A Statistical Method for Comparing Aggregate Data Across A Priori Groups

STEPHEN P. BORGATTI
Boston College

This article introduces a new method for statistically comparing pairs of aggregate data series. Aggregate data series refers to a set of values, each of which is averaged or otherwise aggregated across respondents. The motivating problem is the comparison of aggregate proximity matrices, such as those obtained from pile sort exercises. The standard approach to this problem uses the nonparametric, permutation-based quadratic assignment program (QAP) technique. However, the null distribution that QAP is based on is inappropriate for comparing subsamples of a data set and may lead to misleading conclusions. The new method can yield different results than QAP, results more in line with researchers' intuition. Furthermore, the method can be applied to a variety of data types beyond those appropriate for QAP.

Suppose we ask male and female respondents to rate the similarity of all pairs of items in a cultural domain (Weller and Romney 1988). We average the data separately for men and women and obtain two aggregate similarity matrices, one for each gender. We want to compare the matrices to see if the genders perceive the domain differently. Typically, we are interested in a measure of the magnitude of similarity (or difference) between the matrices and a significance test (typically a statement of the likelihood of obtaining such a high level of similarity or difference given independence of gender and perception). I am particularly interested in the significance test.

The thesis of this article is that the approach most commonly used in these cases is inappropriate and leads to the wrong conclusion. I introduce a new method, based on permutation techniques (Edgington 1969; Good 1994), that is appropriate for comparing any two aggregate data series across a priori groups. By aggregate data series, I mean a set of values for each respondent (e.g., the cells of a proximity matrix), each of which is averaged or otherwise aggregated across respondents.

I gratefully acknowledge the helpful comments of Martin Everett, Lin Freeman, John Gatewood, Penn Handwerker, Jeff Johnson, David Krackhardt, Gery Ryan, Tom Snijders, and an anonymous reviewer on an earlier version of this article.

Field Methods, Vol. 14, No. 1, February 2002 88–107

A few words on the kind of significance test desired are in order. Classical significance tests are fundamentally concerned with evaluating the probability of obtaining a sample statistic (e.g., correlation) as large as the one actually observed due to sampling error, given that in the population the statistic is zero. In other words, the classical significance test is fundamentally about estimating the probability that a sample is deviant if the variables in question are independent in the population (Noreen 1989). This is undeniably useful in the case where we have a random sample and are interested in generalizing to a defined population from which the sample was taken.

However, the cases that I am concerned with are ones in which the sample is not necessarily random or the data are not drawn from a sample at all. Here, the significance we are interested in refers only to the probability of obtaining a test statistic as large as the one actually observed, given that the values of the variables are assigned independently of each other.

For example, suppose we administer a trivia test to boys and girls and find that the average score for the boys is eighty-five and the average score for the girls is seventy-nine. Does this mean that in this group, knowledge is a function of gender? It looks like it, but before answering we should remember that any division of the sample into two groups will yield somewhat different averages, even if the basis for the division had nothing to do with knowledge. If we divided the group by shoe color, for instance (light and dark), one group would almost surely have a higher mean than the other—in fact, it is virtually impossible to divide a sample into groups such that the mean score for each group is exactly the same.

So the real question is, How likely would it be to find a difference as large as actually observed (six points, in our example) if we divided the sample randomly (or more precisely, without regard for test scores)? If individual variability in test scores is high, then large differences in the means of randomly selected groups become relatively likely. To evaluate the observed difference we need to calculate this probability, and it is this probability that we refer to as the significance of the test statistic.

The need for a nonparametric procedure to provide a significance test for correlations between proximity (or other 1-mode) matrices has been noted before, and a standard solution, known as the quadratic assignment program (QAP) (Mantel 1967; Hubert and Schultz 1976), has wide currency today. For example, Boster and Johnson (1989) used QAP to compare pile sorts of fish by novice and expert fishermen. Weller (1984) used QAP to compare pile sorts of illnesses by U.S. and Guatemalan women. Johnson, Mervis, and Boster (1992) used QAP to compare proximities generated by children and adults. And Boster (1987) used the method to compare judged similarities of birds by naïve respondents and ornithologists. However, I believe that using

QAP to compare aggregate data is inappropriate and can yield seriously misleading results. In this article, I present evidence of the problem and offer a possible solution.

ILLUSTRATION OF THE PROBLEM

For a class project, three undergraduate students (Michelle Pirelli, Carolyn Canty, and Jennifer Buchholz) collected pile sort data on the domain of "things that people are afraid of" from male and female undergraduates. The question they posed was simple: Do the sampled men and women perceive the domain differently?

