A HIGH-DENSITY CLUSTERING APPROACH TO EXPLORING THE STRUCTURE OF SOCIAL NETWORKS

James M. Lattin and M. Anthony Wong, Sloan School of Management, M.I.T.

This research was supported in part by the Naval Electronics Systems Command contract number N00039-78-G-0160. We are indebted to Joel Levine of Dartmouth College for his help in obtaining the interlock data from the BARON archive, and for his insightful guidance.

In this paper, we present a technique originally developed for decomposing complex software design problems, and suggest its usefulness for exploring the structure of large, non-directed social networks. The technique, based on a high-density clustering model defined on a graph, is quite efficient on very large, relatively sparse networks and provides a convenient, two-dimensional representation of the global network structure. In particular, we demonstrate the usefulness of the high-density clustering technique on the network defined by the interlocking directors of the 200 largest industrial corporations of the 1970 Fortune 800. The technique enables us to examine the regions of "high-density interlocking" in this corporate subnetwork: regions where any group of firms is quite heavily interlock, and where any one firm in the group is not highly linked outside the group. These and other preliminary results indicate that the high-density clustering technique is conceptually appealing and requires much less time and computational expense than other exploratory methods employed.

I. Introduction

Over a decade ago, Levine (1972) set out to examine a relational network defined by the directorate interlocks among 84 corporations named in a report on the trust activities of commercial banks. Levine's goal was to discover the important characteristics of the network using only the information implicit in the interlocks, "...to 'understand' a large network in a crude, almost a-theoretical sense of being able to represent it, to discern its major outlines, and to distinguish important links from those which are not" (p. 14). His first frustrating attempt with pencil and paper led him to a non-metric multidimensional unfolding approach with hopes of being able to picture the network in some Euclidean space of manageable dimensionality, perhaps the first step in an exploratory data analysis of a social network configuration. The result was an actual picture of this corporate subnetwork, a "spherical" representation that enabled Levine to see the global properties of the network and to guide his subsequent investigations and analysis.

Recent trends and the overall growth and development of social network analysis have presented researchers with the opportunity to investigate much larger networks. Mariolis and Schwarz's archive of almost 800 U.S. corporations has been available to researchers for many years (see e.g., Mariolis (1975), Levine and Roy (1975), and Pennings (1980)) and recently an archive of over 400 multinational corporations (known as WORLDNET) was compiled by Levine. Because of the troublesome implications of imposing an arbitrary boundary (see e.g. Barnes (1975), p. 409ff), network analysts may be hesitant to work with subsets of data that are not delimited by a clear social boundary in reality. Ultimately, the size of the network under study is limited by the finite resources of the researcher for gathering and processing data, and by the capacity and manageability of the techniques available for exploring the data.

When the network consists of several hundred nodes and perhaps almost one thousand arcs, many times the size of Levine's original subset, non-metric multidimensional scaling may have any one of a number of drawbacks:

1. In some cases, the information implicit in the network relations is inconsistent with the formation of a configuration in small-dimensional Euclidean space.

2. When the dimensionality of the solution space is not small (i.e., greater than three or four), the results are difficult to envision and to interpret.

3. For very large problems, non-metric multidimensional scaling can be relatively time-consuming and computationally expensive.

In this paper, we present a high-density clustering technique developed as a part of a systematic design methodology (SDM) for the design of large software systems (see Huff (1979), Wong (1980), Lattin (1981a), and Wong, Madnick and Lattin (1982)). This new approach offers several advantages for the analysis of large social networks:

1. The model considers the context within which any two nodes are linked, thereby taking into account not only the direct relation between two nodes but also the interaction of each with surrounding nodes.

2. The results are presented in a hierarchical clustering trace, which requires no unconventional display and facilitates quick scanning and interpretation.

3. The technique is quite computationally efficient, especially so for sparse networks.

In section II, we elaborate on the high-density clustering technique. In section III, we demonstrate the usefulness of the technique by examining the structure of a single connected network of 171 of the largest 200 industrials listed in the 1970 Fortune 800. Finally, in section IV, we conclude with some observations about the technique and some interesting directions for further application.

