Fred Emery

Causal Path Analysis[1]

"Causal path analysis" is the accepted designation for attempts to empirically identify a system of coupled elements. In social systems, we are seeking to identify producer-product relations. The logical requirements for inferring producer-product relations are

1. Correlations between two or more variables cannot, by themselves, establish that any of these variables are producers or co-producers of any others, i.e., that they constitute even necessary conditions—let alone sufficient conditions—for any other variable.
2. Postulating the state of one variable x as a necessary condition for the state of another variable y, which requires that two judgmental conditions are met:
   (a) That x is a *possible* producer of y. Thus we would not regard x as a possible producer of y if it occurred in time after y, if there is no conceivable path and if there is no conceivable mechanism, e.g., for an unassisted mouse to move a mountain.
   (b) That x is also a *probable* producer of y (i.e., observed concomitance); observed r_{xy} is greater than the r_{xy} possible by chance alone.
3. The conditions stated in requirement 2 are sufficient if it can *also* be established that they are true of x regardless whether any other variable z is also a probable producer of y.

To establish that x is a necessary condition for the state of y regardless of other variables, it is necessary to establish that the probability of x producing y is significantly greater than the probability of any other possible producer producing *both* x and y, i.e., that r_{xy} is significantly greater than $r_{zy} \times r_{zx}$. We are assuming that x has met the conditions specified in requirement 2 of possible and probable production. We are not requiring that x is a more probable producer than z (else we would never identify more than one necessary condition and would exclude the whole notion of co-producers).

If the conditions in requirement 2 are met, but not this condition with respect to x and y, then the observed correlation between x and y is a "spurious correlation."

[1] Excerpted from F.E. Emery and C. Phillips, *Living at Work*. Canberra: Australian Government Publishing Service, 1976.

The Problems of Analysis

Let us take as given a set of n variables for which there are indices of concomitance (e.g., correlations) for the states of all variables within the same individuals at the same moment of time (i.e., within a period of time within which the relation of the variables can be assumed not to have changed, e.g., that in doing a battery of tests growing fatigue has not changed performance in a test).

For this set of n variables, the number of possible direct, one-step, relations R between them is given by $(n - 1)$ factorial, or the following equation:

$$R = \frac{n\,(n - 1)}{2}$$

This number rises rapidly as n increases: $n = 2, R = 1$; $n = 6, R = 15$; $n = 12, R = 66$.

Theoretically, all these relations have to be examined to determine whether the conditions specified in requirement 2 are met, i.e., one has to determine of each observed relation whether one is a *potential* producer of the other, a *possible* producer, a *probable* producer. This has to be done for x with respect to y and for y with respect to x. The possibilities that have to be considered for n variables increase threefold.

The major problem is, therefore, one of sheer complexity, not a problem of logic.

Reduction of complexity is, in the first place, aided by setting levels of statistical significance for the correlation coefficients. Correlations below this are regarded as not probable as the observed relation could be due to chance factors alone.

The next step to reduce complexity has usually been to postulate specific causal models and test these for goodness of fit with the observed correlations. This is a subjective step and usually conservative in that the researcher will tend to postulate models that conform to current belief.

This model of "causal path analysis" is usually associated, in the social sciences, with the name of Blalock (1964). It is a carryover into the social sciences of the statistical method developed, decades before, by Sewell Wright (1934; 1960) to track the transmission of characteristics within domestic animal breeding stocks. The carrying over of this method ignored a basic assumption in Sewell Wright's method. He was justified in accepting the logically stronger form of regression equations (relative to correlation coefficients that average the regression of variable A on variable B and of B on A), because he always knew which of his animals were the parents and which the offspring.

He knew which was the independent variable and which the dependent variable. In the social sciences there is rarely such certainty. We cannot pass lightly over this blindness to basic assumptions because we, in the social sciences, have made the earlier error of uncritically adopting R.A. Fisher's (1935) statistical methods for analysis of variance in yields of small, contiguous agricultural plots to different levels and types of fertilizer treatment (cf. Chow [1992] on "the agricultural model of scientific experimentation"). Fisher could assume without doubt that the crop growth was the dependent variable and the fertilizer treatments the independent variable. He did not have to worry about whether it might be the fertilizer that grew or that the crop might act on whether it approved or took umbrage at the fertilizer treatment (Moghaddam and Harre, 1992; Silverman, 1977). Fisher knew beforehand that randomization of contiguous plots in a small area would overcome micro-climate and soil variations. He knew also that other possible sources of variation, e.g., wandering cattle, were definitely fenced off. It is a very lucky statistician in the social sciences who can be so sure about what is under control and what are the independent and the dependent variables. Social scientists have to confront genuine *multi*variate analysis; and the variables they conceptualize are concrete universals, not the abstract universals that mathematicians and formal logicians work with. Concrete universals have members that also have the characteristics of other concrete universals. Simply put, the concepts of the social sciences cannot be assigned, with assurance, to classes of independent, intervening or dependent variables.

