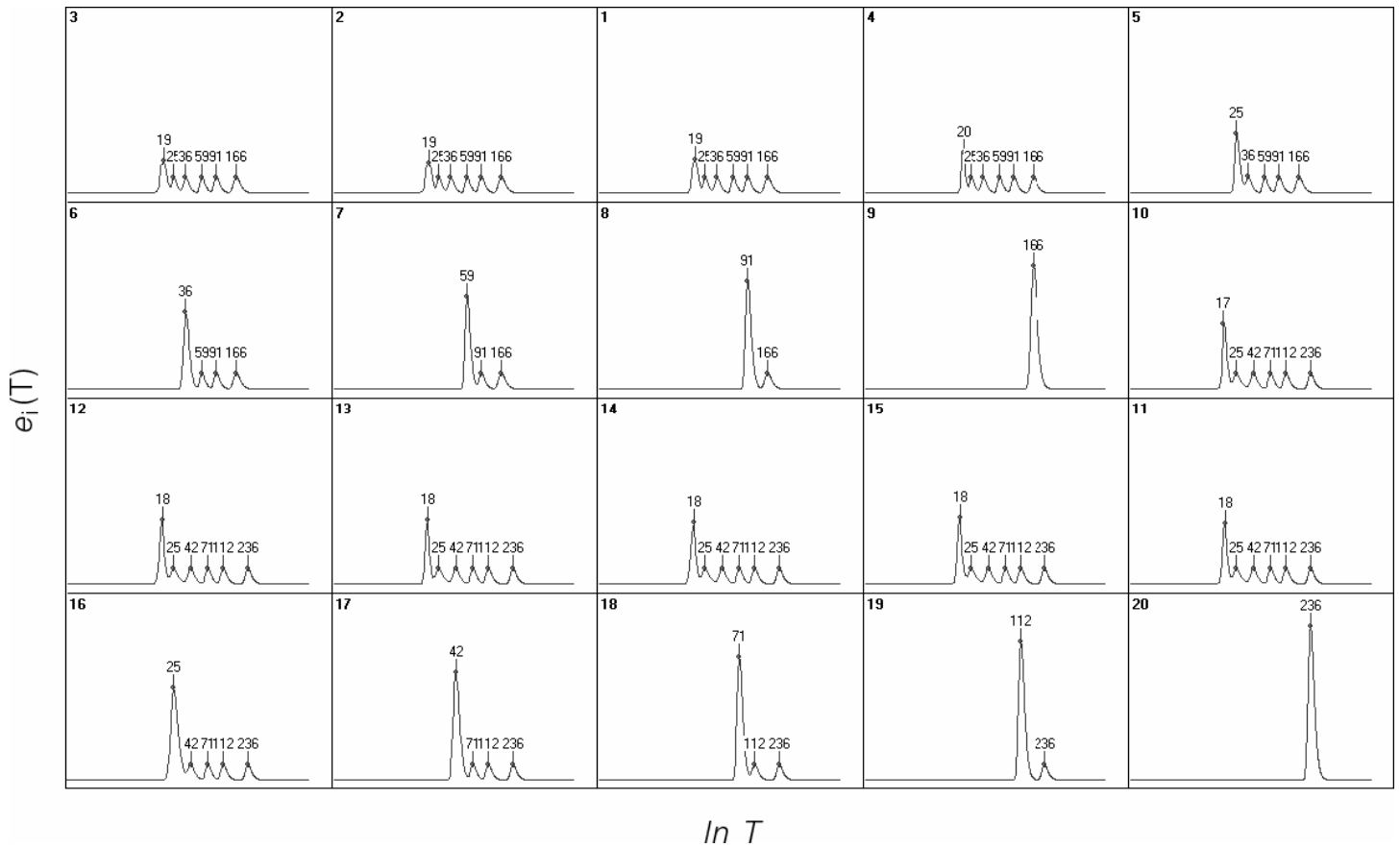


FIG. 2. Individual spectra of the graph described in Table 1.

transformations) of 17-18, vertices 10 through 15 join into one subcluster ($e_i = 0.210 - 0.227$); and vertices 1 - 4 form another subcluster ($e_i = 0.102 - 0.140$) at $T = 19-20$. Thus, we see the formation of two centers that will successively “attract” other

vertices. For instance, the subcluster of vertices 10 - 15 further eventually attracts, one by one, vertices 16, 17, 18, 19, and 20; the subcluster of vertices 1 - 4 further incorporates vertices 5, 6, 7, 8, and 9.