II. The High-Density Clustering Technique

The clustering technique determines regions of "high-density" in a network, i.e., groups of highly or heavily linked nodes separated by other such groups by relatively few, weak links. Systems designers have used the high-density clustering model to focus on the global features of their design specifications, by modeling the design problem as a graph, with the functional requirements of the problem as nodes and the interdependencies between requirements as arcs. The high-density regions of such a design graph suggest well-defined subtasks that exhibit good design characteristics. Just as systems designers use the high-density clustering model to focus on a complicated set of design specifications, so can we use the model to explore the structure of a social network, looking for well-defined regions of highly interrelated nodes that appear to stand apart from the rest.

We choose the high-density clustering model on a graph (Wong (1980); see Lattin (1981a) for implementation and performance evaluation) because of its apparent advantages over other network/graph decomposition techniques:

1. The technique does not require *a priori* specification of the number of subgraphs; rather, it identifies regions of high-density and thereby suggests to the investigator the appropriate number of components to the graph (see e.g., Kernighan and Lin (1970), Christofides and Brooker (1976), both of which require *a priori* specification of subgraph size or the number of subgraphs).

2. The technique utilizes a maximum spanning tree algorithm, which operates very rapidly on large, relatively sparse graphs (see e.g., references in point 1 above; see also McCormick (1972), Huff (1979) for heuristics that take no account of the sparisty of the graph).

3. The high-density clustering model does not rely on a goodness-of-partition measure, which is difficult to specify without somehow favoring extreme partitions (see Wong (1980) for a critique of these methods).

The concept of a density on an arc between two nodes is defined using a neighborhood concept that corresponds to the nearest neighbor density estimation technique in statistics. For an unweighted graph, the density on an arc between any two connected nodes i and j is defined as follows:

$$d_{ij} = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} = \frac{\text{\# of nodes connected to both i and j (plus i and j)}}{\text{\# of nodes connected to either i or j (plus i and j)}}$$

where d_{ij} is the density on the arc between node i and j, N_i is the neighborhood of node i; i.e., node i and all nodes k such that arc (i,k) is in the graph, and $|\cdot|$ denotes the cardinality of the included set.

Wong (1980) proposes the following density measure for weighted graphs:

$$d_{ij} = \frac{2w_{ij} + \sum_{k \in \mathcal{C}} (w_{ik} + w_{kj})/2}{|N_i \cup N_j|} \qquad \text{for all } (i,j) \in A, \qquad (1)$$

where w_{ij} = the weight on the arc between node i and node j, A is the set of all arcs in the graph, and $\mathcal{C}$ is the set of all nodes k (distinct from i and j) such that (i,k) and (k,j) are elements of A. Using this measure, a graph with arc weights in the interval [0, 1.0] will have arc densities in the same interval.

Wong *et. al.* (1982) describe the computational algorithm for finding the tree of high-density clusters for a given graph G = (N,A), where N is the set of nodes in G and A is the set of (weighted) arcs.

STEP 0: For each arc $(i,j) \in A$, compute the density d_{ij} according to equation (1) above.

STEP 1: Find the largest d_{ij} and amalgamate nodes i and j to form a cluster C. Define the density between C and any other node k by $d_{kC} = \max \{d_{ik}, d_{jk}\}$ for all k such that either $(i,k) \in A$ or $(j,k) \in A$.

STEP 2: Repeat STEP 1, treating C as a node that replaces nodes i and j. Continue this step until all nodes are grouped into one large cluster. All clusters C formed in the course of this algorithm are high-density clusters.

Although the actual implementation of the high-density clustering technique differs slightly from the algorithm outlined above, neither differs essentially from the classical minimal (in this case, maximal) spanning tree algorithm. For a published computational algorithm that would produce the MST, see Hartigan (1975).