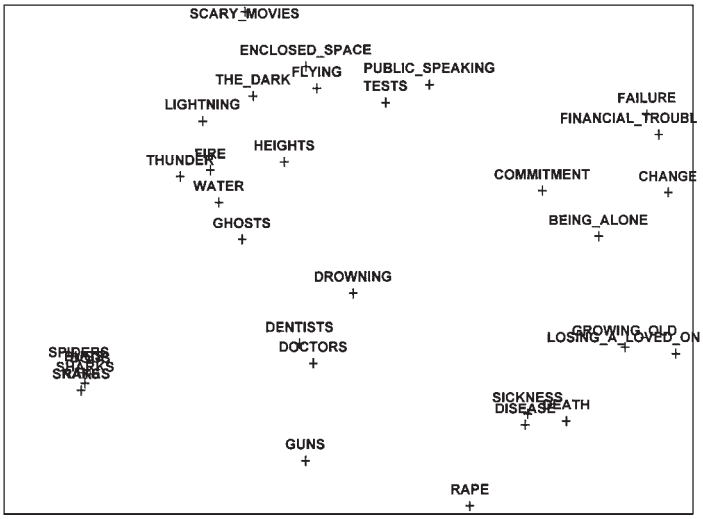
One way to answer this question is qualitative: Generate aggregate proximity matrices from each set of pile sorts, use a multidimensional scaling (MDS) program to represent each proximity matrix in Euclidean space, and compare those representations visually. In addition, to enhance their comparability, we might rotate one picture relative to the other, stretching and contracting axes as needed. Figure 1 gives the results of this approach. Figure 1a gives the women's map, and Figure 1b gives the men's map, rotated and stretched to maximally resemble the women's.[1] Yet another variation would be to stack the two proximity matrices on top of each other and run correspondence analysis, plotting just the row scores. This effectively plots both proximity matrices in the same space, albeit using a slightly different model than ordinary MDS. Figure 2 gives the results of this approach.

Visual inspection of Figures 1 and 2 suggests broad agreement between the men and women, with just a few items in different locations. Whereas the men place "rape" very close to "sickness" and "disease" and toward "death," "growing old," and "losing a loved one" (a cluster we might gloss as serious mishaps that just happen), the women place "rape" a little closer to "guns," "dentists," and "doctors" (perhaps glossed as human or human-directed dangers). Similarly, whereas men place "tests" and "public speaking" with "failure," "financial trouble," "commitment," and "being alone" (sociopsychological things), women move them closer to phobias ("the dark," "flying," "public spaces," and "heights") and natural and supernatural disasters ("lightning," "fire," "thunder," and "ghosts").

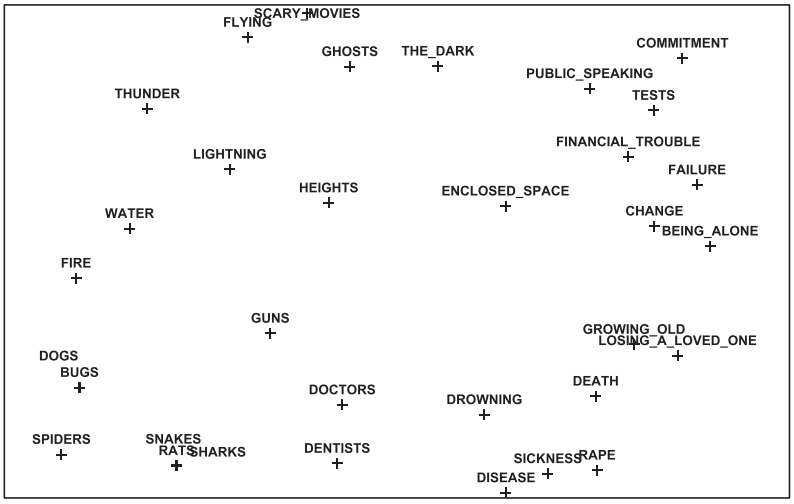
While the qualitative approach can yield important insights into the ways that groups differ in their perceptions, we are typically also interested in measuring the extent and significance of similarity or difference. The standard method for assessing the extent of similarity or difference between two proximity matrices is QAP (Hubert and Schultz 1976). The method involves cor-

FIGURE 1

a. Multidimensional Scaling of "Things People Fear"Domain for Female Undergraduates



b. Multidimensional Scaling of "Things People Fear" Domain for Male Undergraduates (rotated to maximally resemble the female scaling)



relating the two matrices and then using a permutation test to evaluate the likelihood of obtaining the resulting correlation by chance alone.

A QAP analysis of the men's and women's aggregate proximity matrices gives a correlation of 0.823, which is significant at $p < .001$ for ten thousand permutations. Thus, we conclude that despite one or two differences in detail, men and women largely see the domain quite similarly (as indicated by the correlation), and the likelihood that these matrices are independent is quite low (as indicated by the significance level).

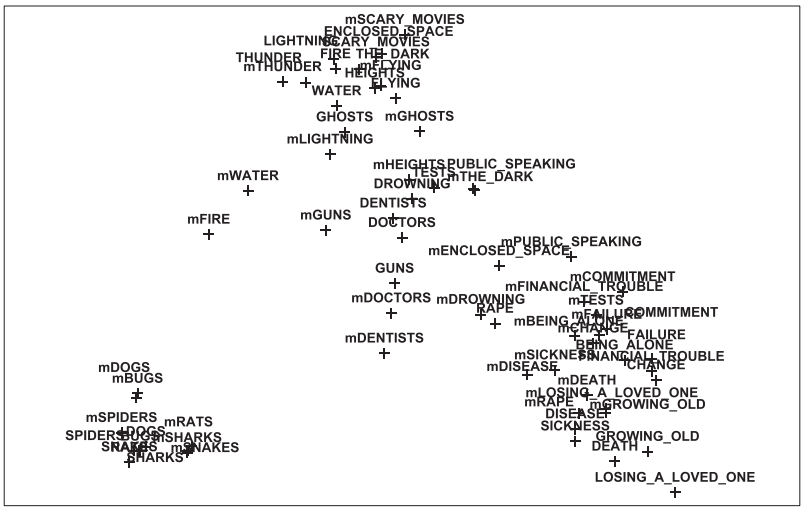
Here is another illustration. Karen Anderson, a student of John Gatewood's at Lehigh University, collected pile sort data from fifty undergraduates on the domain of animals. In addition, based on two biological classification systems, Gatewood generated two pile sorts representing the biological classification schemes. Once again, the question was simple: Do the pile sorts generated by the undergraduates resemble the scientific pile sorts?