Table 2. Numbers of transformations and e_i values at peak maximums of cluster spectra of the graph described in Table 1.

		Numbers of transformations at peak maximums										Total	
		17-18	19-20	25	36	42	59	71	91	112	166= α		236= β
Vertex No.	1		0.110	0.050	0.050		0.050		0.050		0.050		0.360
	2		0.102	0.052	0.054		0.053		0.053		0.053		0.367
	3		0.110	0.052	0.054		0.053		0.053		0.053		0.375
	4		0.140	0.052	0.054		0.053		0.053		0.053		0.405
	5			0.205	0.056		0.053		0.053		0.053		0.420
	6				0.265		0.053		0.053		0.053		0.424
	7						0.318		0.053		0.053		0.425
	8								0.371		0.053		0.424
	9										0.425		0.425
	10	0.220		0.053		0.050		0.050		0.050		0.050	0.473
	11	0.210		0.053		0.053		0.053		0.053		0.053	0.475
	12	0.220		0.053		0.053		0.053		0.053		0.053	0.485
	13	0.221		0.053		0.053		0.053		0.053		0.053	0.486
	14	0.212		0.053		0.053		0.053		0.053		0.053	0.477
	15	0.227		0.053		0.053		0.053		0.053		0.053	0.492
	16			0.317		0.055		0.053		0.053		0.053	0.531
	17					0.371		0.053		0.053		0.053	0.531

	18							0.424		0.054		0.053	0.531
	19									0.0477		0.053	0.530
	20											0.531	0.531

As is seen from Table 2, the formation of two primary conglomerates of objects, as a start of hierarchical clustering, and the subsequent attraction of the remaining objects to one or another primary subcluster display an important peculiarity of the whole process: during the formation of the primary subclusters, the $e_i(T)$ values for individual objects do not depend on each other; whereas at the stage of adjoining additional objects to a primary subcluster, the $e_i(T)$ value of each new adjoining object equals the total of the $e_i(T)$ values of all other objects of that subcluster. For instance, at $T = 25$, vertex No. 5 joins the subcluster of vertices No. 1 – 4 at; its $e_i(25) = 0.205$, whereas $\sum_{i=1}^4 e_i(25)$ equals 0.206.

Hierarchical clustering of the 20-vertex graph described in Table 1 involved ten events of adjunction of vertices to the earlier formed primary subclusters, and in each case of adjunction the aforesaid regularity occurred with an accuracy of 99-100%.

Another important characteristic of thus obtained cluster spectra is the fact that the formation of clusters originating from the two primary groups of objects occurs independently for each group and ends by the adjunction of an object that is least similar to other objects of a respective group. For instance, as is seen from Table 2, the adjunction of the last object, vertex No. 9, to the group of vertices No. 1 – 4, occurs at $T = 166$; while in another group, it takes 236 transformations to adjoin its last object, vertex No. 20. The completion of clustering at the moment of adjunctions of the last object in each group (subcluster) is reflected in individual spectra of each vertex, and the numbers of transformations (in the

discussed case, 166 and 236) resulting in the last object adjunctions are marked as α and β , respectively. The resulting subclusters are marked in the same way, accordingly.

The foregoing is only a brief explanation of the principles of cluster spectroscopy. Cluster spectroscopy of real-life systems of data may involve lots variations revealing the specifics of a system under analysis. For instance, some peaks in cluster spectra may appear to be quite broad – some peak's width at a certain section of height by far exceeding that of other peaks in the same spectrogram – which means that the objects they represent have high affinity to both α - and β -group objects. In some cases, convergence of subclusters within each of the two primary groups occurs concurrently, rather than successively as was demonstrated in Table 2, and the concurrently formed subclusters merge only at the late stages of entire hierarchical clustering. In other words, cluster spectroscopy provides scrupulous and highly accurate insight into relationships between objects of complex systems.

To give an idea of the application potential and the new opportunities provided by cluster spectroscopy in analysis of complex systems, we will present a few examples. Figures 3A – 3D illustrate the partitioning of four simple graphs (α - and β -group vertices are indicated as dark and open circles, respectively). Figures 4 – 7 are the cluster spectra of the graphs, which clearly show which subgroup (α or β) each individual spectrum belongs to and demonstrate the dynamics of subcluster convergence.

