

Multivariate Analysis of Unlabeled and Partially Labeled Structures

Authors:

Carter Butts

Department of Social and Decision Sciences
Center for the Computational Analysis of Social and Organizational
Systems
Carnegie Mellon University

Kathleen Carley

Department of Social and Decision Sciences
Center for the Computational Analysis of Social and Organizational
Systems
H.J. Heinz III School of Public Policy and Management
Carnegie Mellon University

Extended Abstract:

A central problem for the structural analyst is the comparison of disparate social structures in order to identify underlying commonalities between them. This is especially true where large populations of structures are concerned; in such cases, it is often reasonable to ask whether there may be groups of structures which are especially similar to others within the population. Such questions, however, raise fundamental issues regarding the meaning of underlying structure, and of similarity between structures. These issues, ultimately, are central to the problem of describing the distribution of structures within a population.

In a variety of problems of interest, there is no theoretical justification for treating some or all structural elements as a priori distinct from one another. This lack of distinction is equivalent to a notion of exchangeability: certain elements may be exchanged within the structure without doing violence to the theoretical basis for comparison. This situation, common as it is, poses serious difficulties for comparative work. In particular, the methods which exist for directly assessing the differences between structures are based on oriented representations such as labeled graphs; unlabeled or unoriented structures (the sort implied above) cannot be treated in this way.

Previous work by Banks and Carley (1994) and Butts and Carley (1998) has used the Hamming distance (Hamming, 1950) as the fundamental basis for comparison of directed graphs in the labeled case. The reasons for this choice have been detailed elsewhere, and will not be considered at length here; however, it is worth noting that the Hamming distance forms a metric on the set of labeled digraphs, and that the properties of the Hamming distance with respect to various structural measures (such as the central graph)

are reasonably well-understood (see (Banks and Carley, 1994; Butts and Carley, 1998) for more details). Work by Butts and Carley (1998) has further shown that the observed Hamming distance is an unreliable indicator of the difference between unlabeled structures, and have defined a structural distance measure (closely related to the Hamming distance) which is applicable to unlabeled or partially labeled structures. Identifying the structural distance between directed! ! graphs can be accomplished in a number of ways, including heuristic search and canonical labeling approaches; the identified pattern of distances may then be used in identifying the central graph, or in other analyses (Butts and Carley, 1998).

Following this earlier work by Banks and Carley (1994) and Butts and Carley (1998), we here employ the Hamming distance (Hamming, 1950) as our basic measure of difference between structures. In particular, given two labeled digraphs, H_i and H_j with vertex sets $V_i=V_j=V_U$ and edge sets E_i and E_j respectively, we may define a metric distance between them as per Hamming (1950). First, we define an indicator function $\delta_h(x,y)$ such that

$$\delta_h(x,y) = \begin{cases} 1 & \text{if exists } e(v_x,v_y) \text{ in } E_h \\ 0 & \text{otherwise} \end{cases}$$

The function δ_h permits us to count directed edges within a given labeled digraph. To derive the Hamming distance between our two labeled digraphs, then, we simply count the number of directed edges which exist in one graph and not the other. This gives us the following expression for the Hamming distance:

$$D(H_i,H_j) = \sum_{y=1..|V_U|} \sum_{x=1..|V_U|} (|\delta_i(x,y)-\delta_j(x,y)|)$$

As noted above, previous work by Butts and Carley (1998) has shown that the observed Hamming distance between two labeled graphs may be decomposed into a minimal, structural distance which depends only on the underlying unlabeled graphs, and an Additional labeling distance which is a function both of the underlying unlabeled graphs and their respective labelings. For $H_i = L_i(G_i)$ and $H_j = L_j(G_j)$ (where L represents a labeling on the unlabeled graph G), this decomposition gives us

$$D_O(L_i(G_i),L_j(G_j)) = D_S(G_i,G_j) + D_L(L_i(G_i),L_j(G_j))$$

where D_L represents the labeling distance and the structural distance

$$D_S \text{ is given by } D_S(G_i,G_j) = \min(D_O(L_a(G_i),L_b(G_j))) \text{ for all } a,b$$

Minimization of the labeling distance (D_L) between pairs of graphs can be achieved as described above via a canonical labeling algorithm, but other approaches are available for problems of dyadic comparison. As Butts and Carley (1998) suggest, heuristic search techniques such as Monte Carlo sampling and genetic algorithms provide possible alternatives for finding the structural distance between a given pair of graphs, and have the distinct advantage of being tunable to adjust performance based on the problem under study.

While previous research, then, has given us a valuable set of tools for the dyadic comparison of social structures and for one type of point estimate of central tendency within graph sets (the central graph), we do not have, as yet, techniques for addressing the broader question of general tendencies towards similarity within large graph sets. For instance, given some large collection of social structures, we would like to be able to determine the extent to which the larger set may be expressed in terms of a smaller set of highly similar structures. In traditional data analysis, such questions are commonly answered via methods of cluster analysis (Johnson, 1967), which identify groups of observations satisfying various properties of within-group similarity and between-group difference. Another, related, question is that of the identification of archetypal structures in terms of which the structures of a given data set may be expressed; this problem is roughly analogous to those addressed classically by principal component and factor analyses. In general, then, we find that a number of problems arise in social network analysis which have clear classical analogues, but for which classical methods cannot generally be applied due to their differing assumptions (e.g., exchangeability of observations, distributional requirements (especially normality), non-permutability of labels across variables).

Here, we propose a number of techniques whose basis lies in traditional multivariate analysis for the purposes of analyzing both labeled and unlabeled structural data. Unlike typical network analyses, of course, our “data points” will often represent entire social structures (rather than vectors of attribute values), a fact which will introduce some special concerns. Nevertheless, we here demonstrate that multivariate analysis of both labeled and underlying (i.e., unlabeled or unoriented) social structures is both possible and informative, and extend our initial results using both a sample analysis of data collected on work teams in a Carnegie Mellon University information systems program (Carley et al., 1993) and simulation of method behavior across a range of conditions.

Banks, D.L. and Carley, K. (1994) “Metric Inference for Social Networks,” *Journal of Classification*, 11, 121-149.

Butts, C. and Carley, K. (1998) “Canonical Labeling to Facilitate Graph Comparison.” ICES Research Report 88-06-98, Carnegie Mellon University.

Carley, K., Kiesler, S., and Wholey, D. (1993) “Learning Teamwork: Studies of Training in Software Development,” *Proceedings of the 1993 Coordination Theory and Collaboration Technology Workshop*. Symposium conducted for the National Science Foundation, Washington, D.C.

Hamming (1950) “Error Detecting and Error Correcting Codes,” *Bell System Technical Journal*, 29, 147-160.

Johnson, S.C. (1967) “Hierarchical Clustering Schemes,” *Psychometrika*, 32, 241-253.