Figures 3 and 4 give the MDS and correspondence analysis maps. This time, the differences between the two groups are a little more pronounced. For example, undergraduates group animals primarily by where the animals are found: land, water, and sky. Consequently, starfish and whales are placed near each and not far from frogs and salamanders. The scientific classifications, in contrast, place whales with other mammals, and put starfish far from everything else.

A QAP analysis of the two aggregate proximity matrices gives a correlation of 0.553 ($p < .001$ for ten thousand permutations). Thus, despite our qualitative impression that the pile sorts showed important differences, the QAP analysis indicates that overall the proximity matrices are quite similar (as indicated by the correlation) and that this is unlikely to have occurred by chance (as indicated by the p value). We would normally respond to this result by speculating about the dependence between these groups—diffusion, common perceptual processes, and so forth.

These results are not unusual: In my experience, virtually every pair of aggregate proximity matrices derived by splitting a sample of pile sorts according to some attribute of the informants is significantly correlated based on the QAP procedure. For example, Boster and Johnson (1989) computed forty-five separate QAP correlations among aggregate proximity matrices. All but one was significant. Johnson, Mervis, and Boster (1992) used QAP to correlate two aggregate proximity matrices and found them significantly similar. Weller (1984) correlated two aggregate proximity matrices and found them significantly related. Boster (1987) computed thirty different QAP correlations among aggregate proximity matrices and found every single one of them significant. An interesting case was provided by Garro

FIGURE 2
Correspondence Analysis of Stacked Proximity Matrices: Prefix of "m" Indicates Males, Absence of Prefix Indicates Females



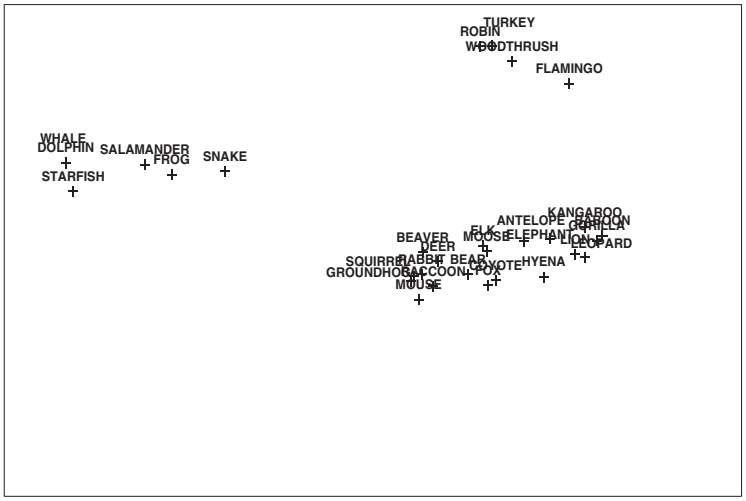
(1986). She compared two similarity matrices, one constructed from the ten informants who had the highest agreement with the sample as a whole (called the most reliable) and one constructed from the ten informants who had the lowest agreement with the sample as a whole (called the least reliable). Garro wrote:

> In the multidimensional scaling plot, the most reliable are those closest to the center, and the least reliable ones are scattered around the edges. . . . However if two new data matrices are created, one for the reliable group and one for the less reliable group . . . and the quadratic assignment program [QAP] is run again, the resulting z-score of 10.45 ($p < 0.01$) indicates that the pattern of responses is highly similar for the two groups. . . . Thus, the less reliable and the more reliable subgroups are at once both different and the same. (P. 364)

Garro seemed to be wrestling with a situation where the matrices should have been dissimilar—in fact, were dissimilar—yet QAP said they were more similar than we would expect by chance.