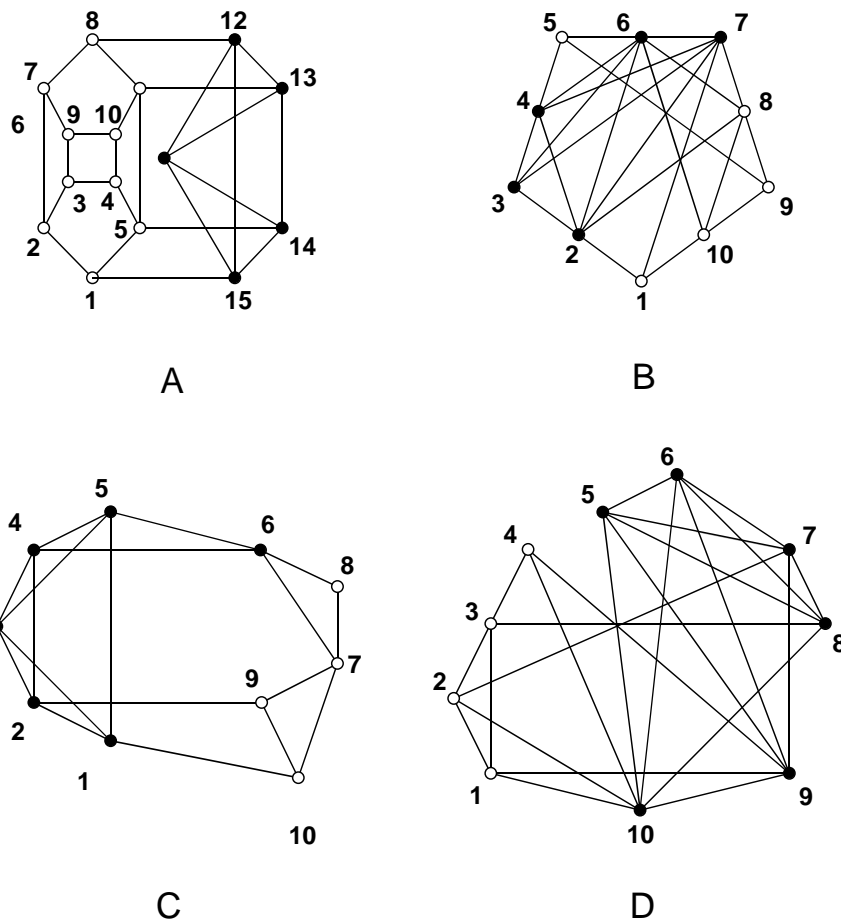


FIG. 3A – 3D. Four graphs whose partitioning was performed by cluster spectroscopy (open and dark circles in each graph indicate different subclusters).

