A SPECTRAL ANALYSIS OF RELATIONS

Klaus Krippendorff
University of Pennsylvania, Philadelphia

August 1976

Introduction

The concept of "relation" is probably most central to the social sciences. For example, "descendency," a simple biological relation between exactly three individuals, does give rise to the most elaborate kinship systems. Kinship terms denote social relations, transcend the individuals involved and are an important part of social reality. "Communication" is another relation linking socially two or more individuals as senders and receivers. The process of communication has been recognized as the glue that holds complex organizations together. And when one speaks of "organizational structure" one always has some network of role relations in mind whether its defining feature is authority, control, prestige, friendship or the like. Sender and receiver, and in fact all social roles, are defineable only relative to each other, in the context of manifest social relations. Relations are also manifest in the messages transacted within society. The aim of linguistics is to reveal structures in language. And for a semantic example, paragraphs of law delineate in considerable detail procedures and consequences as a function of the co-occurance of circumstances, events and people and thereby structure ongoing social processes. Almost anything social is transacted and has relational qualities rather than individual properties that could be observed in isolation.

In view of the theoretical importance of this concept, it is surprising that methods in social research are particularly poor when it comes to analysing relational data, especially when the relations inherent in data are complex and escape the bias of natural language towards polar opposites, pair comparisons and binary relations generally. The purpose of this paper is to focus attention on complex (many-valued) relations, to present a formal technique for analysing such relations in multi-variable data and to provide a calculus for assessing the degree to which these

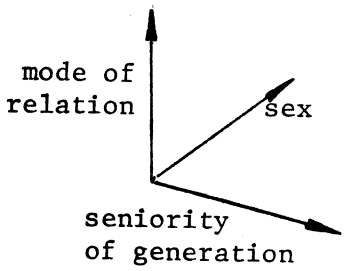
relations account for given data. The proposed technique is formally analogous to spectral analysis although its content differs in fundamental ways from physics. In conclusion the paper suggests that many multivariate techniques in social research, despite their implicit claim, ignore what are perhaps the most important characterisics of multi-variable data: higher-order relations.

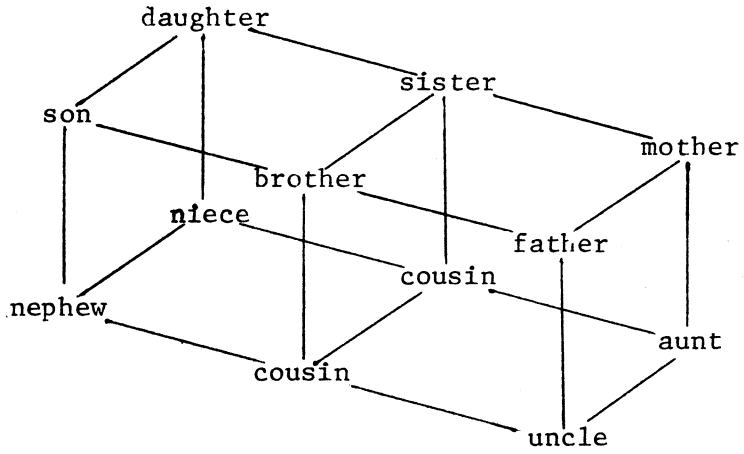
The paper responds to the pressing need for synthesizing and analyzing complex relations in a variety of systems. This need has been established in Simon's work on the "Architecture of Complexity" (1969) which elaborates on essentially four problems of complexity: that of understanding the evolution of organization, that of its adequate description, that of the decomposability of relations, and that of hierarchy. In this paper, decomposability stands in the foreground, hierarchy being a particular form of decomposition and description being the final aim of understanding the process of the evolution of complexity and of its product: complex structures.

Relations

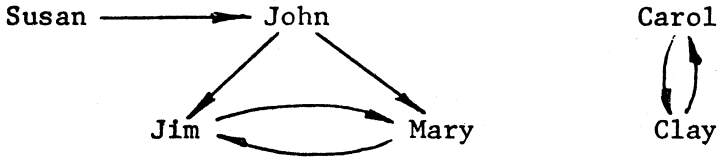
The description and analysis of relational data is not a simple matter, especially when the manifest relations are complex in the sense of involving many values, arguments, dimensions or variables. For the sake of clarity, let me give a few simple examples of relations, show how they all conform to a basic conception and then outline three principal tasks for the analysis of relations.

The first example may be taken from English kinship terminology, a fraction of which is conveniently represented within three semantic dimensions relative to the user of these terms. Its form of representation is common in anthropological linguistics:





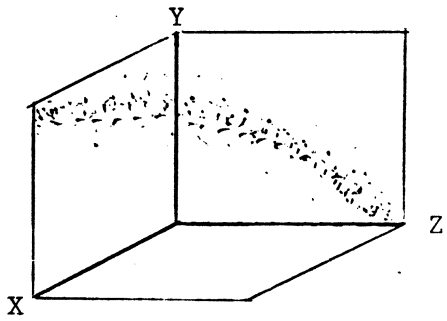
A second example may be taken from sociometric analysis of preferential choices among the members of a social group:



A third example may be taken from causal analysis in which coefficients α_{ij} express x_i's influence on x_j:

$$
\begin{aligned}
x_1 &= \alpha_{21}x_2 + \alpha_{31}x_3 + \alpha_{41}x_4 \\
x_2 &= \alpha_{32}x_3 + \alpha_{42}x_4 \\
x_3 &= \alpha_{43}x_4
\end{aligned}
$$