FIGURE 3

a. Multidimensional Scaling of Animal Domain among Undergraduates



b. Multidimensional Scaling of Animal Domain Based on Biological Taxonomic Systems (rotated to resemble the undergraduate picture)

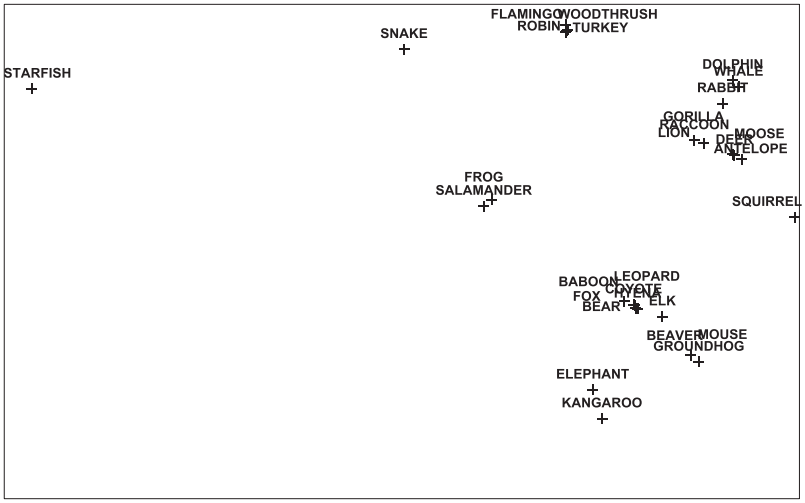
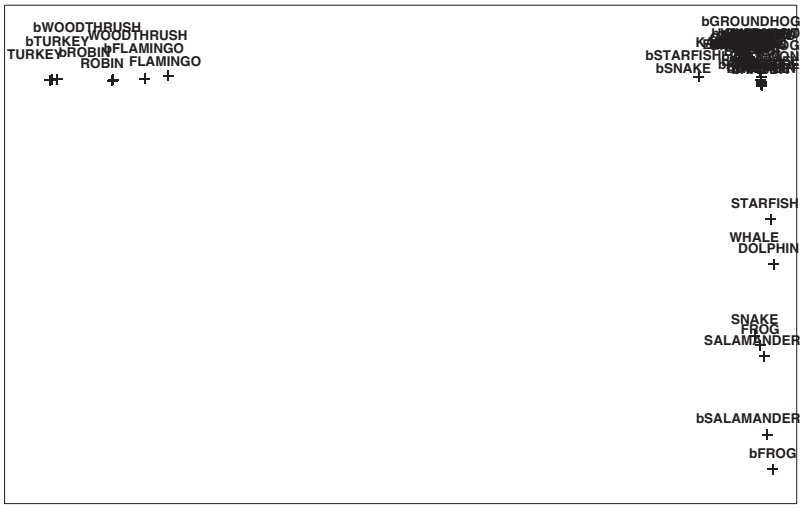


FIGURE 4
Correspondence Analysis of Stacked Proximity Matrices:
Prefix "b" Indicates Biological Taxonomic Systems



If representative, this body of experience raises questions about the validity or applicability of QAP methodology for the special purpose of comparing aggregate proximity matrices. The possibility that QAP may not be appropriate for this special case should not alarm QAP enthusiasts: All methods have their boundaries, and various limitations of QAP have been discussed before (Faust and Romney 1985; Krackhardt 1986). In the next section, I explore the assumptions of QAP in relation to comparing aggregate proximity matrices and then suggest an alternative methodology that I believe is more appropriate for this task.

QAP METHODOLOGY

QAP is a special case of randomization tests, a powerful class of statistical tests going back to Fisher (1934) that have enjoyed a sharp increase in popularity since the spread of very fast computing (Noreen 1989). The generic

model of randomization tests is simple. Suppose we want to compare two means, as in comparing test results for boys and girls. We start by computing the mean for each group and then subtract one from the other. This difference in means is our test statistic, and we refer to the actual difference between means among boys and girls as the observed value of the test statistic.

Now we generate a distribution of differences in means that would be obtained if test scores were wholly independent of gender. That is, we determine what kinds of results we could obtain (and with what probability) if test scores were assigned without regard to gender. To generate this distribution, we randomly reassign the scores to individuals, without regard to gender (or, to keep it simple, to any other characteristic). We then recompute the difference in means between the genders and store this difference away. This random reassignment and recomputation of differences is repeated tens of thousands of times,[2] generating a distribution of differences in means. We then count the proportion of these differences that are as large as the observed value of the test statistic.

This proportion or *p* value is literally the significance of the test and indicates the probability of obtaining a difference as large as actually observed, given that gender and test ability are actually independent. By convention, if the probability is less than .05, we regard the result as significant and conclude that gender may in fact be a cause of test performance (or, at least, we cannot rule it out).

In comparing two proximity matrices, the usual test statistic is the Pearson correlation between corresponding cells of the two matrices. The distribution of the test statistic under the null hypothesis of independence is obtained by randomly permuting the rows (and corresponding columns) of one matrix, then recomputing the correlation, and repeating this process thousands of times (or, for small problems, all possible permutations). The result is that the QAP test compares the observed correlation with the distribution of correlations of random matrices that contain the same collection of values that characterize the observed proximity matrices.

Furthermore, by permuting rows and columns rather than individual cells in the matrices, QAP preserves any hidden dependencies among the cells of the matrices. For example, the observed proximity matrices may be subject to transitivities such that if items *a* and *b* are very similar, and *b* and *c* are very similar, then *a* and *c* must be fairly similar as well. That is, if $s(a, b) > k$ and $s(b, c) > k$, then $s(a, c) > m$, where m is not much less than k. If both proximity matrices were generated by processes that imposed such transitivity, it would inflate the correlation between them. Thus, comparing that correlation with the correlations of random matrices without such transitivities would always result in *p* values that were, in a certain sense, too low. QAP avoids this prob-

lem by taking permutations of the rows and columns rather than individual cells, which preserves all such transitivities. So, QAP would seem rather appropriate for comparing aggregate proximity matrices.

