Improving the Measurement of Shared Cultural Schemas with Correlational Class Analysis *

Andrei Boutyline

University of California, Berkeley

WORKING PAPER

Please do not cite without permission.

August 2, 2014

Keywords: survey data, network analysis, correlation networks, cultural measurement.

* This research was supported by a fellowship from the National Science Foundation Graduate Research Fellowship Program. I thank Neil Fligstein, Amir Goldberg, Monica Lee, Fabiana Silva, Stephen Vaisey and Robb Willer for feedback on the paper. I am also grateful to Amir Goldberg for generously discussing RCA and making its software implementation available online. Direct all correspondence to Andrei Boutyline at Department of Sociology, 410 Barrows Hall, University of California, Berkeley, CA 94720. Email: boutyline@berkeley.edu

ABSTRACT

The task of measuring shared cultural schemas is among the most central methodological challenges in sociology of culture. Relational Class Analysis (RCA) is a recently developed technique for identifying such schemas in survey data. It uses a novel measure called "relationality" to quantify the extent to which two respondents organize their attitudes according to a shared cultural schema. However, since existing work lacks a formal definition of such schemas, RCA's accuracy is difficult to assess. In this paper, I build on the theoretical reasoning behind RCA to arrive at this formal definition. In doing so, I discover that shared cultural schemas should result in linear dependencies between rows in a survey dataset—the same relationship as measured by Pearson's correlation. I thus compare relationality and correlation across a broad range of simulations. The results indicate that switching from "Relational" to "Correlational" Class Analysis (CCA) reliably increases accuracy, so much so that, when the methods disagree, the odds that CCA's results are more accurate exceed 16:1. I then revisit a prior RCA analysis of the 1992 GSS musical tastes module with CCA, and find that the more accurate method contradicts the controversial RCA finding that a substantial portion of respondents organize their tastes according to the traditional distinction between "highbrow" and "lowbrow" genres. Instead, CCA documents two classes of respondents whose tastes echo the exclusionary omnivorousness described by Bryson (1996), in which a handful of musical tastes closely associated with low-education listeners are distinguished from tastes for all other genres. I conclude by highlighting further areas for methodological improvement.

Improving the Measurement of Shared Cultural Schemas with Correlational Class Analysis

The task of revealing intelligible structures of meaning beneath complex collections of cultural data is among the most central methodological challenges posed by sociology of culture (Mohr and Rawlings 2012; Mohr 1998). From the perspective of culture and cognition (DiMaggio 1997, 2011), this task is a search for "cultural schemas"—abstract shared structures which specify relationships between cultural elements. In a high-profile recent work, Goldberg (2011) proposes an innovative methodology for identifying groups of survey respondents who share such cultural schemas, which he terms Relational Class Analysis (RCA). RCA has garnered a substantial amount of attention across diverse domains of study including cultural tastes (Goldberg 2011), public opinion (Goldberg and Baldassarri forthcoming; Wu forthcoming), organizational behavior (Miranda, Summers, and Kim 2012) and economic sociology (DiMaggio and Goldberg 2010). However, existing work has not yet provided a clear formal definition of cultural schemas, the central concept under investigation. This is a crucial limitation as, without such a definition, it is impossible to quantify how well RCA measures what it purports to measure. In this paper, I build on Goldberg's (2011) theoretical reasoning to arrive at this missing definition, which I then use to develop a substantially more accurate, computationally faster, and analytically more parsimonious alternative I call Correlational Class Analysis (CCA).

Goldberg (2011) introduces RCA using a case study of musical tastes. Taking a cue from relational theories of meaning (e.g., Mohr 1998; Saussure 1916 [2013];

Emirbayer 1997), RCA searches for cultural schemas that define not the musical tastes themselves, but the relationships between these tastes—that is, which genre tastes are perceived as similar and which as opposed. This kind of schema can be found in the implicit agreement between an individual who likes musical genres A and B but dislikes genre C, and another who dislikes A and B but likes C: though the two hold no tastes in common, they nonetheless agree that A "goes with" B, whereas C is "opposed to" A and B. On the other hand, an individual who likes all three of the genres A, B and C has two tastes in common with the first individual, but does not agree that C is the opposite of A and B. Thus, under this definition, the first two individuals arrange their tastes according to exactly the same cultural schema, while the third one does not.[1] The goal of RCA is to partition the survey population into classes of respondents that share such cultural schemas.

