A Measure of Past Collaboration[1]

Stephen P. Borgatti & Candace Jones
Carroll School of Management, Boston College

Suppose you are interested in studying project teams. In particular, you would like to know if teams composed of people who have worked together on other project teams will be more successful than teams in which the members have never worked together. But rather than simply classifying teams into two categories (1=some members have worked together before; 0=none have worked together), you would like to define a continuous variable indicating the extent to which group members have worked together.

There are many approaches that can be taken here. One is a multidimensional set of variables that separately measures different aspects of working togetherness, such as:

- The number of individuals that have experience with any other member of the team
- The number of pairs, triples and other combinations of members that have worked together before
- The number of times (projects) that any combination of members has worked together in the past

Another approach is to create a single measure that is intended to capture all of these aspects. This is the approach we take here.

As networkers, our every instinct says 'for each team, construct a member-by-member matrix X of past collaborations, where x_{ij} gives the number of projects that members *i* and *j* have collaborated on in the past, then sum all the entries in the matrix'. This measure could also be normalized by dividing by the number of pairs, giving a kind of density of interaction.

The trouble with this measure is that it double-counts projects in which more than one pair of present members collaborated. For example, if a past project included three members of the current team, then it would be counted for *i* and *j*, for *i* and *k*, and again for *j* and *k*. While it is not completely clear that this is undesirable, we wanted to develop a measure that did not have this "feature".

Perhaps the easiest way to do this is to first create a measure of non-collaboration and then reverse it by taking either its additive or multiplicative inverse.

Consider a team composed of individuals A, B, C, and D. According to their résumés, these individuals have previously worked on 1, 2, 3, and 4 projects respectively. If none had ever collaborated before, then all these projects would be distinct, and the total number of projects in their collective experience would

[1] ***Techniques*** *is a regular column devoted to techniques of data construction, management, interpretation and analysis. Contributions are appreciated.*

be 1+2+3+4=10. By looking at which projects each individual actually worked on, we can count how many different projects there really were. A reasonable measure of non-collaboration can be constructed by dividing the actual number of distinct projects by the maximum possible (in this case, 10). This yields the following definition:

$$\eta = \frac{|A \cup B \cup C \cup ...|}{|A|+|B|+|C|+...}$$

where the letters A, B and C refer to individuals (who are represented by the sets of projects in their past), and the notation | X| indicates the size of a set X. The numerator of the formula, then, is the number of different projects in the collection of all résumés, and the denominator is the sum of the sizes (*i.e.*, number of projects) of each résumé.

A measure of collaboration can then be defined as one minus non-collaboration as follows:

$$\zeta = 1 - \eta$$

When the résumés are completely disjoint, η achieves its maximum value of 1/ 1=1.0, and ζ achieves its minimum value of 1-1=0.0. When the résumés are identical, ζ achieves its maximum value of 1-0=1, which occurs when the team members have always and only worked with each other in the past.

One difficulty with this measure concerns its maximum value, which is determined by the minimum value of η. If we take as given the size of each individual résumé, then we can see that the minimum value of the numerator of η is equal to the size of the largest vita. In other words, even when there is perfect past collaboration, the fewest possible number of distinct projects is at least as big as one individual's own set of distinct projects.

For example, consider Figure 1 which depicts the projects belonging to the résumés of three people (A, B and C). A is the set {1,2,3,4}, B={4}, and C={1,2,4}. According to our formula, η = 4/ 8 = 0.5, and ζ = 1 - 0.5 = 0.5 which suggests middling overlap in résumés. Yet the projects of these individuals overlap as much as possible given that some individuals were on fewer projects than others.

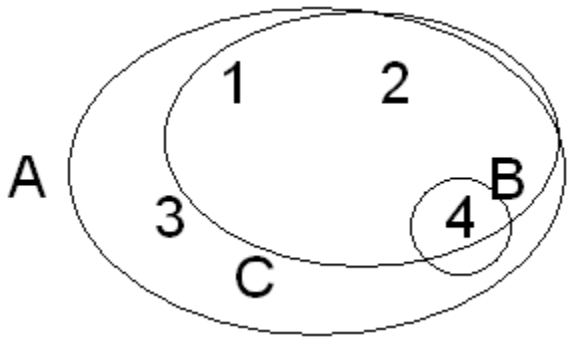

Figure 1. Circles are resumes of persons A,B,C. Numbers are ID codes of their past projects.

To correct this problem, we can subtract the minimum possible number of distinct projects from both the numerator and the denominator of η. This minimum value is equal to the size of the largest vita. The corrected formula for η is:

$$\eta' = \frac{|A \cup B \cup C \cup ...| - Max(|A|,|B|,|C|...)}{|A|+|B|+|C|+... - Max(|A|,|B|,|C|...)}$$

and

$$\zeta' = 1 - \eta'$$

When applied to the situation in Figure 1, the corrected η' equals (4-4)/ (8-4)=0, and ζ'=1.0.