Yet, experience suggests that virtually every QAP correlation between two aggregate proximity matrices (such as men's and women's) is significant. At first glance, this may seem an odd thing to be concerned about. First, if the correlation is high, then it is high. That is all there is to it. The correlation measures the extent of similarity. Second, if it is a high correlation, why should we be surprised if it is also significant? Large correlations don't occur by chance! This would seem obvious, but, in fact, there is a problem here that has to do with both the fact that we are dealing with aggregations and with the implicit null hypothesis.

Consider the case of perceived similarities of items in a well-established cultural domain such as animals. For the sake of argument, let us grant that all members of a given culture will categorize animals in grossly similar ways. In fact, for any given pair of animals, such as worm and elephant, we might find that no respondent places them in the same pile (similarity score of zero for each individual). Hence, the aggregate proximity matrix contains a zero for that pair of animals. In the QAP procedure, many (probably most) permutations will place nonzero similarity scores in that cell, effectively constructing a matrix that simply does not occur empirically. This means that the distribution of correlations for the null hypothesis includes correlations with matrices that are unlike anything in the observed data. Consequently, actual correlations among proximity matrices aggregated from human perceptions are nearly always higher than these unnatural matrices, and this tends to produce statistically significant results. Ideally, we would want some way of generating the null distribution from the kinds of matrices that we actually observe empirically.

Let me approach this same idea from another direction. Suppose we construct separate proximity matrices for boys and girls and observe some sociologically meaningful differences. The correlation is very modest, but the QAP significance test shows that the boys' and girls' matrices are significantly similar because it is comparing their similarity against the similarity of random matrices. The test does not take into account that, given the overall similarity of all the individual matrices, the correlation between the average boy's matrix and the average girl's matrix may actually be surprisingly small. That is, when we divide the respondents into two groups according to other criteria, such as hair color, height, religion, or simply randomly, the similarity between the resulting aggregate matrices tends to be just as high as when we do it by gender. So gender is not special: The speciously high correlation we observe would have been obtained by virtually any division into

groups, and for the purpose of understanding the relation to gender, we would like this correlation not to be significant.

In other words, QAP tests a different hypothesis than the one we are actually interested in when comparing aggregate matrices—that is, it compares the observed correlation against barely constrained random matrices, and this is the wrong baseline distribution for our purposes. This is not a problem with QAP itself but with our use of it. That QAP should find that two aggregations of a given data set are significantly correlated is to be expected, as these are clearly not independent. For instance, if we consider the cognitive domain of animals, it would not surprise us that human beings the world over see the domain very similarly (Boster 1987). Humans, if not primates in general, share essentially the same perceptual and cognitive apparatuses. Their observations are not really independent. Hence, QAP correctly returns a result of nonindependence. It is simply the wrong technique for the research question.

However, before moving to another approach, we should address a possible confusion. When we compare boys' and girls' aggregate matrices, it is ordinarily because we think gender might make a difference in how people see the world. The null hypothesis is one of sameness between the genders; if this is rejected, we say that gender makes a difference in how people perceive the domain in question. But the QAP procedure is normally thought of as a correlation, with the null hypothesis being a lack of similarity between the matrices. Is this the root of our problem? No. The QAP procedure is entirely about significance and is perfectly general with respect to the measure of association used. We often use correlation today, but QAP was originally formulated using other measures (Mantel 1967; Hubert and Schultz 1976). Hence, in place of the correlation coefficient, we can just as well use a measure of dissimilarity, such as one minus the correlation, or Euclidean distance. Doing this does not change any of the results under discussion here. For example, when applied to the animal data, a one-tailed QAP test (as they always are) of the null hypothesis of no difference (by a variety of standard dissimilarity measures) between matrices produced by students and biologists yields a p value of approximately 1.000. (If we like, we can think of it as a significant p value for a test of the alternative hypothesis.) In other words, just as before, QAP rejects the hypothesis of difference.

A NEW APPROACH

The approach I propose is simple. To compare the boys' aggregate proximity matrix with the girls' aggregate matrix, we begin by correlating the two

matrices (or computing a dissimilarity measure). This is our observed test statistic. Then we go back to the individual-level data and divide the respondents into two groups at random. We then aggregate the matrices separately for each group, obtaining an aggregate proximity matrix for each group. Next, we correlate these matrices (or compute dissimilarity measure) and store the result. This process is repeated thousands of times to generate a distribution of (dis)similarities under the null hypothesis of independence (i.e., judged proximities are independent of gender). We then count the proportion of correlations (or dissimilarity measures) that are as small (or as large) as the observed measure. The question is whether any random division into two groups would show as much difference as did our boys and girls. The proportion of correlations as small as the observed (or, equally, the proportion of dissimilarity coefficients as large as the observed) gives the *p* value: the likelihood that the difference we see could be obtained by chance, that is to say, without regard for gender.

A key difference between the new method and the QAP technique is that the new method utilizes the individual-level data matrices along with a vector indicating to which group each individual belongs. In contrast, QAP is applied to the two aggregate proximity matrices and does not require (and has no way to take account of) individual-level data.