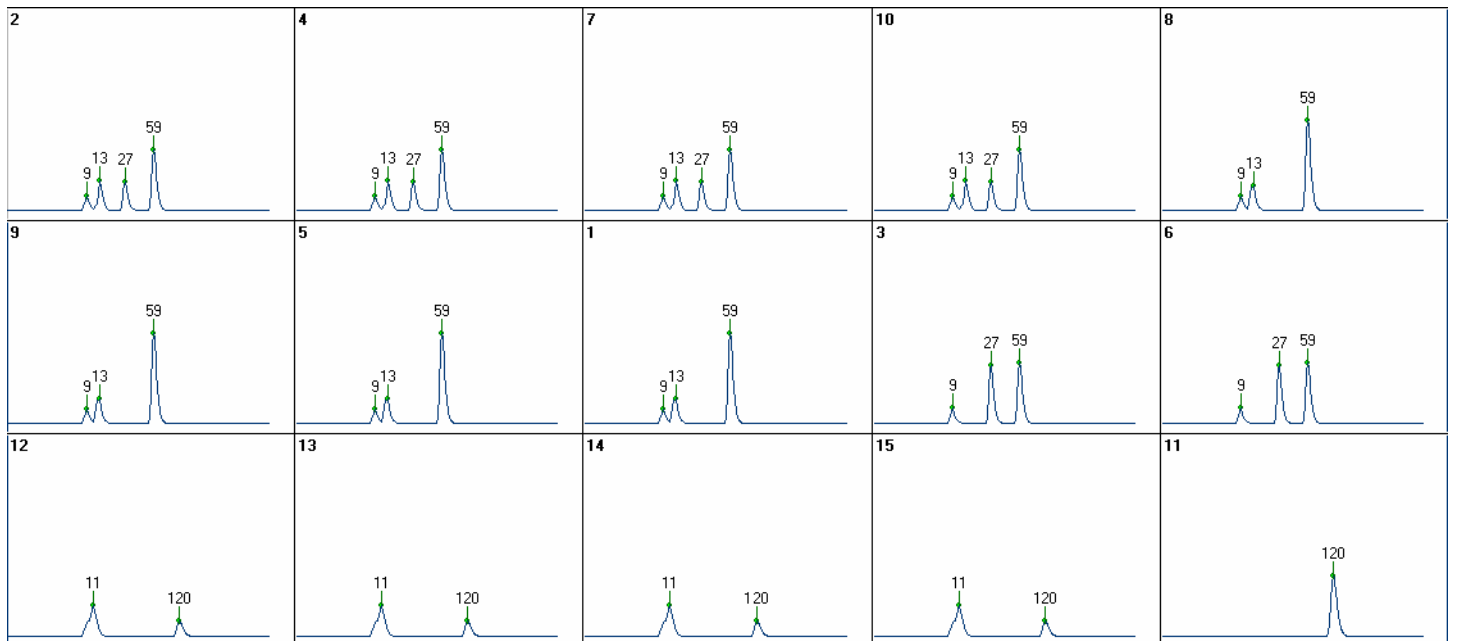


FIG. 4. Individual cluster spectra of the graph shown in Fig. 3A.

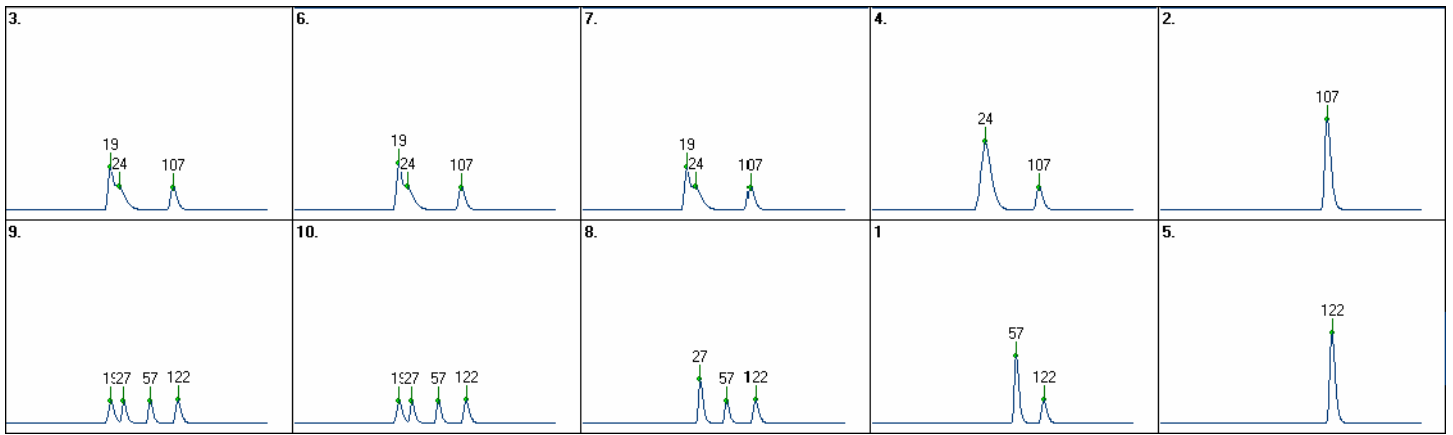


FIG. 5. Individual cluster spectra of the graph shown in Fig. 3B.