In order to illustrate the concept of an arc density, and the output of the high-density clustering technique, we select a small subset of seven interlocked corporations from the data compiled by Mariolis and Schwarz, which are currently available on-line in the BARON archive at Dartmouth College. Figure 1 shows the subset configured as a weighted network according to the following weighting scheme:

$$w_{ij} = \frac{|C_i \cap C_j|}{\min\{|C_i|, \ |C_j|\}} , \qquad (2)$$

where w_{ij} is the weight on the arc connecting i and j, C_i is the set of directors on the board of corporation i, and $|C_i|$ denotes the cardinality of set C_i. The measure is quite similar to the one proposed by Bearden et. al. (1974), who used $(|C_i| \cdot |C_j|)^{1/2}$ in the denominator instead of $\min\{|C_i|, \ |C_j|\}$. In practice, the difference between the two schemes does not appear to be substantial. The densities for each arc, calculated according to (1), are shown in the figure in parentheses.

Any value d* ε [0, 1.0] defines a density contour, delimiting the high-density clusters at level d*. These are the regions of "high-density interlocking" in the corporate network, where firms are linked to other firms within the group with a density at least d*. These density contours have a hierarchical nature; each lower contour encircles all the nodes in the contours above it. Figure 1 shows an example of the hierarchy of high-density clusters.

Had we selected a much larger subset of corporations for the illustration in Figure 1, the depiction of density contours might have become quite difficult. For large graphs we appeal to a more concise representation, which is the modified form shown in figure 2 of the standard clustering tree output (see e.g., Hartigan (1975) for standard output). Although this representation contains less information than that of Figure 1, it is somewhat easier to see that General Telephone (19), Continental Can (51), and Textron (57) are clustered at a much higher density level than any of the other firms in the sample. A vertical line at density level d* = 0.0675 delimits the high-density clusters at that level ({19, 51, 57}, {138, 192}) from the low-density nodes ({191}). The fact that the point between Warner Lambert (138) and Otis Elevator (192) does not appear as a sharp peak to the right of the line d* = .0675 in the figure indicates that the pair may not stand by itself as a strong, well-defined cluster.

We cannot, however, say anything further about the interlocking behavior of these seven firms in a more global setting. A concern voiced by Barnes (1975) and shared by all network analysts is that "this structure may well be merely an artifact of our procedure for delimiting the graph" (p. 409); that is, the rather aribitrary choice of these particular seven firms from the set of tens of thousands in the U.S. may distort the picture of their actual interlocking structure. In the context of the entire Fortune 500 industrials, for example, any one of the seven might be more heavily interlocked outside the sample than inside it. Adding another several hundred firms to this network might well change its characteristics completely.

Because we use the full extent of the interlocking activity of a firm in our density measure, we must attempt to present the network in the most complete form possible. Otherwise, we run the risk of distorting certain areas in the network by failing to include firms that might contribute either to the weight assessed to the link between two firms (through a direct or an indirect interlock) or to the density measure on some arc (i,j) (e.g., by increasing the number of firms in the neighborhood of either i or j). In theory, this requires the inclusion of all existing firms into the network; in practice, we compromise by establishing some boundary around a large number of firms, and hope that in this collection we capture the main effects of the interlocking phenomenon.

III. Applying the Technique

In order to demonstrate the usefulness of the new technique in exploring the structure of large social networks (but recognizing that this is a first cut at the data with a new approach), we focus on the largest 200 industrials in the 1970 Fortune 800 data. Of these firms, 171 form a single connected network with 437 weighted arcs, while the remaining 29 firms are isolates. Further details summarizing the characteristics of this particular subnetwork are in Lattin (1981b).

The results of the high-density clustering analysis are shown in Figure 3. The formation of the clustering trace shown in the figure is quite rapid, accomplished by an algorithm which runs in O(A) operations, where A is the number of arcs in the network. The technique, as currently implemented on an IBM 370/168, requires less than one second of CPU time to form the clustering trace. The calculation of the arc densities runs in $O(A^2/N)$ operations, or k^2N^3 where $k = 2A/N^2$ is a fraction representing the sparsity of the graph. Although this step is potentially very time consuming for large, complete networks, in this example the value of k = .03 and the technique requires less than two CPU seconds to calculate the arc densities. For additional detail on the actual implementation and efficiency of the algorithms in the high-density clustering technique, see Lattin (1981a).