A more important difference lies in the nature of the null hypothesis being tested—that is, the reference distribution that the observed similarity measure is being compared to. The QAP procedure asks whether the two aggregate matrices are independent of each other. It compares the observed correlation to correlations among nearly all matrices that have the same collection of values and intercell dependencies. Where a set of homogeneous (e.g., all human) respondents is asked to respond to the same stimuli, the results are a foregone conclusion: There is a basic commonality—a lack of independence—among all of the matrices due to the shared perceptual apparatuses, not to mention culture. And aggregating subsets of respondent matrices before correlating can sharply accentuate that similarity.

In contrast, the method presented here tests a different hypothesis. This method asks whether the (dis)similarity between two aggregate matrices is independent of a respondent attribute. The observed correlation is compared to a reference distribution of correlations among a set of aggregate matrices that are created using the same process as the original matrices, automatically excluding impossible matrices. The reference distribution here is a conditional one: We compare the observed aggregate matrices with the set of all aggregate matrices that can be constructed from the collected set of individual matrices. Thus, we take as given the basic similarity among all the individual matrices and do not allow that to trigger a significant result.[3]

Let us now apply the new method to the Anderson and Gatewood data set. Let us take no difference between undergrads and biologists as the null hypothesis and use one minus the correlation coefficient as a measure of dissimilarity. The observed dissimilarity is 0.447. The p value is .034, which is significant by conventional standards. In other words, randomly splitting the fifty-two respondents into any two groups of identical size as our treatment groups (undergrads and biologists) is unlikely to produce a dissimilarity between the aggregate proximity matrices as high as 0.447. So we conclude, contrary to the QAP analysis, that occupational status is related to perception of animals.

Applied to the fears data collected by Pirelli et al., the new method (again using one minus the correlation coefficient to measure association) yields a p value of .062—borderline significant by most standards. We interpret the result as saying that it is somewhat unlikely (6% chance) that you could obtain such a large difference by chance alone. Had the probability been just a little lower (5%), we would conventionally feel comfortable rejecting the null hypothesis of no difference. Given the actual result, we would note that in this case the results do not clearly contradict the QAP results, which was unequivocal in accepting the null hypothesis of no difference.

To illustrate the method in another setting, we turn to an example using social network analysis. Krackhardt (1987) asked all twenty-one managers of a high-tech firm to indicate who was friends with whom among all twenty-one. This results in a data set consisting of twenty-one different 21-by-21 matrices. To obtain an aggregate view of the network, one can simply average the twenty-one matrices. The resulting values could be interpreted as a strength of tie: If manager a was a close friend of manager b, many people would be likely to have noticed and report it, while if a's and b's relationship were not very close, few people would be likely to report a friendship tie between them.

The data set also contains some information about each manager, including their age. I divided the respondents into two groups based on age: older than the mean (39.7 years) and younger than the mean. Then I created separate aggregate matrices for each group and correlated them via QAP. The correlation was 0.704, significant at the $p < .0001$ level with ten thousand permutations. The Euclidean distance was significant at the $p \approx 1.00$ level, which is to say significant in the other direction, the direction of no difference. Obviously, these results indicate very high agreement, and the significance level all but rules out the possibility of independence.

But why should we expect independence? The managers and the employees work together, interacting on a daily basis. Every pair of matrices is bound to be similar, and comparing their correlation to correlations among

matrices that could not be obtained by any aggregation scheme doesn't seem very useful. The more appropriate test is the one presented in this article. The results are in marked contrast to the QAP test. Using one minus the correlation as the measure of dissimilarity, the dissimilarity between aggregate matrices was significant at the 0.036 level. This indicates that aggregating by age class yields matrices that are much less similar than we expect by chance, given the overall pattern of similarity. Again, I stress that the results contrast with QAP not because there is something wrong with QAP, but because QAP is not trying to answer the same question we are asking—it is a mistake to use QAP to compare aggregate matrices.

BEYOND PROXIMITIES

Although my motivation for writing this article was the comparison of aggregate proximity matrices, there is nothing in the method that limits its application to proximities. Whereas QAP is limited to comparing whole matrices, in the method presented here the data could just as well consist of a vector of values, such as the responses of two groups of respondents to any set of survey questions.

A particularly interesting application is the comparison of word frequencies in texts. For example, Jang and Barnett (1994) obtained a matched sample of annual reports from Japanese and American companies and looked at word counts in the letter from the CEO. A basic question they asked was whether Japanese and American companies used different words in their reports. Jang and Barnett used correspondence analysis to see patterns in the data and also used discriminant analysis to identify a set of words whose frequency helped predict whether the text in which they were found was associated with an American or Japanese company. The method presented here gives us a statistical test to determine whether the differences in frequencies of words could have (with reasonably probability) arisen by chance alone.

Table 1 shows the relative frequencies of the fifty-eight most frequently used words (not including articles such as *the* and connectors such as *and*) by national origin. The correlation between the U.S. and Japanese columns is 0.319, and the Euclidean distance is 1.949. The correlation may seem high (and the Euclidean distance low), but it turns out that the probabilities of obtaining such a low correlation and such a high Euclidean distance are both smaller than 1/10,000. In other words, given the high level of similarity between all the texts, it is (statistically) surprising how different the Japanese and American texts are. That is, there is a cultural effect.