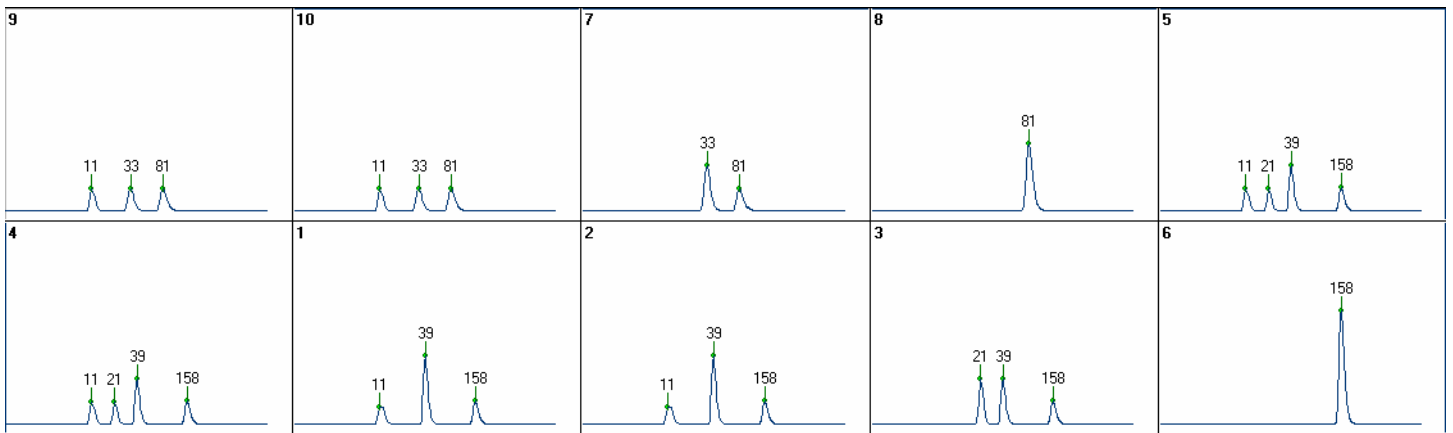


FIG. 6. Individual cluster spectra of the graph shown in Fig. 3C.

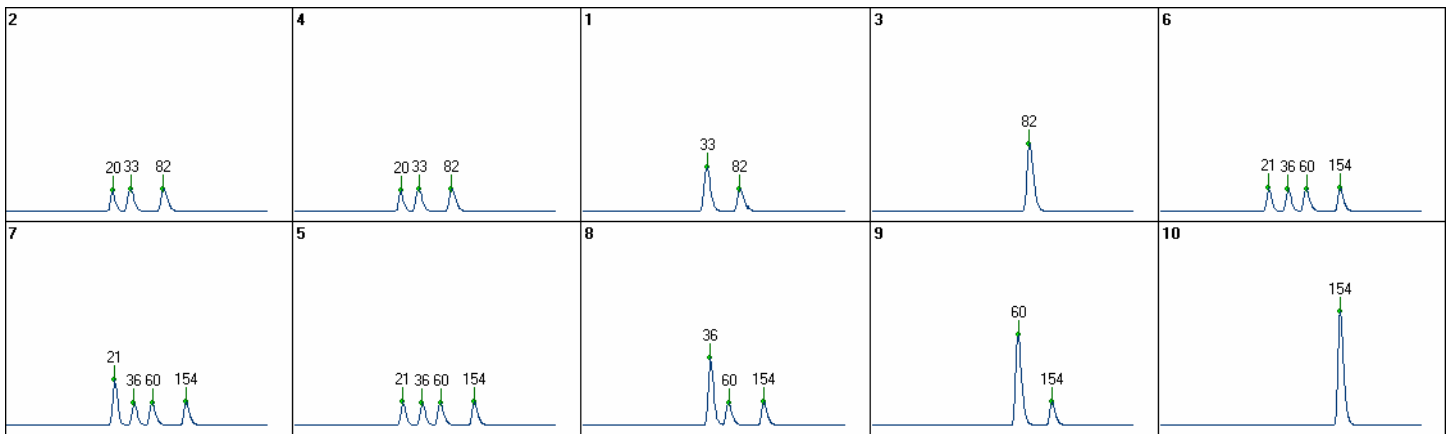


FIG. 7. Individual cluster spectra of the graph shown in Fig. 3D.

Cluster spectra provide a powerful visualization tool that allows a data analyst to abstract their mind from numbers that are often impossible to be objectively interpreted. Cluster spectroscopy can be valuable in analysis of any system to which the law of “the total larger than a sum of its parts” applies. To illustrate the above, we will refer to a case study based on public opinion poll analysis.

Table 3 shows selected results from The Los Angeles Times National Poll (the selected data from Study # 443, July 31, 2000, available at: <http://www.latimes.com>, are reproduced by courteous permission of Ms. Susan Pinkus, Director of The Los Angeles Times Poll). A similarity matrix for these data was computed by using the X-metric (see Methods).