The jagged peaks to the right of the vertical line indicating density level d* = .030 are the high-density clusters at this particular level. Space limitations make it impossible to label every point in the clustering trace with the associated firm, so only the fifty-five firms in the six largest clusters (about one-third of the network) have been identified. The twenty remaining unlabeled clusters include a total of fifty-one other firms, leaving sixty-five firms unclustered at level d* = .030.

The effect of increasing or decreasing the density level d* is much the same as raising or lowering the water level on a three-dimensional topological map. When the water level is relatively high (corresponding, say, to a density level of d* = .075), only the tops of the very tallest peaks on the map will show above the water level. These are the high-density clusters; the remaining land underwater represents the unclustered nodes. As the water level decreases, more and more of the mountain tops begin to show, and as the water drops below the level of the highest valleys, some of the mountains begin to join together. Finally, when the water level is quite low (e.g., d* = .001), all of the land mass is above water, leaving one large cluster containing all the nodes in the network.

Using a fully annotated clustering trace (which requires more space than we have available to reproduce here), we can examine the regions of "high-density interlocking" in this corporate subnetwork and proceed with further analyses armed with a picture of the network's clustering structure. The high-density clustering technique makes it possible for the researcher to search out that "arresting finding" which is the stuff of an exploratory data analysis, a significant insight that may help direct his or her subsequent investigations.

In this particular case, the overall shape of the clustering trace led to an interesting approach to the network that might not have been possible without the representation afforded by the high-density clustering technique. Lattin (1981b) noticed an underlying uniformity to the clustering trace of the corporate subnetwork and posed the following question: what if such a structure might just as well have arisen from some completely random process of director choice? He went on to model a plausible process of corporations choosing directors without regard to their established affiliations, and then compared the clustering trace from the model network thus generated to the clustering trace of the actual network. The results are astonishingly similar, although there exists no test to differentiate the two with any degree of statistical significance. It may be valuable to test other models of social interaction in this way.

IV. Conclusion

The high-density clustering technique has had but a few applications in the relatively new area of systematic design of software systems, and with the exception of this one example, it is virtually untested as a technique for exploring social network data. Nonetheless, it has provided us with some useful insight in our study of the corporate interlock data, and may have some advantages as well for other researchers studying large social networks.

In our own work, two promising directions for further analysis have arisen. One approach is to probe further to uncover some structural difference between the model network described above and the actual interlock network. This might be done by examining the networks of direct and indirect interlocks to identify any second order structural differences. Another research direction involves examining the structural changes in the corporate network as it grows in size. There may be certain size-dependent characterisitics of the interlocking phenomenon that will show up by comparing the clustering traces of larger and larger networks. In either case, the high-density technique should prove to be a helpful tool.

Notes

For any further information regarding the high-density clustering programs or the technical reports listed in the references, please contact: James Lattin, E53-336 Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139.

Bibliography

Barnes, J. A. 1975. Network analysis: orienting notion, rigorous technique or substantive field of study? in Holland, P. W. and Samuel Leinhardt, eds. Perspectives on Social Network Research. New York: Academic Press.

Bearden, J. et. al. 1974. The nature and extent of bank centrality in corporate networks, presented at the Annual Meeting of the American Sociological Association in San Francisco.

Christofides, N. and P. Brooker. 1976. The optimal partitioning of graphs, SIAM Journal of Applied Math. 30: 55-69.

Hartigan, J. A. 1975. Clustering Algorithms. New York: John Wiley.

Huff, S. L. 1979. A systematic methodology for designing the architecture of complex software systems, unpublished Ph.D. dissertation, Sloan School of Management, M.I.T.

Kernighan, B. W. and S. Lin. 1970. An efficient heuristic for partitioning graphs, Bell System Technical Journal. 49: 291-307.

Lattin, J. M. 1981a. Implementation and evaluation of a graph partitioning technique based on a high-density clustering model, Technical Report #15, Sloan School of Management, M.I.T.

1981b. Examining the interlocking structure of the corporate network, Technical Report #16, Sloan School of Management, M.I.T.

Levine, J. H. 1972. The sphere of influence, American Sociological Review. 37: 14-27.

Levine, J. H. and W. S. Roy. 1975. A study of interlocking directorates: vital concepts of organization, in Holland, P. W. and Samuel Leinhardt, eds. Perspectives on Social Network Research. New York: Academic Press.