TABLE 1
Relative Frequencies of Words in Annual Reports, by Nation

Word	*Overall*	*United States*	*Japan*
Business	94	100	88
Products	94	94	94
New	91	94	88
Growth	86	83	88
President	83	67	100
Sales	83	72	94
These	83	72	94
Product	83	83	82
Global	80	78	82
Which	80	78	82
Continue	80	89	71
During	77	72	82
States	77	78	76
Market	77	72	82
United	77	78	76
Net	77	67	88
Company	77	83	71
Results	74	72	76
Continued	74	67	82
World	74	72	76
Chairman	74	94	53
Billion	74	83	65
Economic	71	67	76
Their	71	83	59
Operating	71	78	65
Well	71	94	47
Performance	71	83	59
Years	71	83	59
Time	71	83	59
All	71	78	65
Markets	71	67	76
Financial	71	83	59
Support	69	61	76
Strong	69	83	53
Corporate	69	61	76
Share	69	78	59
Customers	69	78	59
U.S.A.	69	83	53
Operations	69	56	82
Management	66	78	53
Environment	66	44	88
Businesses	66	89	41
One	66	83	47

TABLE 1 Continued

Word	*Overall*	*United States*	*Japan*
Million	66	67	65
Industry	63	72	53
Income	63	39	88
Customer	63	72	53
Term	63	56	71
Major	63	72	53
Development	63	56	71
Japan	63	33	94
High	60	50	71
Economy	60	50	71
Further	60	61	59
Up	60	56	65
Despite	60	56	65
Worldwide	60	72	47
Long	60	61	59

NOTE: Limited to words occurring in at least 60% of texts and not including articles and connector words.

A similar application of the method is to the comparison of free-list data. By free-list data, I mean responses to questions of the type "Tell me all the kinds of ____ you can think of," where "____" is a cognitive domain such as animals, illnesses, or occupations. For example, I asked approximately one hundred undergraduate students to list all the "vacation destinations" they could think of. A total of 352 locations were given. Table 2 gives the relative frequencies of destinations named by 10% or more of the sample.

The Pearson correlation between the frequencies of the males and those of the females is 0.846. The probability of obtaining a correlation as low as that (given that gender is independent of word recall) is approximately 0.49. Thus, the hypothesis of gender difference in free listing of vacation destinations is not supported: The frequencies of one gender cannot be statistically distinguished from the other. It is important to realize that this indistinguishability cannot be deduced from the high correlation alone: In another domain (that of "bad words"), the correlation between boys' and girls' frequencies was 0.882, but the p value was a significant .02, while sample size and domain size were virtually the same as in the vacation study. The difference in p values is due to differences in the overall level of agreement among respondents in each study. The bad words domain is a high concordance domain in which even fairly small differences in frequency between two groups are unlikely to occur by chance. In contrast, the vacation destinations

TABLE 2
Vacation Destinations

Destination	*Girls*	*Boys*
Hawaii	0.68	0.76
Bahamas	0.45	0.63
Cancun	0.53	0.52
Jamaica	0.42	0.52
California	0.42	0.48
Florida	0.45	0.45
Paris	0.34	0.47
Australia	0.39	0.40
Bermuda	0.37	0.34
London	0.39	0.31
Disney World	0.24	0.29
Puerto Rico	0.16	0.32
Italy	0.13	0.32
France	0.18	0.27
Spain	0.13	0.31
Miami	0.29	0.21
New York	0.26	0.21
Rome	0.18	0.26
San Francisco	0.18	0.23
New York City	0.16	0.23
Los Angeles	0.21	0.19
Mexico	0.21	0.18
Egypt	0.11	0.24
Grand Canyon	0.13	0.23
Las Vegas	0.18	0.18
Canada	0.16	0.18
Caribbean	0.13	0.19
Aruba	0.13	0.19
Colorado	0.18	0.16
Cape Cod	0.16	0.18
New Orleans	0.18	0.15
Virgin Islands	0.21	0.13
Montreal	0.16	0.16
Chicago	0.18	0.13
Ireland	0.21	0.11
Alaska	0.16	0.15
Maine	0.16	0.13
Japan	0.13	0.15
Europe	0.16	0.13
D.C.	0.24	0.08
Amsterdam	0.18	0.10
Boston	0.13	0.13
Orlando	0.13	0.13

TABLE 2 Continued

Destination	*Girls*	*Boys*
China	0.11	0.13
Disneyland	0.13	0.11
Germany	0.11	0.13
San Diego	0.16	0.10
Africa	0.05	0.16
Florence	0.08	0.13
New Zealand	0.16	0.08
England	0.03	0.16
Venice	0.08	0.13
Cayman Islands	0.13	0.10
Vermont	0.05	0.15
Brazil	0.08	0.13
Hong Kong	0.16	0.08
St. Thomas	0.13	0.08

NOTE: Values are sample proportions.

domain is a low concordance domain; hence, larger differences between groups can be obtained by chance alone.