At the core of RCA's approach is a novel similarity measure termed "relationality," which Goldberg contends can quantify the extent to which two respondents organize their attitudes according to such a shared cultural schema. RCA calculates the relationality scores for each pair of respondents and interprets the result as an adjacency matrix for a valued network, where the nodes are individual survey respondents and ties between them are their pairwise relationalities. After adjusting tie strengths for a bias in the relationality measure and dropping weaker ties, it partitions the network via Newman's (2006) eigenvector-based modularity maximization algorithm, which assigns individuals to groups so that relationality scores are high within groups, and lower between groups.

Goldberg (2011) developed relationality as a tool to detect shared cultural schemas in survey data—a novel methodological task that is, in itself, a bold conceptual innovation. However, existing work on RCA does not provide a formal definition of shared cultural schemas, making relationality's performance difficult to analyze. As I demonstrate below, however, the theoretical reasoning behind RCA easily yields such a definition. Moreover, this formalization reveals that, in order to detect shared cultural schemas, relationality must measure the degree of linear dependency between two vectors of responses—the same quantity for which Pearson's correlation is a long-established measure. When I revisit Goldberg's introductory example (Goldberg 2011:1404-1405) with Pearson's correlation, I discover that it produces a substantially more accurate result than Goldberg's relationality.

To verify that this difference in accuracy generalizes beyond this example, I simulate 10,000 further test cases. The results confirm that this switch from "Relational" to "Correlational" Class Analysis (CCA) reliably increases the accuracy of the technique, so much so that, when the two disagree, the odds that CCA's result is more accurate exceed 16:1. RCA's low relative accuracy degrades to 23:1 when the simulated data violate a strong distributional assumption introduced by relationality. To examine the substantive consequences of switching to the more accurate correlation-based method, I also revisit Goldberg's (2011) RCA analysis of the 1993 GSS musical tastes data, and find that CCA results differ in substantive ways from RCA's. I conclude by discussing the remaining limitations of the method, and by highlighting further areas for improvement.

SHARED SCHEMAS

Although Goldberg (2011) does not formally define what it means for a set of respondents to share a cultural schema, he illustrates such a relationship by way of an example and accompanying diagram, which I recreate as Figure 1 below. Describing this figure, Goldberg states that A and B have "identical" patterns of musical tastes, C's pattern is "almost a mirror image" of A's and B's, and D's is "different but not antithetical." He thus concludes that "respondents A, B, and C exhibit the same logic of musical taste construction ... as they all exhibit the same structure of relevance and opposition" (Goldberg 2011:1405), whereas respondent D does not.

To arrive at a formal definition of schematic similarity, I expand Goldberg's discussion of this introductory example by writing out the implied algebraic operations. Respondent A likes pop, blues and rock, strongly likes classical and opera, and is indifferent towards bluegrass and country: $A = \{4,4,4,5,5,3,3\}$. Respondent B, on the other hand, dislikes pop, blues and rock, is indifferent towards classical and opera, and strongly dislikes bluegrass and country: $B = \{2,2,2,3,3,1,1\}$. Except for an overall downward shift in the appraisal of all the genres, this pattern of tastes is identical to that of the first respondent: $B = A - 2$.

[Figure 1 about here]

In contrast to A and B, respondent C is indifferent towards pop, blues and rock, strongly dislikes classical and opera, and strongly likes bluegrass and country: $C = \{3,3,3,1,1,5,5\}$. These tastes again follow the same relative pattern as A and B, except all tastes are vertically shifted, inverted and amplified: $C = 2 * (-1) * A + 11$, or,

equivalently, $C = 2 * (-1) * B + 7$. Finally, respondent D strongly dislikes pop and rock, strongly likes blues, likes classical, opera and bluegrass, and dislikes country: $D = \{1,5,1,4,4,4,2\}$. Unlike A, B and C, this respondent construes an opposition between bluegrass and country, but not between bluegrass and opera. No series of inversions, multiplications or shifts of this pattern can transform it into the one exhibited by A, B and C. We thus conclude that, while respondents A, B and C follow the same schema, respondent D does not.