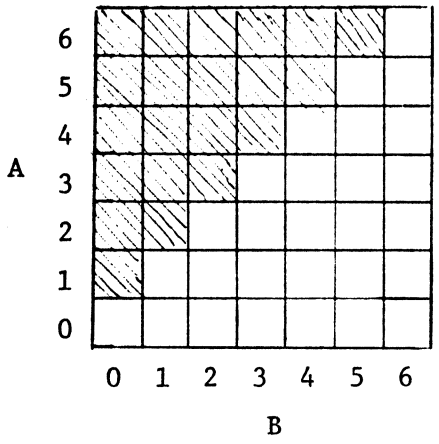
A fourth example may be taken from elementary statistics in which data define point distributions in multi-dimensional spaces and correlation coefficients assess the stregth of an association between pairs of variables:



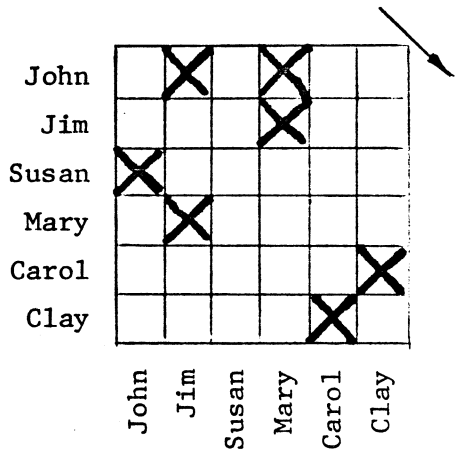
A final example may be taken from mathematics where numerous expressions include relational operators, among the most elementary of which is the "larger than"-sign:

$$a > b$$

Despite their distictively different appearances and historical origins, it can be shown that all of the above examples of relations conform to the general idea, initially advanced by Wiener (1914) who intended to give the then somewhat blurred and mystical notion of relation a more definite mathematical representation. His proposal, now accepted everywhere, was to identify a relation with a proper subset of a product set, to which I must add that this subset not be decomposable into its components. So, in the example of "a>b" the product set is the set of all pairs of numbers, here integers, and the subset is the set of pairs satisfying the definition of >, here shaded in the following diagram:



Similarly can the sociometric preference data be represented in matrix form:



In either example, the list of integers or the list of names by themselves do not indicate how pairs of them are related, hence neither subset is decomposable. The product set for the kinship relations and that for the causal network, each require four-dimensional representations, the former within the three semantic dimensions and the set of names, the latter within the four variables occuring in the system of equations.

In probabilistic terms, relations are manifest when probability distributions within a product set, are uneven and, what is more important, not explainable by the probability distributions in its components. The point distribution

above illustrates the case. This particular three-dimensional distribution of data points cannot be explained in terms of or reproduced from the probabilities in either dimension. Data points have to be taken in triples or at least in pairs, else irrevocable losses are incurred.

Relational data then take the form of m-tuples, vectors with m components or a row of m ordered values:

$$< a, b, c, d, \ldots , x_m >$$

each element pertaining to a different variable or dimension, $a \varepsilon A$, $b \varepsilon B$, $c \varepsilon C$, etc. and each occurring in the sample with a certain frequency or probability. Accordingly, the data on the above kinship example may be listed as a set of quadruples:

Name	Seniority	Mode	Sex
< father	senior	affineal	male >
< mother	senior	affineal	female >
< uncle	senior	consanguineal	male >
< aunt	senior	consanguineal	female >
< brother	same	affineal	male >
< sister	same	affineal	female >
< cousin	same	consanguineal	male >
< cousin	same	consanguineal	female >
< son	junior	affineal	male >
< daughter	junior	affineal	female >
< nephew	junior	consanguineal	male >
< nice	junior	consanguineal	female >

In the analysis of kinship terminology frequency considerations might be unimportant except that the 120 quadruples that are excuded from the list are considered to occur with zero probability.

In the analysis of relational data, essentially three tasks are identifiable. First is the task of defining a relation that would account for the data. This means either finding a name that designates the subset within the variables of the data, specifying a formal test that decides for each datum whether or not the relation holds, or designing a mechanism for generating all those, and only those, elements belonging to the relation. The paradigm for the eleven different kinship terms exemplifies the naming (though somewhat complex) of what a kinship relation means. The equations for the causal network exemplifies a formal test. A definition by name, test or generative principle is adequate to the extent it approximates given data.

Second is the task of identifying relational properties, such as reflexivity, transitivity, symmetry. Knowledge of such properties is important in guiding theory construction. The relation "greater-than" is transitive, for example, while on the hand the sociometric choice data do not exhibit transitivity.

Third is the task of locating relations in data and of establishing their magnitude. This is the pricipal task of the spectral analysis of relations to be developed below.

It might be noted in passing that these tasks are not entirely independent. An adequate definition of a relation logically implies the properties that the second task aims to identify. And, an adequate identification of all relevant relational properties includes those that are required to locate a relation or the component of a relation. It follows that the latter is the least ambitious of the three. As will be seen, it is nevertheless not without problems, but it possibly provides the key to accomplishing the second, and ultimately the first, of the three tasks. This hope is taken as the motivation for the following proposal.