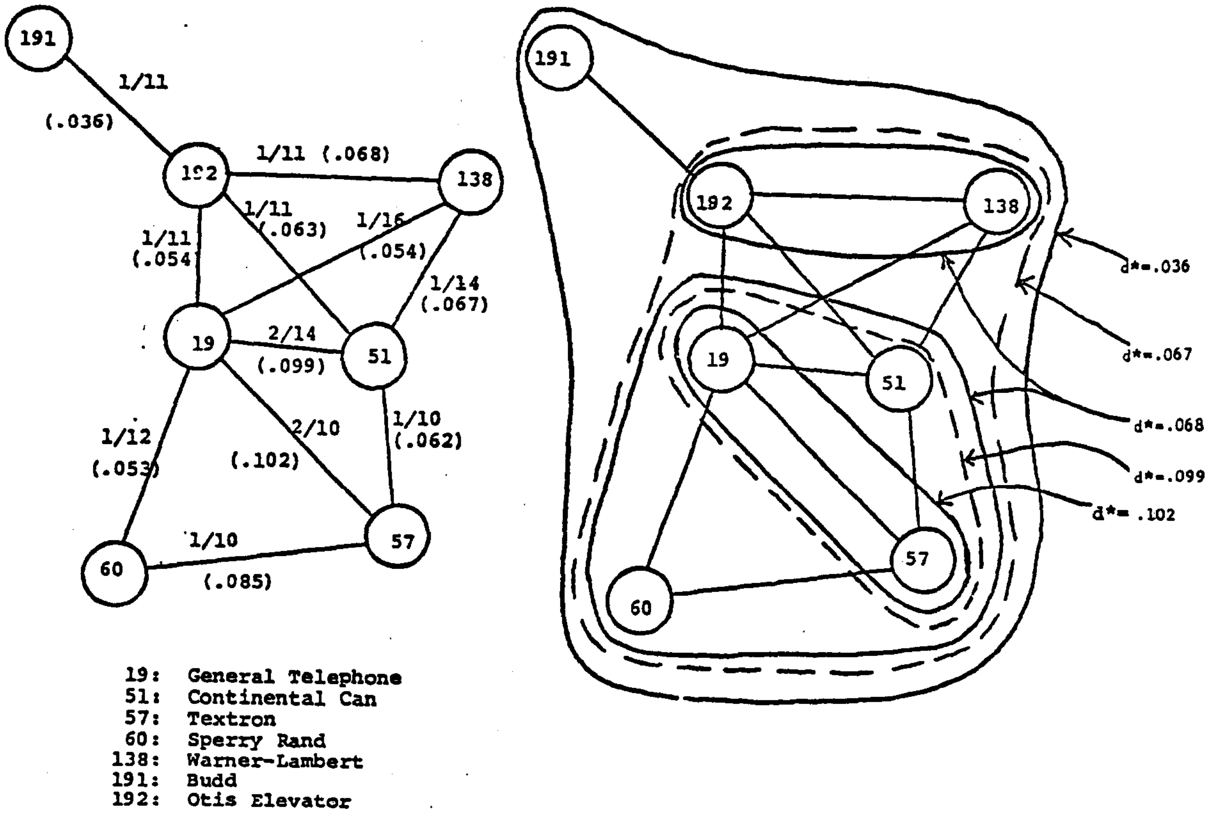
Mariolis, P. 1975. Interlocking dirèctorates and control of corporations, Social Science Quarterly. 56(3): 425-439.

McCormick, W. T., P. J. Schweitzer, and T. W. White. 1972. Problem decomposition and data reorganization by a clustering technique, Operations Research. 20(5): 993-1007.

Pennings, J. M. 1980. Interlocking Directorates. San Francisco, CA: Jossey-Bass.

Wong, M. A. 1981. A graph decomposition technique based on a high-density clustering model on graphs, Technical Report #14. Sloan School of Management, M.I.T.

Wong, M. A., S. E. Madnick, and J. M. Lattin. 1982. A graph-partitioning method for decomposing complex software design problems into manageable subproblems, unpublished manuscript, Sloan School of Management, M.I.T.