A simple example might indicate the difficulties that statisticians face in the social sciences. In a study of smoking and stress (Emery et al., 1968), we were expecting to find patterns like the following:

social class $\rightarrow$ personal stress $\rightarrow$ smoking

We expected this because social class is so often the most independent, least easily changed, of all the variables. As it turned out, for young women (not for other sex and age groups) the pattern that emerged from the observed correlations was

personal stress $\rightarrow$ social class $\rightarrow$ smoking

What was indicated by this result was that social class, in this instance, was acting as an intervening cultural variable, not as an independent life conditions variable. That is, whether a young woman under stress smoked depended on whether she was in a culture that accepted this behavior or in a middle-class (English) culture that frowned on smoking by young women.

Using the regression-based model of causal paths, we would have tested the data for the first—the expected—pattern. This might, or might not, have given an acceptable statistical fit. In any case, we would have missed the explanation of "class as culture." The evidence in our data would have been wasted, and social science all the poorer.

As an alternative, we have sought to manage complexity by the step-by-step procedures evolved by McQuitty (1960) to graph matrices of correlations according to their ordinal properties (elementary linkage analysis) and gradually to reduce a matrix by successively combining variables most like each other.

A Solution

The usual method of causal path analysis, postulating a particular pattern of causal relations between all of the variables and then testing for goodness of fit with the observed correlations, has serious weaknesses. One is that the number of possible patterns between n variables increases enormously as n gets beyond 4 or 5, and the method soon becomes inapplicable unless one wishes to make strong a priori assumptions about what are the causes, and thus delimit the number of patterns that will be considered.

The method is inherently conservative. It is not conservative in the narrow sense of tending to reduce the ratio of "false positive" to "false negative" errors. It is conservative in the much more debilitating way of constraining search to the replication of existing findings. For publication, researchers usually, not always, look for positive findings. As the method handles so few variables, and hence so few patterns, one naturally tries first to see if the well-founded (and respected) patterns appear in one's own data.

Knowing the restrictions on the number of variables that can be handled by this method of analysis, a competent survey designer is likely to reduce costs by studying less of the context of the phenomena being studied.

There is one property of causal paths that can be utilized to approach the problem in a different way. If three variables, A, B and C, are correlated so that $r_{AC} > r_{AB}$ or r_{BC} and $r_{AB} > r_{BC}$, then any causal inferences must consist of putting arrows into the following graph (Blalock, 1964):

$$B—C—A$$

Any causal hypotheses that implies a variation in the structure of the graph, e.g.,

$$A—B—C \text{ or } A—C—B—$$

would be contrary to the observed correlations.

The mathematical technique of elementary linkage analysis, devised by

McQuitty for cluster analysis, enables any matrix of correlations between variables (or subsequent reiteration of this matrix—by McQuitty's method of hierarchical linkage analysis) to be represented by a unique graph or set of graphs which have the property discussed in 2b (p. 328).

Following this procedure, one is no longer restricted to a handful of variables and, of particular value, the process of making causal inferences is clearly separated out as a stage that comes after the nonsubjective stage of ordering variables simply according to the observed values of the correlations. The strength of this approach is that it eliminates a great number of possible models that would be contradicted by the data. However, the analysis yields nothing which can tell us which way the arrows go, if indeed there are grounds for inferring any arrows. Such judgments must be based on other considerations (see 2a, p. 328).

If there are causal relations between the variables then they will have to correspond to the graph yielded by McQuitty's elementary linkage analysis. Two further points need to be made:

- Analysis of the prime matrix may not reveal a graph linking together the variables with which one is concerned. In this case, the matrix must be reiterated by McQuitty's method of Hierarchical Linkage Analysis. No further reiterations are required when the level is reached where the variables of interest are represented in the one connected graph (or where it is clear that no further reiterations will bring them together) *and* the graph is internally consistent, i.e., for variables i and j, n steps removed from each other, r_{ij} is less than the correlation in the sequence i to j between i, j or any other variables less than n steps removed from each other. That is, the magnitude of the correlation should be smaller between variables further removed from one another in the graph.
- The McQuitty elementary linkage analysis graphs the matrix in terms of one-step relations, e.g., $A—B—C$. It does not indicate whether, for instance, A has any direct effect on C over and above its primary influence via B. This has to be determined by an independent operation to assess whether $(r_{ac} - r_{ab} \times r_{bc})$ is significantly greater than zero. If it is, the relation of AC should be entered in the graph as a broken line and, if possible, a causal arrow inferred.