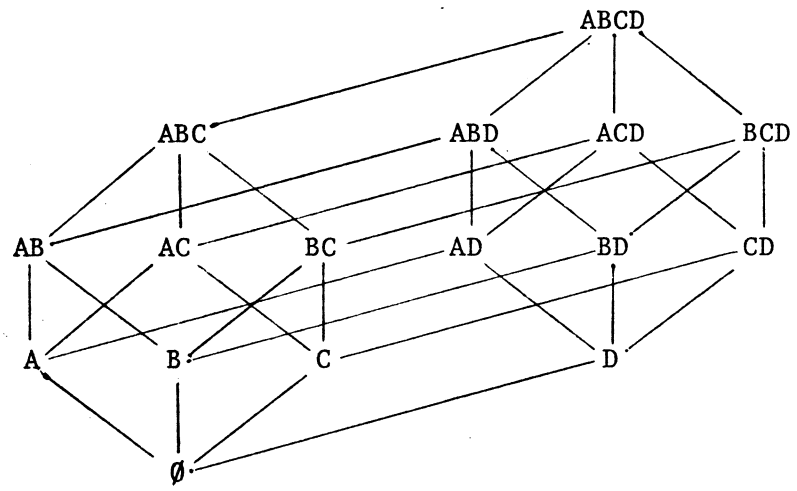
Spectral Analysis

In physics, spectral analysis denotes a technique for analysing a source of energy into several additive components, of which each is identifiable qualitatively and quantitatively. Accordingly, a light source may be seen as decomposable into several basic wavelengths, or pure light sources emitting red, yellow, or blue, for example, which are thought to contribute in individually measurable magnitudes. The synthesis of a comparable light source by means of these basic wavelengths can then serve the purpose of validating the analysis. The same idea is built into a Fourier analysis of oscillations, whether of musical, psychological or social origin. For example, the observed fluctuations of a complex economic indicator might be viewed as a superimposition of several separate sources of variation of which each is assumed to have different origins, known characteristics and measurable effects. Common to these examples is the notion that some relevant property identifiable within given data is accounted for in terms of (i) several qualitatively distinct (logically independent) components, (ii) each component is associated with a magnitude, and (iii) all component magnitudes add up to or at least approximate a measure of the property being assessed. The term spectral analysis is chosen here because the proposed technique shares these features with the above.

Relations, however, are different from the color of a light scource or the modulations in music in that they hold between or among individuals, in that they tie two or more parties to a common course of action, or in that they result from mutually imposed constraints. But, just as a physicist may want to know to what extent his basic components contribute to the source under investigation, so may the social scientist want to know where relations operate within a social group, message, or within any social data of interest, what ordinality they have, whether he can dispense with complex forms of explanation in favor of simple ones, how much he

would lose when restricted to a particular analytical method or form of verbal discourse, et..

The qualities of interest in a spectral analysis of relational data consist of all possible relations of equal or lower ordinality than inherent in given data. The possible relations may be depicted as lattices. For example within four variables A, B. C and D, all nodes of the following lattice may contain relations of one kind or another:



Generally, for relational data within m variables there are:

1	zero-order relation
m	1st-order relations (properties)
$\frac{m(m-1)}{2}$	2nd-order relations (binary relations)
$\frac{m(m-1)(m-2)}{6}$	3rd-order relations (tertiary relations)
⋮	
$\frac{m!}{(m-r)!r!}$	rth-order relations
⋮	
1	mth-order relation

The total number of relations involved is:

$$\sum_{r=0}^{r=m} \frac{m!}{(m-r)!r!} = 2^m$$

This can be a large number, even with a moderate number of variables, and might set computational limits to any analysis of relational data.

The crucial problem now is to define a quantitative measure of the magnitude of a relation's manifestation in data. As indicated above, such measures must be additive to conform to the idea that a complex relation may be the result of a superimposition of several lower order relations. Such measures must be zero when a relation is absent and monotonically increasing with the "strength" of the relation involved. And such measures must be defined on the probability distributions within relational data.

With $a \varepsilon A$, $b \varepsilon B$, etc., probabilities in these subspaces are defined below, e.g.:

p_{abcd}	4th-order probabilities
$p_{abc} = \sum_D p_{abcd}$	3rd-order probabilities
$p_{ab} = \sum_C \sum_D p_{abcd}$	2nd-order probabilities
$p_a = \sum_B \sum_C \sum_D p_{abcd}$	1st-order probabilities
$1 = \sum_A \sum_B \sum_C \sum_D p_{abcd}$	zero-order probability

Two variables, A and B, are regarded as independent when the joint probability distribution is fully explainable from its individual probability distributions, i.e., for all values $a \varepsilon A$ and $b \varepsilon B$: $p_{ab} = p_a p_b$ or $p_{ab}/p_a p_b = 1$. If such a condition prevails this would indicate the absence of a binary relation between A and B. The test is common and requires no further justification.

However, tests for higher-order dependencies are not so straightforeward. A tertiary relation may or may not be decomposable into three binary relations just as it is conceivable that all three binary relations are absent but the tertiary relation is not. Testing for which of these alternatives apply to given data will provide answers to the question of how an analysis of such data might proceed. For 2 by 2 by 2 contingency tables a test for tertiary interactions is provided by Bartlett (1935). It is extendable to higher-order interactions but does not differentiate between decomposable and non-decomposable relations. A test for the absence of a non-decomposable relation should discount what its lower-order relations can account for by themselves. For tertiary relations such a test is provided by the condition that for all values in A, B and C: $p_{abc}/p_{ab}p_{ac}p_{bc}/p_a p_b p_c = 1$. This condition is independent of the conditions $p_{ab}/p_a p_b = 1$, $p_{ac}/p_a p_c = 1$ and $p_{bc}/p_b p_c = 1$. It is a condition for the uniqueness of the combination of three values a, b, and c. With an extension of such conditions to higher-order relations in mind, let me express the rth-order probabilities in terms of 2^r products, each of which denotes a test for the absence of a non-decomposable relation:

$$1 = 1$$

$$p_a = 1\, p_a$$

$$p_{ab} = 1\, p_a\, p_b\, \frac{p_{ab}}{p_a p_b}$$

$$p_{abc} = 1\, p_a\, p_b\, p_c\, \frac{p_{ab}}{p_a p_b}\, \frac{p_{ac}}{p_a p_c}\, \frac{p_{bc}}{p_b p_c}\, \frac{p_{abc}}{p_{ab}p_{ac}p_{bc}/p_a p_b p_c}$$