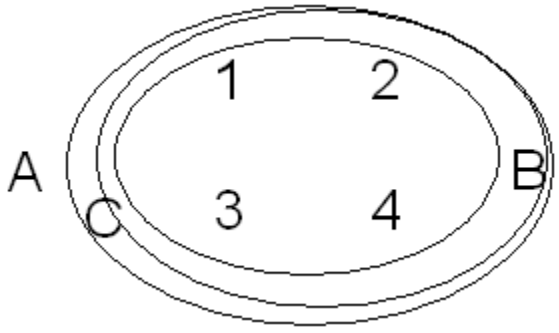

Figure 2. All three individuals have identical résumés.

Both ζ and ζ' are useful measures: neither is better than the other in all situations. For

example, consider the situation depicted in Figure 2. The new measure is again 1.0, and so does not distinguish between the situations in Figure 1 and Figure 2. In contrast, the old measure only achieves 1.0 in the second figure, indicating more past collaboration in Figure 2.

Which is better depends on the circumstances of the research in which they will be used. For example, if non-collaboration is being used as an index for the amount of exposure that team members have had to "foreign" ideas learned in other projects (i.e., a diversity index), it seems likely that the normalized measure η' is the preferred one; in Figure 1, none of the members have worked on wholly outside projects, and $\eta' = 0.0$.

In contrast, if the argument is that the more that all of the people work together the better their team performance, then we want a measure that distinguishes between Figures 1 and 2, since in Figure 2 more of the people have worked more often together.

EXAMPLE

Consider the well-known dataset collected by Davis, Gardner and Gardner (1941) in the which the rows are women and the columns are events as shown here:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
EVELYN	1	1	1	1	1	1	0	1	1	0	0	0	0	0
LAURA	1	1	1	0	1	1	1	1	0	0	0	0	0	0
THERESA	0	1	1	1	1	1	1	1	1	0	0	0	0	0
BRENDA	1	0	1	1	1	1	1	1	0	0	0	0	0	0
CHARLOTTE	0	0	1	1	1	0	1	0	0	0	0	0	0	0
FRANCES	0	0	1	0	1	1	0	1	0	0	0	0	0	0
ELEANOR	0	0	0	0	1	1	1	1	0	0	0	0	0	0
PEARL	0	0	0	0	0	1	0	1	1	0	0	0	0	0
RUTH	0	0	0	0	1	0	1	1	1	0	0	0	0	0
VERNE	0	0	0	0	0	0	1	1	1	0	0	1	0	0
MYRNA	0	0	0	0	0	0	0	1	1	1	0	1	0	0
KATHERINE	0	0	0	0	0	0	0	1	1	1	0	1	1	1
SYLVIA	0	0	0	0	0	0	1	1	1	1	0	1	1	1
NORA	0	0	0	0	0	1	1	0	1	1	1	1	1	1
HELEN	0	0	0	0	0	0	1	1	0	1	1	1	1	1
DOROTHY	0	0	0	0	0	0	0	1	1	1	0	1	0	0
OLIVIA	0	0	0	0	0	0	0	0	1	0	1	0	0	0
FLORA	0	0	0	0	0	0	0	0	0	1	0	1	0	0

Let us pretend that the events are ordered chronologically, so that event 1 was the first event and event 14 was the last. This is not a requirement of the method, but shows how to apply the method in the case of time-ordered data (such as film projects, basketball games, and so on). To calculate the η and ζ coefficients for each event, we start from the left and consider all events up to but not including the current event. The coefficients cannot be calculated for the first event since at that point all women have empty event histories. For, say, event 3, η=2/ 6, so ζ=.67. The corresponding normalized figures are η'=(2-2)/ (6-1)=0.0 and ζ'=1.0. Statistics for all the events (except #1) are as follows:

	Raw		Normalized	
Event	η	ζ	η'	ζ'
2	0.500	0.500	0.000	1.000
3	0.333	0.667	0.000	1.000
4	0.375	0.625	0.000	1.000
5	0.250	0.750	0.000	1.000
6	0.250	0.750	0.000	1.000
7	0.363	0.727	0.059	0.941
8	0.194	0.806	0.033	0.967
9	0.286	0.714	0.048	0.952
10	0.286	0.714	0.091	0.909
11	0.556	0.444	0.200	0.800
12	0.240	0.760	0.050	0.950
13	0.350	0.650	0.971	0.929
14	0.333	0.667	0.959	0.941

All of the events have high collaboration scores, which means that they tend to be attended by people who have attended many other events together. An exception is event 11, which brings together people that tend to attend different events (when they attend at all).

References

Davis, A., B.B. Gardner and M.R. Gardner. 1941. Deep South: A Social Anthropological Study of Caste and Class. Chicago: University of Chicago Press.