19:	General Telephone
51:	Continental Can
57:	Textron
60:	Sperry Rand
138:	Warner-Lambert
191:	Budd
192:	Otis Elevator

Figure 1

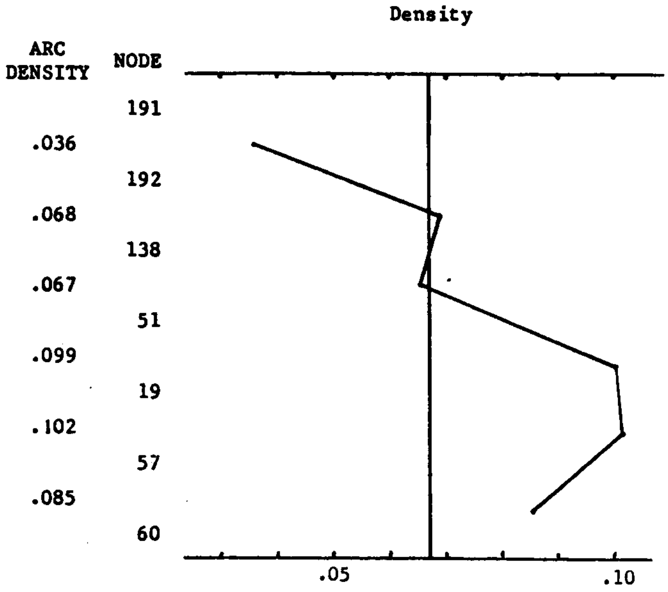
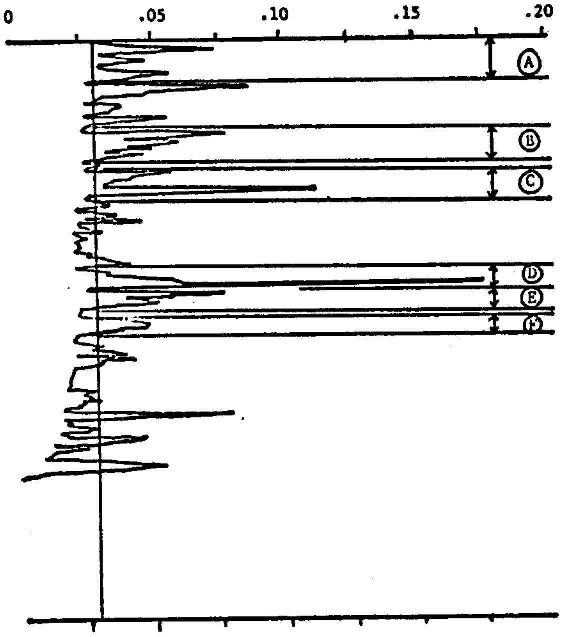


Figure 2

Clustering trace for example network of seven industrials.
Vertical line delimits high-density clusters at level d*=.0675



Ⓐ - General Motors, Alcoa, Heinz, Gulf, PPG, LTV, General Foods, Whirlpool, Std. Oil NJ, Mobil, Del Monte, Caterpillar, Northwest Ind., Time, Borg-Warner

Ⓑ - Olin, Squibb Beech-nut, Avco, Republic Steel, Std. Oil Ohio, White Motor, Firestone, Rockwell, TRW, Kodak, Std. Brands

Ⓒ - ARCO, Bristol Meyers, Johns-Manville, American Standard, NCR, American Can, Allis Chalmers, Xerox, Teledyne

Ⓓ - Celanese, Anaconda, Texaco, National Steel, Burroughs, Eaton, Studebaker, Babcock and Wilcox

Ⓔ - Sperry Rand, Textron, General Telephone, Continental Can, Warner-Lambert, Otis Elevator

Ⓕ - Colgate Palmolive, McDonnel Douglas, Bethlehem Steel, CPC International, Western Electric, Uniroyal

Firms listed in order of appearance in the clustering trace, from top to bottom.

Clustering trace for 171 industrials. High-density clusters defined at level d*=.030. Circled letters identify principal clusters detailed at right.

Figure 3