$$p_{abcd} = 1\, p_a\, p_b\, p_c\, p_d\, \frac{p_{ab}}{p_a p_b}\, \frac{p_{ac}}{p_a p_c}\, \frac{p_{ad}}{p_a p_d}\, \frac{p_{bc}}{p_b p_c}\, \frac{p_{bd}}{p_b p_d}\, \frac{p_{cd}}{p_c p_d}\, \frac{p_{abc}}{p_{ab}p_{ac}p_{bc}/p_a p_b p_c}\, \frac{p_{abd}}{p_{ab}p_{ad}p_{bd}/p_a p_b p_d}\, \frac{p_{acd}}{p_{ac}p_{ad}p_{cd}/p_a p_c p_d}\, \frac{p_{bcd}}{p_{bc}p_{bd}p_{cd}/p_b p_c p_d}\, \frac{p_{abcd}}{p_{abc}p_{abd}p_{acd}p_{bcd}/p_{ab}p_{ac}p_{ad}p_{bc}p_{bd}p_{cd}/p_a p_b p_c p_d}$$

$$p_{\text{rth-order}} = \prod_{i=0}^{r} \prod_{\text{all ith-order in r}} \frac{p_{\text{ith-order}}}{\prod_{\text{all (i-1)th-order in i}} p_{\text{(i-1)th-order}}}\, \frac{1}{\prod_{\text{all (i-2)th-order in i}} p_{\text{(i-2)th-order}}} \cdots \frac{1}{1}$$

$$1 = 1$$

$$p_a = 1\, p_a$$

$$p_{ab} = 1\, p_a\, p_b\, \frac{p_{ab}}{p_a p_b}$$

$$p_{abc} = 1\, p_a\, p_b\, p_c\, \frac{p_{ab}}{p_a p_b}\, \frac{p_{ac}}{p_a p_c}\, \frac{p_{bc}}{p_b p_c}\, \frac{p_{abc}}{\frac{p_{ab} p_{ac} p_{bc}}{p_a p_b p_c}}$$

$$p_{abcd} = 1\, p_a\, p_b\, p_c\, p_d\, \frac{p_{ab}}{p_a p_b}\, \frac{p_{ac}}{p_a p_c}\, \frac{p_{ad}}{p_a p_d}\, \frac{p_{bc}}{p_b p_c}\, \frac{p_{bd}}{p_b p_d}\, \frac{p_{cd}}{p_c p_d}\, \frac{p_{abc}}{\frac{p_{ab} p_{ac} p_{bc}}{p_a p_b p_c}}\, \frac{p_{abd}}{\frac{p_{ab} p_{ad} p_{bd}}{p_a p_b p_d}}\, \frac{p_{acd}}{\frac{p_{ac} p_{ad} p_{cd}}{p_a p_c p_d}}\, \frac{p_{bcd}}{\frac{p_{bc} p_{bd} p_{cd}}{p_b p_c p_d}}\, \frac{\frac{p_{abcd}}{\frac{p_{abc} p_{abd} p_{acd} p_{bcd}}{p_{ab} p_{ac} p_{ad} p_{bc} p_{bd} p_{cd}}}}{p_a p_b p_c p_d}$$

$$p_{r\text{th-order}} = \prod_{i=0}^{r} \left(\text{all } i\text{th-order in } r \right) \frac{p_{i\text{th-order}}}{p_{(i-1)\text{th-order}}\ \text{all } (i-1)\text{th-order in } i \;\; p_{(i-2)\text{th-order}}\ \text{all } (i-2)\text{th-order in } i \;\; \vdots \;\; 1}$$

In the last expression, the "in" under the serial product signs is inclusive.

While much may be learned about the nature of the data by testing these conditions one by one, these products do not render the quantitative account needed in a spectral analysis. The only measuring function that is known to satisfy the conditions set forth above and that does not assume any metric for the variables involved is Shannon's p log p function. When applied to the probabilities one obtains primary entropy measures for each subspace of interest:

$$H(A) = -\sum_{A} p_a \log_2 p_a$$

$$H(A,B) = -\sum_{A}\sum_{B} p_{ab} \log_2 p_{ab}$$

$$H(A,B,C) = -\sum_{A}\sum_{B}\sum_{C} p_{abc} \log_2 p_{abc}$$

$$H(\text{r variables}) = -\sum_{\text{all r variables}} p_{\text{rth-order}} \log_2 p_{\text{rth-order}}$$

And, taking the 1st order probabilities of the expansion to the other side of the equal sign, and applying the logarithmic function as above, one obtains an expression of the total amount of relation (relatedness, constraint, information transmission, association, etc.) in the data to be accounted for:

$$T(A:B) = -H(A,B) + H(A) + H(B)$$

$$T(A:B:C) = -H(A,B,C) + H(A) + H(B) + H(C)$$

$$T(A:B:C:D) = -H(A,B,C,D) + H(A) + H(B) + H(C) + H(D)$$

$$T(\text{r variables:}) = -H(\text{r variables}) + \sum_{i=1}^{r} H(i)$$

As has been argued above, the quantities in terms of which these totals are to

be accounted for must be based on the quantities associated with each possible relation. In effect this means distributing the above totals over the remaining $2^r-(1+r)$ nodes of the lattice of all possible relations in r variables, all of which appear in the expansion of the rth-order probabilities. The logarithmic function applied to these products yields what McGill (1954) calls interaction terms that are expressed here in terms of the primary entropy measures H:

$$
\begin{aligned}
Q(A:B) &= -H(A,B) + H(A) + H(B) \\
Q(A:B:C) &= -H(A,B,C) + H(A,B) + H(A,C) + H(B,C) - H(A) - H(B) - H(C) \\
Q(A:B:C:D) &= -H(A,B,C,D) + H(A,B,C) + H(A,B,D) + H(A,C,D) + H(B,C,D) \\
&\quad - H(A,B) - H(A,C) - H(A,D) - H(B,C) - H(B,D) - H(C,D) \\
&\quad + H(A) + H(B) + H(C) + H(D)
\end{aligned}
$$

$$
Q(\text{rth-order}) = \sum_{k=1}^{r} \Delta_{rk} \sum_{\substack{\text{all kth-} \\ \text{order in r}}} H(\text{k variables}) \quad \text{where} \quad \Delta_{rk} = \begin{cases} -1 \text{ for even } (r-k) \\ +1 \text{ for uneven } (r-k) \end{cases}
$$

Since the logarithm of one is zero, if a product of the expension of the rth-order probabilities is unity for all values a, b, c, ..., then the corresponding measure Q(A:B:C: ...) is zero and indicates the absence of an rth-order relation.