Table 3. Responses to question “What is your personal view on gun control laws? Generally speaking, do you think they ought to be more strict than they are now, or less strict, or do you think the laws we have now are about right?”, from The Los Angeles Times National Poll, Study # 443, July 31, 2000 (“don’t know” answers not included).

	More strict	Less strict	About right
Republicans	31	11	54
Conservatives	36	15	45
Men	41	12	44
Independents	54	7	34
Moderates	59	3	35
Women	64	4	28
Democrats	69	4	24
Liberals	69	4	23

From these data, it is obvious, for instance, that the position of republicans is different from that of democrats and liberals; however, even for such a small dataset, it is not easy to uncover all of the relationships among the eight groups of respondents. Fig. 8

shows the individual cluster spectra of each of the eight groups of respondents, and Table 4 provides the $e_i(T)$ values for the peak maximums, showing the order in which the subclusters are formed.

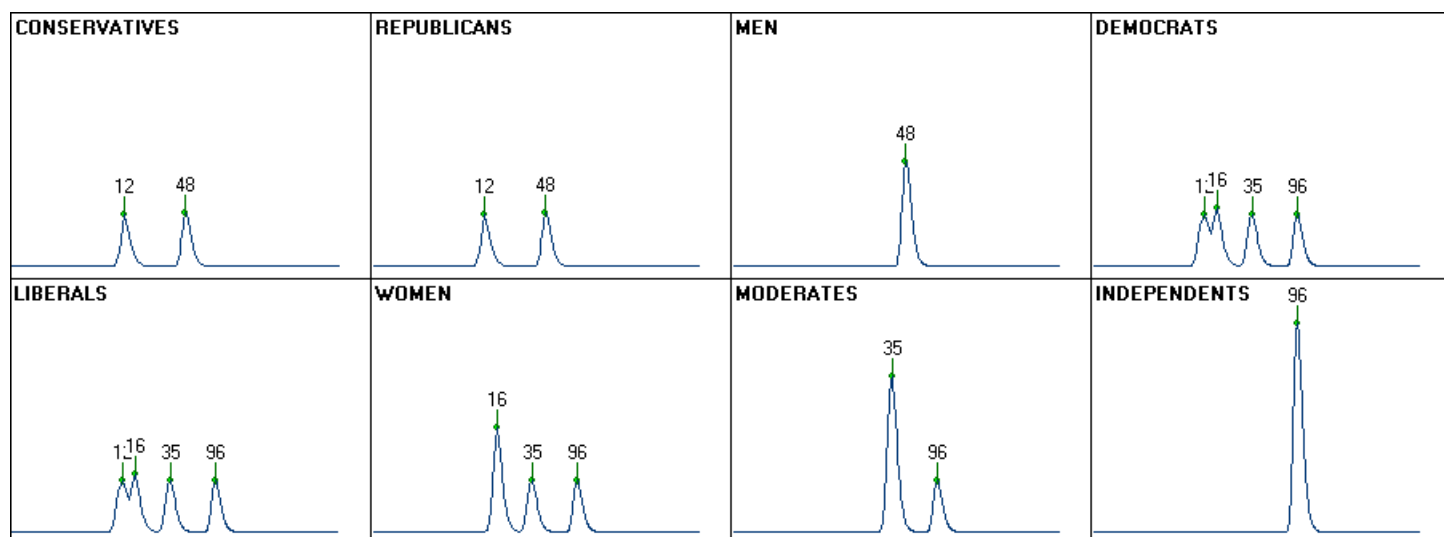


Fig. 8. Cluster spectra of the groups of respondents in the poll according to data in Table 3.

Table 4. Numbers of transformations and e_i values at peak maximums of cluster spectra of the groups of respondents in the poll according to data in Table 3.

	12	16	35	48	96	Total
Republicans	0.130	--	--	0.130	--	0.260
Conservatives	0.132	--	--	0.133	--	0.265
Men	--	--	--	0.265	--	0.265
Liberals	0.132	0.148	0.132	--	0.133	0.545
Democrats	0.132	0.147	0.132	--	0.133	0.544
Women	--	0.260	0.132	--	0.133	0.525
Moderates	--	--	0.400	--	0.133	0.533
Independents	--	--	--	--	0.530	0.530

As is seen, the first objects to join into the α -group are ‘republicans’ and ‘conservatives’, and the first objects that form the β -group are ‘democrats’ and ‘liberals’. Formation of the α -group is complete when it adjoins ‘men’. Formation of the β -group involves the adjunction of ‘women’, then ‘moderates’, and finally ‘independents’, after which the cluster formation is complete. One can also see that in the β -group, on object that is most close to the α -group objects is ‘independents’. The data in Table 4 visualize the

dynamics of subcluster convergence occurring in the process of analysis of a real-life dataset. Here, one can see the same regularities that are shown in Table 2 based on the artificially constructed graph.