Matrix of Intercorrelations of Major Variables

Table 1 presents the bare bones of an example taken from the same study (Emery and Phillips, 1976). Initial analysis of the 2,000 questionnaires, by the methods outlined above, revealed eight items and eight scales, composed of empirical clusters of items, to be a closely related set (Table 1).

Reiteration of Table 1 yielded the unique graph presented in Figure 1. The

TABLE 1 Intercorrelation of the 16 Major Variables Entering into Figure 1 (sample size = 2,000)*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	—	38	35	32	37	31	25	30	28	12	30	18	21	27	20
2	38	—	31	35	51	38	34	34	31	8	38	17	20	28	17
3	35	31	—	27	28	21	18	20	20	−1	59	14	15	24	20
4	32	35	27	—	38	30	26	26	28	7	28	11	15	15	11
5	37	51	28	38	—	56	32	35	32	7	30	21	20	21	17
6	31	38	21	30	56	—	25	29	32	9	20	14	20	10	5
7	25	34	18	26	32	25	—	63	21	15	25	27	18	16	13
8	30	34	20	26	35	29	63	—	27	16	23	29	24	19	16
9	28	31	20	28	32	32	21	27	—	9	16	15	28	5	4
10	12	8	−1	7	7	9	15	6	9	—	−2	4	6	4	5
11	30	38	59	28	30	20	25	23	16	−2	—	12	14	32	21
12	18	17	14	11	21	14	27	29	15	4	12	—	16	5	5
13	21	20	15	15	20	20	18	24	28	6	14	16	—	15	13
14	27	28	24	15	21	10	11	19	5	4	32	5	15	—	50
15	20	17	20	11	17	5	13	16	4	5	21	5	13	50	—
16	19	17	20	9	12	7	11	11	4	6	26	5	15	41	30

* 1. Objective quality of work scale (Qs, 2a, 4a, 5a, 5d, 18a, 35). 2. Subjective quality of work (Qs, 1a, 1b, 1d, 17, 22a). 3. Supervision (Q.41). 4. Social climate scale (Qs, 8, 10, 15, 21). 5. Job satisfaction (Q.19). 6. Potential labour turnover (Q.20). 7. Happy with life (Q.34, a, b, d, e, f). 8. Hopeful of life (Q.34, h, g, i). 9. Conditions of work (Q.22, less items a, h, and j.). 10. Health (Q.31, a, c, and f, Q.32). 11. 'Taylorism scale' (Qs, 6, 12, 14, 16). 12. Life expectations (Q.28). 13. Adequacy of income (Q.33). 14. Occupational status (Q.53b). 15. Educational background (Q.51b). 16. Actual income (Q.54).

arrows added to the graph are the only elements left to the judgment of the researchers.

Using Oscar Lange's (1965) terms it will be noted that: elements 2, 7 and 8 are *boundary* elements; 3, 4 and 5 *interior* elements and 1, 6 and 9 *environing* elements. Because it possesses boundary elements it describes an *open* system. Items 3, 4, 7 and 5 and items 4, 7 and 8 form closed chains (loops) of couplings.