This leads to the fundamental accounting equation for the spectral analysis of relations, according to which the total amount of relation in m-valued relational data is expressed as the algebraic sum of qualitatively distinguished magnitudes Q. The $2^m-(1+m)$ possible relations constitute the spectral qualities and the Q-measures associate with each relation a magnitude that indicates the extent of their presence in data. This fundamental accounting equation is defined as follows:

$$T(A:B) = Q(A:B)$$

$$T(A:B:C) = Q(A:B:C) + Q(A:B) + Q(A:C) + Q(B:C)$$

$$T(A:B:C:D) = Q(A:B:C:D) + Q(A:B:C) + Q(A:B:D) + Q(A:C:D) + Q(B:C:D) + Q(A:B) + Q(A:C) + Q(A:D) + Q(B:C) + Q(B:D) + Q(C:D)$$

$$T(:m \text{ variables}:) = \sum_{r=2}^{m} \sum_{\substack{\text{all } r\text{th-} \\ \text{order in } m}} Q(r\text{th-order})$$

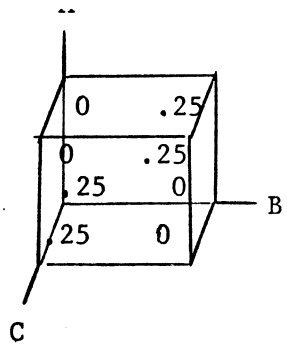
With this fundamental accounting equation it is possible to locate within the lattice of possible relations those that contribute much to the structure inherent in given data and need to be examined to gain insights and those that contribute little or nothing to the total quantities and might therefore be ignored.

The idea of decomposing complex relations into simple ones where possible is based on Ashby's work (1964) who formulated a non-statistical method of analysing relations into subrelations. The quantitive part of this spectral analysis also owes much to the groundwork laid by him in various extensions of information theory (1965)(1969) which were in turn influenced by McGill's work on multivariate information transmission (1954).

Examples, Properties

Some properties of the Q-measures require further elaboration and the application of the accounting equation needs to be demonstrated here on some very simple examples.

First, the Q-measures assess the magnitude of the corresponding relation's account for the data. If it turns out to be zero or approximately zero, then the relation so assessed need not be looked into any further, and if it equals the total, then that relation is the only one that provides a basis for explanations. Thus, relative to the measure T for the total amount of relation in data, the Q-measures locate a relation within the lattice of all possible relations and assess the degree to which it is worth seeking explanations of the data in terms of these relations. In the three variable case, for example, with probabilities p_{abc} entered into a 2 by 2 by 2 contingency matrix:



$$Q(A:B) = 1$$
$$Q(A:C) = 0$$
$$Q(B:C) = 0$$
$$\underline{Q(A:B:C) = 0}$$
$$T(A:B:C) = 1$$

The data are fully accounted for by one binary relation between A and B. The variable C is unrelated to either and might be omitted, for the three-dimensional data cube can be reduced to two dimensions without loss:

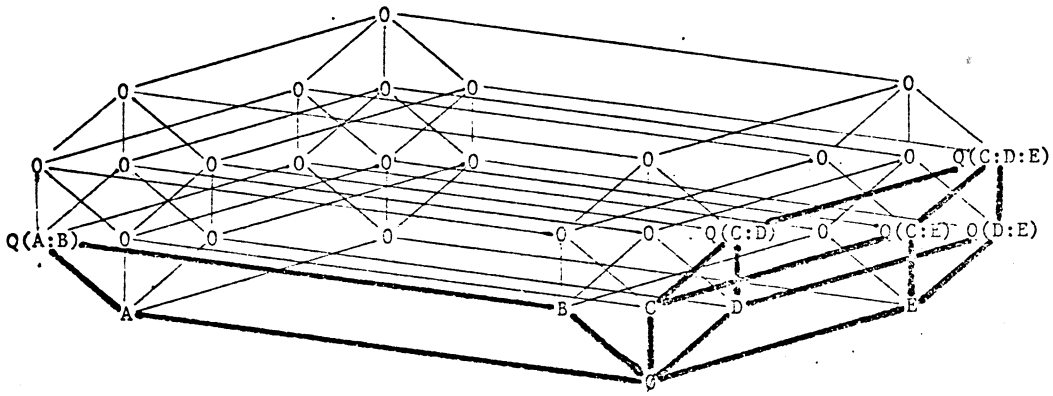
A \ B		
	0	.5
	.5	0

A \ C		
	.25	.25
	.25	.25

B \ C		
	.25	.25
	.25	.25

$$T(A:B:C) = Q(A;B)$$

Second, according to a theorem proven by Ashby (1965), if two sets of variables are statistically independent, then all Q-measures containing variables of both sets of variables are zero. For example, suppose the two variables A and B are independent of the three variables C, D and E in the sense that $p_{abcde} = p_{ab}p_{cde}$ for all a, b, c, d and e, then only five of the 26 Q-measures into which T(A:B:C:D:E) can be analysed may be non-zero. This may be illustrated by the following lattice in which heavy lines are used to indicate which variables and relations are then still worthy of further explorations, while thin lines connect those nodes that drop out.