Methods

Evolutionary transformation of similarity matrices

The ETSM-method [1] consists in the processing, in one and the same fashion, of each cell of a similarity matrix so that a similarity coefficient between each pair of objects in a data set is replaced by a ratio of a similarity coefficient between each of objects in a pair and a mean value of similarities between each of two objects whose replacement similarity coefficient is under computation and all other objects of a matrix. The algorithm of the process of evolutionary transformation of a similarity matrix is based on the following formula:

$$[S_{ij}]_{T+1} = \text{Aver} ((\text{Min}([S_{in}]_T, [S_{jn}]_T) / \text{Max} ([S_{in}]_T, [S_{jn}]_T), n) \quad (3),$$

where Aver is a geometric (or arithmetic) mean value function; T is a number of similarity matrix transformations; $[S_{ij}]_{T+1}$ is a pair-wise similarity between i and j objects after T+1 transformations; and n is a size of a square similarity matrix. The algorithm for such transformation is repetitively applied to a similarity matrix till each of similarities between objects within each of the clusters reaches 1 and no longer changes. In the end, the process of successive transformations results in convergent evolution of a similarity matrix. First, the least different objects self-join into sub-clusters; then, major sub-clusters merge as necessary, and, finally, all objects appear to be distributed among the two main sub-clusters, which automatically ends the process. Similarities between objects within each of the main sub-clusters equal 1, and similarities between objects of different sub-clusters equal a constant value which is less than 1.

Similarity matrix computation

Similarity matrices were constructed by method [2] that involves computation of monomer matrices according to each parameter of the objects under analysis, and the following hybridization of the obtained monomer matrices into a full matrix according to the equation:

$$S_{ij} = \text{Aver}(M(k)_{ij}, v) \quad (4),$$

where $M(k)_{ij}$ is a pair-wise similarity between objects i and j in a monomer matrix computed according to parameter k; S_{ij} is a pair-wise similarity between objects i and j in a hybrid matrix computed based on a totality of a v-number of parameters; and Aver is a geometric mean function. This SM computation technique allows the processing of any parameters, both numeric (in positive or negative values) and binary, as well as the computation of $S_{i,j}$ by assigning any weights to parameter “k”.

Metrics

Monomer similarity matrices were computed by using the metrics specially designed for the ETSM-method: a metric for Shape (XR) and a metric for Power (R) [2].

According to the XR-metric,

$$S_{(k)ij} = B^{|k_i - k_j|} \quad (5),$$

where $S(k)_{ij}$ is similarity between objects i and j compared by parameter k; k_i and k_j are values of parameter k for objects i and j; and B constant is >1 . The value of B practically has no influence on the result of processing and is usually 1.50. In this work, $B=1000$.

According to the R-metric,

$$S(k)_{ij} = \min(k_i, k_j) / \max(k_i, k_j) \quad (6),$$

where k_i and k_j are values of parameter k for objects i and j.

SM contrasting

The SM contrasting technique [1] consists in proportional decreasing, by several orders of magnitude, of each similarity coefficient (S) in a similarity matrix, which does not change an outcome of the evolutionary transformation of the SM but provides the visualization of processes occurring in a matrix when, in the course of its evolutionary transformation, its pair-wise similarities are reaching values close to 1. The contrasting formula is:

$$[S]_C = \frac{\exp[(\exp S - 1)^{0.082 \cdot x^C}] - 1}{\exp(e - 1)^{0.082 \cdot x^C} - 1} \quad (7),$$

where S and $[S]_C$ are, respectively, the original and contrasted similarities, and C is the contrasting index (in this study, a contrasting index of 210 was applied).

REFERENCES

- [1] Leonid Andreev. Unsupervised automated hierarchical data clustering based on simulation of similarity matrices. U.S. Patent No. 6,640,227.
- [2] Leonid Andreev. High-dimensional data clustering with the use of hybrid similarity matrices. U.S. Patent Application Ser. No. 10/622,542.

For questions or comments on cluster spectroscopy, contact us at info@matrixreasoning.com.