From this example, we can surmise that two responds follow exactly the same schema if (i) their attitudes are identical, (ii) their attitudes are exact inverses of each other's, (iii) the attitudes of either respondent are uniformly more extreme than those of the other, (iv) the attitudes of either respondent are uniformly more positive than of the other, or (v) any combination of (ii), (iii) and (iv). These conditions specify the mathematical operations of identity, inversion, scaling and vertical shift, and can thus be captured by a single algebraic statement: two respondents X and Y follow exactly the same schema *if and only if there exists a linear transformation that can produce one response pattern from the other one*, or, more formally, if there exist such constants *b* and $k \neq 0$ that $Y = kX + b$. It is therefore intuitively clear that any measure of schematic similarity between two respondents should obtain its maximum value when such *k* and *b* exist, and should otherwise capture the degree to which one pattern can be approximated by linear transformations of the other.

Relationality

Goldberg (2011) offers relationality R_{ij} as a measure of this schematic similarity. It is computed by first taking the vector of attitudes belonging to a single respondent and calculating the pairwise differences between each pair of attitudes in that vector, so that each row in the survey dataset produces a square matrix of differences between variables in that row. Then, to calculate the relationality between a pair of respondents *i* and *j*, the absolute values of their respective matrices are element-wise subtracted from each other, and each element of the resulting matrix is assigned a sign based on whether the original differences were in the same or in opposite directions. Finally, the elements of this matrix are summed into a single variable, which is then rescaled and recentered to range from 1 (same direction) to -1 (opposite direction)[2]. The distinction between positive and negative relationalities, however, is not useful for RCA, as either extreme of the measure indicates that respondents follow the same schema. Thus, in all further analyses, RCA uses only the absolute values of the relationality $|R_{ij}|$, which ranges from 1 (same schema) to 0 (unrelated schemas).

The values $|R_{ij}|$ obtains in the introductory example should thus clearly identify that A, B and C follow exactly the same schematic pattern, but D does not. However, relationality's difficulties at this task are evident in Figure 1B of the original paper (Goldberg 2011:1405). I depict the relevant parts of this diagram in Figure 2A below[3]. Relationality achieves its maximum value for the respondent pair A and B (AB = 1.00), thus clearly indicating that the two follow the same schema. Conversely, the absolute relationality between the pair A and D is approximately 0.2, which is appropriately low as A and D follow different schemas. Since C follows the same schema as A and B, the

absolute relationalities AC and BC should optimally be equal to the same value as AB (1.00). Unfortunately, this is not the case: both AC and BC have absolute relationalities of approximately 0.3, which is far closer to the relationality of the unrelated pair AD (0.2). Thus, relationality appears to grossly understate the schematic similarity between respondent C and respondents A and B.

[Figure 2 about here]

To determine whether this inaccuracy causes RCA to yield an incorrect solution, I created a dataset consisting entirely of rows A, B, C and D, each repeated 200 times for a total of 800 rows. I then analyzed it with the RCA software provided by Goldberg (Goldberg and Zhai 2013). To produce the correct solution, RCA would have to partition this population in two classes, the first containing all the copies of rows A, B and C (600 rows total), and the second all copies of D (200 rows). However, RCA instead produced an erroneous solution consisting of three distinct classes, with the copies of C incorrectly assigned to their own class, separate from copies of A and B[4]. Since Goldberg uses this example to introduce relationality as a tool for detecting shared schemas, its failure at this task is especially troubling.