The account for the total amount of relation then reduces to:

$$T(A:B:C:D:E) = Q(A:B) + Q(C:D) + Q(C:E) + Q(D:E) + Q(C:D:E)$$

or to:

$$T(A:B:C:D:E) = T(A:B) + T(C:D:E)$$

If such conditions are found to hold, tremendous analytical advantage can be taken by partitioning the set of m variables into two with lower ordinality in each. The advantage lies in the fact, suggested by the lattice above, that the numerosity of relations is greatly reduced if variables can be analysed separately. In case a partitioning of variables is justified, the advantage is rooted in the following inequality:

$$2^r + 2^{m-r} << 2^m \quad .$$

Thus, given that:

$$T(A:B:C:D:E) = T(A:B) + T(C:D:E),$$

the spectral analytical account for the total amount of relation reduces from:

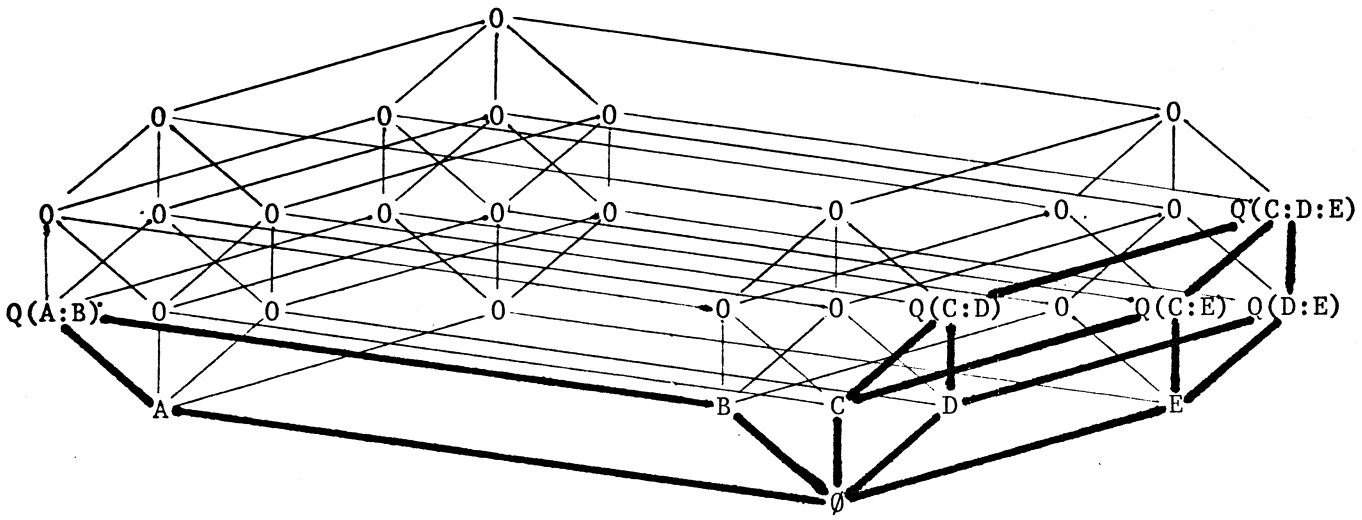
$$\begin{aligned} T(A:B:C:D:E) = {} & Q(A:B) + Q(A:C) + Q(A:D) + Q(A:E) + Q(B:C) \\ & + Q(B:D) + Q(B:E) + Q(C:D) + Q(C:E) + Q(D:E) \\ & + Q(A:B:C) + Q(A:B:D) + Q(A:B:E) + Q(A:C:D) + Q(A:C:E) \\ & + Q(A:D:E) + Q(B:C:D) + Q(B:C:E) + Q(B:D:E) + Q(C:D:E) \\ & + Q(A:B:C:D) + Q(A:B:C:E) + Q(A:B:D:E) + Q(A:C:D:E) + \\ & + Q(A:B:C:D:E) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + Q(B:C:D:E) \end{aligned}$$

to:

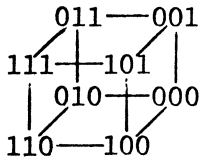
$$T(A:B:C:D:E) = Q(A:B) + Q(C:D) + Q(C:E) + Q(D:E) + Q(C:D:E).$$

If such conditions are found to hold, tremendous analytical advantages can be taken by partitioning the set of m variables into two subsets, involving, say r and m-r variables, with lower ordinality in each. The advantage lies in the fact, suggested by the lattice and by the two accounting equations above, that the numerosity of relations is greatly reduced if variables can be analysed separately. ~~The~~ In case a partitioning of variables is justified, the advantage is rooted in the following inequality:

$$2^r + 2^{m-r} << 2^m \quad .$$


Q(C:D:E)
Q(A:B)
Q(C:D)
Q(C:E)
Q(D:E)
A
B
C
D
E
Ø

Third, as McGill (1954) noted, Q-measures for tertiary and higher order relations may assume negative values. This need not be disturbing. It occurs whenever lower-order relations exist that, when taken in conjunction, overdetermine the relation in question. For example, the following three-valued data, expressed as 3rd-order probabilities with values 0 or 1 in each variable:



$p_{111} = .5$

$p_{000} = .5$

yield the following account:

$$\begin{aligned} Q(A:B) &= 1 \\ Q(A:C) &= 1 \\ Q(B:C) &= 1 \\ \underline{Q(A:B:C) = -1} & \\ T(A:B:C) &= 2 \end{aligned}$$

Here from any two binary relations, depicted in form of contingency tables below:

A \ B	0	1
0	.5	0
1	0	.5

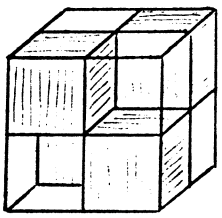
A \ C	0	1
0	.5	0
1	0	.5

B \ C	0	1
0	.5	0
1	0	.5

the whole three-dimensional distribution can be constructed and the third binary relation is implied. The measure associated with the third binary relation is then redundant and is compensated for by a negative measure for the tertiary relation. Generally, if a Q(rth-order)-measure is negative, then its $2^r-(2+r)$ lower-order relations overdetermine the rth-order relation in question. Conversely, if a Q(rth-order)-measure is positive, but the analysis of the data proceeds on the basis of lower-order relations only, then these lower-order relations underdetermine the relation manifest in data. The latter is a common inadequacy of many multivariate techniques as will be discussed below.