The method can be extended for use with more complicated kinds of aggregation as well. For example, consider the case of a multiple-choice test underlying the theory of cultural consensus developed by Romney, Weller, and Batchelder (1986). In their approach, a multiple-choice test is given in which the answer key is unknown. They use agreement among respondents to estimate the amount of knowledge they have, and these estimates are in turn used to guess the answer key. A simple procedure, among many alternatives, for estimating the answer key is to take the modal response for each question, where respondents are weighted by their knowledge. Given the possibility of both individual differences and the existence of subcultures that may have different culturally correct answer keys for the same set of questions, we can ask whether a given a priori division of the respondents into two groups reveals systematic differences between them.

Using the techniques described in this article, it is a simple matter to test the similarity or difference between the answer keys for each group. Essentially, we calculate the answer key for each group, correlate, then randomly reassign respondents to groups, reestimate the answer keys, correlate, and repeat a few thousand times. In other words, as long as we use the same procedure at all times for constructing an aggregate answer key from a set of respondents, it doesn't matter for the purposes of this test how exactly the aggregation is performed: Any aggregation method can be used.

CONCLUSION

The objective of this article has been to propose a new method for statistically comparing pairs of aggregate data series. The motivating problem was the comparison of aggregate proximity matrices, such as obtained from pile sort exercises. This problem is central to cognitive research in which we seek to compare the perceptual maps of different groups. The standard approach to this problem uses QAP. However, the null distribution that QAP is based on is not appropriate for correlating subsamples of a data set. QAP explicitly compares the correlations between subsamples with correlations among a set of matrices that include many that are quite different from the observed. Consequently, when applied to subsamples, the QAP approach can achieve a significant p value too often. In contrast, the method proposed here compares the observed correlation with the correlations among matrices that are of the same kind as the observed.

It should be noted that the method presented here is limited to comparison between two groups. There are clearly ways to generalize to the case of multiple groups, but I leave that as an avenue for future research.

NOTES

1. A minor variation on this approach would be to run the INDSCAL model (Carroll 1972).

2. In principle, we would like to repeat for all possible reassignments, but this is only feasible for small numbers of respondents. Instead, we sample randomly from the space of all possible reassignments.

3. The situation is analogous to that of using a classical significance test on data that are not drawn from a random sample (or in any other way violate the assumptions of the test). If you obtain a significant result, there is no way to know whether it is because the variables are not independent or because the sample is not random. This point is made very clearly by Noreen (1989).

REFERENCES

Boster, J. 1987. Agreement between biological classification systems is not dependent on cultural transmission. *American Anthropologist* 89:914–20.

Boster, J., and J. C. Johnson. 1989. Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist* 91:866–89.

Carroll, J. D. 1972. Individual differences and multidimensional scaling. In *Multidimensional scaling: Theory and applications in the behavioral sciences. Vol. 1. Theory*, edited by R. N. Shepard, A. K. Romney, and S. Nerlove, 105–55. New York: Seminar.

Edgington, E. S. 1969. Approximate randomization tests. *Journal of Psychology* 72:143–49.

Faust, K., and A. K. Romney. 1985. The effect of skewed distributions on matrix permutation tests. *British Journal of Mathematical and Statistical Psychology* 38:152–60.

Fisher, R. A. 1934. The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society A* 98:39–54.

Garro, L. C. 1986. Intracultural variation in folk medical knowledge: A comparison between curers and noncurers. *American Anthropologist* 88:351–70.

Good, P. 1994. *Permutation tests: A practical guide to resampling methods for testing hypotheses.* New York: Springer-Verlag.

Hubert, L. J., and J. Schultz. 1976. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology* 29:190–241.

Jang, H., and G. A. Barnett. 1994. Cultural differences in organizational communication: A semantic networks analysis. *Bulletin de Methodologie Sociologique* 44:31–50.

Johnson, K. E., C. B. Mervis, and J. S. Boster. 1992. Development changes in the structure of the mammal domain. *Developmental Psychology* 28:74–83.

Krackhardt, D. 1986. A caveat on the use of the quadratic assignment procedure. Unpublished manuscript.

———. 1987. Cognitive social structures. *Social Networks* 9:104–34.

Mantel, N. 1967. The detection of disease clustering and a general regression approach. *Cancer Research* 27 (2): 209–20.

Noreen, E. 1989. *Computer intensive methods for testing hypotheses: An introduction*. New York: Wiley.

Romney, A. K., S. Weller, and W. Batchelder. 1986. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist* 88 (2): 313–38.

Weller, S. C. 1984. Cross-cultural concepts of illness: Variation and validation. *American Anthropologist* 86:341–50.

Weller, S. C., and A. K. Romney. 1988. *Systematic data collection*. Beverly Hills, CA: Sage.

STEPHEN P. BORGATTI is an associate professor of organizational behavior at the Carroll School of Management, Boston College. His research focuses on social networks and cognitive domains, most recently with application to knowledge management in organizations (in collaboration with IBM's Institute for Knowledge Management). Recent publications include the following: "Beyond Answers: Dimensions of the Advice Network" (by R. Cross, S. P. Borgatti, and A. Parker, Social Networks*, 2001) and "Models of Core/Periphery Structures" (by S. P. Borgatti and M. G. Everett,* Social Networks*, 1999).*