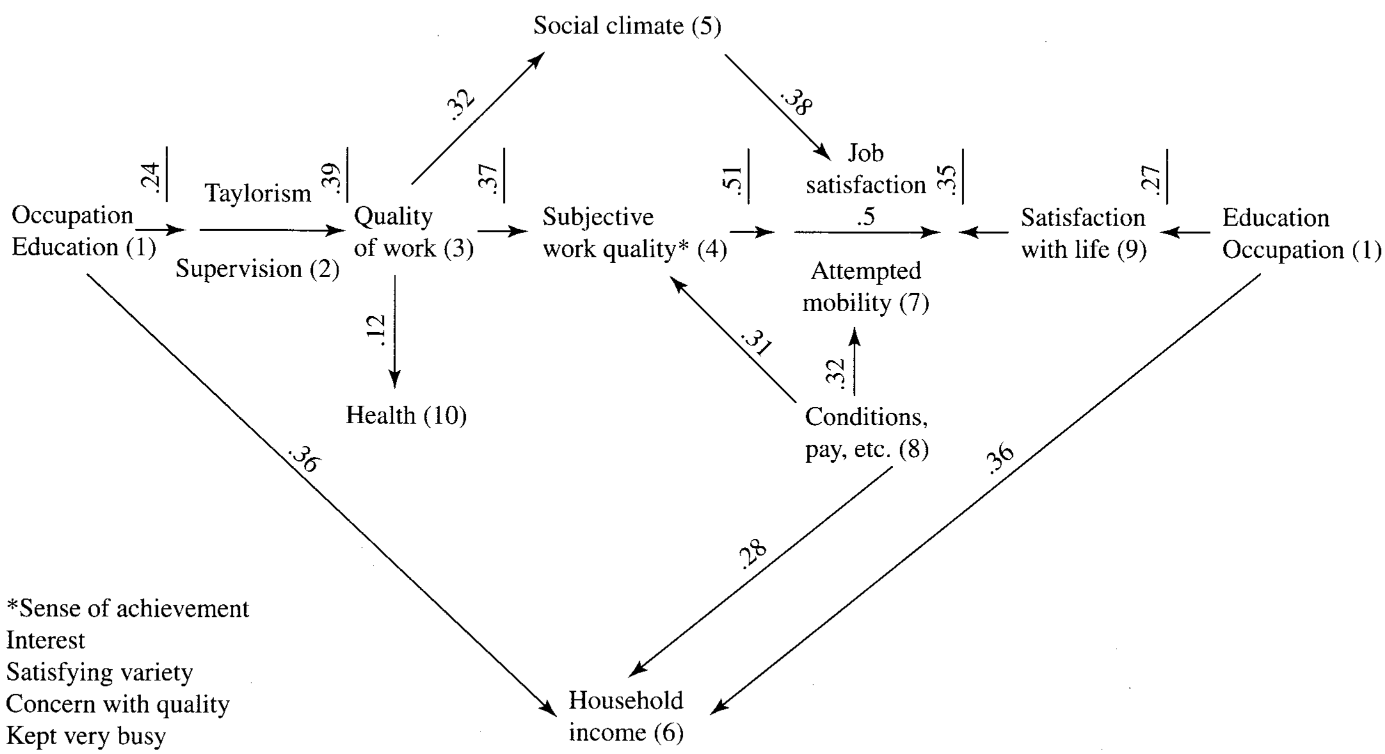

Figure 1. Graph derived from Table 1.

The graph shown in Figure 1 is a typical outcome of this kind of analysis of this kind of data. As an undirected graph (i.e., without the arrows) it tells us what interpretations the observed data restrict us to. The graph tells us that a great range of possible interpretations are ruled out by this particular body of facts.

By also presenting the matrix of correlations it is open to others readily and speedily, without recourse to even a calculator, to check for arithmetic errors in the transformation from the matrix to the graph.

If there are no arithmetic errors, then debate about the interpretation of the results is restricted to debate about whether the arrows should point one way or the other, or both ways (mutual determination) or whether there should be no arrows (mere concomitance or variables that are synonymous). Beyond that, is debate about the design of the study. It is always open to debate whether a study has failed to include measures of some potentially critical variable. This is less likely when the method of analysis places no constraints on the number of variables that can be analyzed. But it can and will still happen.

At the practical level, these graphs can be read like nonmetrical road maps (e.g., the famous London Underground train maps). If one wishes to go from *A* to *B* (to influence changes at *B* by making changes at *A*) the map tells you what stations you have to go through and where you might have to change trains.

References

Blalock, H.M. 1964. *Causal Inferences in Non-Experimental Research.* Chapel Hill: University of North Carolina Press.

Chow, S.L. 1992. "Positivism and Cognitive Psychology: A Second Look." In *Positivism in Psychology*, edited by C.W. Tolman. New York: Springer-Verlag.

Emery, F.E., E.L. Hilgendorf and B.L. Irving. 1968. *The Psychological Dynamics of Smoking*. London: Tobacco Research Council.

Fisher, R.A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Lange, O. 1965. *Wholes and Parts: A General Theory of System Behaviour*. London: Pergamon.

McQuitty, L.L. 1960. "Capabilities and Improvement of Linkage Analysis as a Clustering Method." *Educational and Psychological Measurement,* 24:441–56.

Moghaddam, F.M. and R. Harre. 1992. "Rethinking the Laboratory Experiment." *American Behavioral Scientist*, 36:22–38.

Silverman, I. 1977. *The Human Subject in the Psychological Laboratory.* New York: Pergamon.

Wright, S. 1934. "The Method of Path Coefficients." *Annals of Mathematical Statistics*, 5:161–215.

———. 1960. "Path Coefficients and Path Regressions: Alternative or Complementary Concepts?" *Biometrica*, 16:189–292.