Fourth, Q-measures assess logically independent qualities as required by any spectral analysis. They neither contain each other nor influence each other except where they need to compensate for overdetermination by lower-order relations. In other words, the absence of an rth-order relation that an appropriate Q-measure will detect does not lead to any inferences either about relations of an order lower than r nor about relations of an order higher than r. The assumption that lower-order associations are a prerequisite of higher-order associations is a common mistake made by many multi-variate techniques and is easily refuted by the following example:

$p_{111} = .25$
$p_{100} = .25$
$p_{010} = .25$
$p_{001} = .25$



$Q(A:B) = 0$
$Q(A:C) = 0$
$Q(B:C) = 0$
$Q(A:B:C) = 1$
$T(A:B:C) = 1$

If the analysis of such data were restricted to analysing binary relations only, then the analyst would jump to the conclusion that his data contains no structure because all projections of this cube onto its three two-dimensional representations show 2nd-order probabilities to be uniformly distributed. However the spectral analysis of higher-order relations reveals that whatever can be said about the data is manifest in a strong and non-decomposable tertiary relation. The analysis does not assume that higher-order relations are decomposable into lower-order relations. It considers this to be a question that only data can answer.

Finally, a more obvious point is that Q-measures are symmetrical, for example:

$$Q(A:B:C) = Q(A:C:B) = Q(C:A:B) = Q(B:C:A) = \ldots$$

The proof lies in the definition of Q in terms of sums of H which are freely permutable. A positive value for Q(rth-order) simply indicates which r variables need to be considered in a suitable theoretical account for the data.

Distributions

Probably the most important remaining problem associated with the proposed spectral analysis of relations is the lack of knowledge of the distribution of its Q-measures and the consequent difficulty of testing statistical hypotheses. The fact that each additional variable adds to the degree of freedom, makes a comparison of the corresponding Q-measures actually difficult. The expression of the Q-measures in percentages of the total amount of relation, T, accounted for might falsely suggest that only the magnitudes of these measures matter. In general, if the values of two different Q-measures are the same, that associated with the higher-order relation is more likely to be significant than that associated with the lower-order relation. But, to my knowledge, distributions for Q-measures have not been worked out, and, if they are, they are known not to be chi-square like distributions because Q-measures may have negative values. However, the situation is not quite as desperate as this might indicate.

McGill (1954) has shown that large sample distributions of the likelihood ratio λ may be used to find approximate distributions for the total amount of transmission, T. In particular he showed that, assuming independence of the variables in T, and with n denoting the sample size:

$$-2 \log_e \lambda = 1.3863\ n\ T(:r\ variables:)$$

For large samples, $-2\log_e\lambda$ is known to have approximately a chi-square distribution provided the null hypothesis is true. Hence, 1.3863 n T(:r variables:) is distributed approximately as chi-square if the population T(:r variables:) equals zero. With α_i denoting the number of values of the ith variable, the degrees of freedom for the total amount of relation is:

$$df_{T(:r\ variables:)} = \left(\prod_{i=1}^{r} \alpha_i - 1 \right) - \sum_{i=1}^{r} \left(\alpha_i - 1 \right)$$

To take advantage of this approximation to the asymptotic distribution, one is forced to obtain T-measures by summing the appropriate Q-measures. To test null-hypotheses for T-measures rather than for Q-measures is quite appropriate when it is the task to formulate an analytical construct, model, function, or explanation and the analysis aims at identifying the variables within which relations are manifest. With this limited task in mind, further advantage can be taken from the decomposition of variables into separate sets of variables in which case the total amount of relation can be expressed as the algebraic sum of the amounts of relation within each set and between the sets. For example:

$$T(A:B:C:D:E) = T(A:B) + T(C:D:E) + T(A,B:C,D,E) .$$

The latter quantity expresses the degree to which binary relations hold between the two product sets A×B and C×D×E. If the null-hypothesis can be accepted for this binary relation which means that the two product sets are independent, then, according to Ashby's theorem concerning Q-measures, the null-hypothesis can be accepted for all Q-measures containing variables of both sets.

This indirect procedure is weak, however, when it comes to the decomposition of a relation into its lower-order relations (as opposed to the decomposition of variables into mutually exclusive sets). For example, if T(C:D:E) and T(C:D), T(C:E) and T(D:E) are all found to be significantly above zero, then it remains uncertain whether the loss Q(C:D:E) incured by describing the three-valued data in terms of the three binary relations is significant. The picture changes, however, if some of the T-measures turn out not to be significantly different from zero. So, the testing of null-hypotheses for T-measures is not entirely powerless.

Bi-ordinal Multivariate Techniques

Basic to the following is that not all relations are decomposable into relations of lower ordinality without loss, and, Relations of lower ordinality do not imply relations of higher ordinality. Both points have been adequately demonstrated above. This section attempts to show that many multivariate techniques, by the assumptions they maintain, ignore what are perhaps the most important characteristics of many-valued relational data: higher-order relations.

The most elementary multivariate technique involves graphs to depict and to analyse group structures, communication networks, coding processes, mapping functions and transformations (see Harary, Norman and Cartwright, 1965, for examples). A graph can be thought of as a collection of lines, each of which connects two of a possibly large number of nodes. Nodes may be representative of individuals as in sociometric choice experiments, of symbols in the case of coding processes, of a system's states in the case of transformations, etc. Graphs can be mapped into two-dimensional matrices as demonstrated in the second example above. As subsets of the possible entries in such a matrix, graphs satisfy the definition of a binary relation. And there is simply no way that unique combinations of three or more nodes could be depicted as a graph. Graph theory is therefore powerless in the face of non-decomposable higher-than-2nd-order relations, which is well recognised in the axioms of that theory.

It is important to realize that the limitation to bi-ordinality does not change with the denotation of the entries in a two-dimensional matrix.

For example, clustering procedures lump, put together or group the nodes of a two-dimensional matrix in such a way that maximally similar nodes or those with greater proximity turn out in the same cluster. Insofar as it proceeds from measures of

similarity, distance or proximity, clustering is a bi-ordinal technique regardless of the size of the emerging clusters.

Similarly, most factor analyses start out or compute as a first step a two-dimensional matrix of correlations. By operating on pairwise correlations only, factor analysis reveals itself unable to respond to the possible presence of higher-than-2nd-order correlations.

Systems of linear equations of the kind given in the third example above do not reach higher-order relations either. The coefficients α_{ij} merely provide another kind of entry in a two-dimensional matrix whose rows and columns are the variables x_i and x_j. Incidentally, the factor loadings resulting from a factor analysis can be interpreted in terms of α_{ij}.

Analysis of variance is another case in point. Variance, usually conceptualized as the sum of the squared deviations from the mean, can be expressed in numerous ways, among them as the average squared difference between all pairs of data points in a sample:

$$V(Y) = \sum_{y \varepsilon Y} p_y (y-\bar{y})^2 = \frac{1}{2} \sum_{b \varepsilon Y} \sum_{c \varepsilon Y} p_b p_c (b-c)^2$$

Variance analysis offers a convenient calculus for differences within and between variables and ascertains how these differences can be explained. But, whichever differences are compared, differences are expressed only between pairs. Tertiary or higher-order differences between unique combinations of three or more values are difficult to conceptualize and indeed do not enter variance analysis.

Network analysis, cluster analysis, factor analysis, analysis of variance and many others accept many-valued relational data and indeed consider many-dimensional

point distributions of those data, hence the attribute "multivariate," but their built-in assumptions prevent them from exploiting the multi-ordinal characteristics that such data are powerful enough to contain.

In the social sciences, higher-order relations have not been entirely ignored, however. For example, the analysis of Markov processes is not restricted to bi-ordinality. If 2nd-order probabilities fail to explain or to reproduce the process as recorded, it is customary to proceed with 3rd-order probabilities, 4th-order probabilities, etc. until a satisfactory explanation is found. Because lower-order probabilities do not imply higher-order probabilities, each increase in ordinality requires new references to data and cannot be obtained from lower-order representations. Another example is the use of conditional product-moment correlations $r_X(Y:Z)$. These are often employed to check on spurious correlations. Such measures express nothing but correlations between two variables relative to a third and do not give equal weight to all variables involved. However, considering how Q-measures that are expressed relative to a variable j can give rise to Q-measures of higher ordinality, including j:

$$Q(r+1\text{th-order}) = Q_j(r\text{th-order}) - Q(r\text{th-order})$$

conditional correlations are a step in this direction. Other examples are provided by the use of systems of non-linear equations involving products and exponents (rather than sums only) of the values of three or more variables.

If the decomposition of complex relations into several binary relations were indeed empirically justifiable, then this would result in considerable analytical savings in terms of smaller sample sizes, simple units of enumeration, and theory that is closer to verbal discourse. However, I believe that such justifications are rarely found in fact. Attempting to find simple examples of 2 by 2 by 2 contingency

tables, I computed all configurations of frequencies up to n=19 and could find many cases in which the configuration could be explained in terms of two binary relations but none in which the tertiary Q-measure is zero and all three binary relations positive. Ashby (1965) gives the following example for the latter:

$$
\begin{aligned}
p_{010} &= .276906 \\
p_{011} &= .169281 \\
p_{100} &= .138453 \\
p_{101} &= .346133 \\
p_{110} &= .069227 \\
&\overline{1.000000}
\end{aligned}
$$

all Q(2nd-order) > 0
Q(3rd-order) = 0 to 6th decimal

Higher-order relations can be ignored, for example, when people communicate with each other only by phone and without reference to a third party, when all cooperative ventures in society involve no more than two individuals, when social processes are chain-like with each event dependent on one predecessor only, when only distances, differences and pairwise comparisons matter, etc.. Given the breadth and complexity of social reality however, such restrictions are rarely satisfied and yet bi-ordinal techniques force their user to see the world that way.

If higher-order relations are not decomposable into binary ones, as is to be expected when communication involves more than two individuals, when organizations incorporate hierarchies, when social processes coordinate many activities, etc., then bi-ordinal techniques are simply not powerful enough to capture what seems to be the essence of social reality.

Though it does not provide all the answers, the proposed spectral analysis of relations goes a step beyond most multivariate techniques toward a fuller utilization of the relations in multi-variable data collected from complex social situations.

References

Ashby, W.R., "Constraint Analysis of Many-Valued Relations," General Systems, 9: 99-105, 1964.

Ashby, W.R., "Measuring the Internal Information Exchange in a System," Cybernetica, 8,1:5-22, 1965.

Ashby, W.R., "Two Tables of Identities Governing Information Flows Within Large Systems," Am. Soc. for Cybernetics Communications,1,2:3-8,1969.

Bartlett, M.S., "Contingency Table Interactions," Supplement to the Journal of the Royal Statistical Society,2:248-252,1935.

Harary, F., Norman, R.Z. and Cartwright, D., Structural Models: An Introduction to the Theory of Directed Graphs, New York: Wiley, 1965.

McGill, W.J., "Multivariate Information Transmission," Psychometrica,19:97-116, 1954.

Rummel, R.J., "Understanding Factor Analysis," Journal of Conflict Resolution, 11,4: 444-480, 1967.

Simon, H.A., "The Architecture of Complexity," pp.84-118, in his The Sciences of the Artificial, Cambridge Mass.: MIT Press, 1969.

Wiener, N., "A Simplification of the Logic of Relations," Proceedings of the Cambridge Philosophical Society, 17:387-390, 1914.