Correlation

This shortcoming means that there is merit in trying out a different measure of schematic similarity. Recall that two respondents X and Y exactly follow the same schema if there exist such constants $k \neq 0$ and b such that $Y = kX + b$. Thus, provided that X and Y have a finite non-zero variance, it can be easily shown that the absolute Pearson's

correlation $|r|$ between X and Y equals 1 if and only if they follow exactly the same schema. As the two responses become more and more linearly independent of one another—that is, as the best possible linear transformation of X leaves an ever larger percentage of Y's variance unexplained—the value of $|r|$ decreases monotonically towards 0. Finally, $|r|$ will be equal to 0 if and only if $k = 0$ gives the best linear approximation of Y, or, in other words, if the best linear approximation of Y ignores the contents of X altogether. This is why Pearson's correlation is often interpreted as the "measure of the degree of linear relationship between two variables" (Stockburger 2007; see Rodgers and Nicewander 1988 for a detailed treatment). Thus, Pearson's correlation appears to be a perfect candidate for this task.

Figure 2B demonstrates the results obtained by applying Pearson's correlation to the same problem. The absolute correlations AB, AC and BC all equal 1, whereas AD, BD and CD equal 0.25. The absolute correlations between responses that follow the same schema are thus at their theoretical maximum, while the ones between members of different schematic classes are closer to their minimum. Thus, correlation appears to produce a far clearer depiction of the schematic relationships between these respondents than relationality. To examine whether this improvement results in a correct partition into classes, I implemented an algorithm for detecting schematic classes based on absolute row correlations instead of relationalities, which I term Correlational Class Analysis (CCA; see Appendix A for details). And indeed, when I applied CCA to the same 800-row dataset, it correctly assigned all the rows into the two schematic classes present in the

data. Thus, while RCA failed to correctly recover the schematic classes in Goldberg's example, CCA produced a perfect answer.

Therefore, though Goldberg asserts that relationality's far greater computational complexity makes it "more sensitive to interdependencies" between the variables than correlation (Goldberg 2011: Appendix A), the analysis thus far suggests that quite the opposite may be the case. The theory of schematic similarity implies that the quantity these measures need to capture is the degree of linear dependency between responses, which is exactly what is measured by Pearson's correlation. Thus, it is far from certain that measuring any other kind of interdependency should aid in the search for these schema. Moreover, Goldberg never actually provides an example of any schematic similarity which relationality can detect but correlation cannot. On the other hand, this introductory example presents a clear case where relationality fails to detect a schematic similarity that is perfectly detected by correlation.

SIMULATION

The above analysis suggests that CCA is a more accurate tool than RCA for detecting groups of respondents whose answers follow the same cultural schema. However, one may rightfully object that a single example does not provide a sufficient basis for drawing such a broad conclusion. To rule out the possibility that CCA's apparently superior performance is due to features specific to this introductory example, I turn to simulation to carry out a more thorough and realistic test.

The formal definition of shared schemas can serve as the basis for such simulations. Since two response vectors exactly follow the same schema if and only if they are linear transformations of one another, the schema specifying relationships between *N* tastes can itself be specified with a vector $\rho = \{\rho_1, \rho_2, \ldots, \rho_N\}$.[5] Such a schema can be randomly generated by drawing a vector of integers from an appropriate probability distribution. In turn, each response pattern X that exactly follows this schema can be generated by randomly drawing a pair of linear transformation constants *k* and *b*, which can invert, rescale or shift the pattern in ways consistent with the theory discussed earlier. However, since real survey respondents do not perfectly reproduce cultural schemas, a simulated response must also include substantial stochastic deviations from the schema. These can be introduced via an independent error vector ϵ of the same length as the original pattern, so that $X = k\rho + b + \epsilon$. This is the basic formula behind my simulations.

To ensure that the simulations cover a wide range of potential cases, each simulation run consists of three randomization steps. The first step randomizes the broad characteristics of the data to be simulated in this run, such as the ranges and variances that will be used to generate the values of ρ, k, b and ϵ, as well as the number of distinct taste schemas behind the responses. The second step generates these schemas using the variance parameters produced in the first step. Finally, the third step generates a random number of respondents following each of these schemas by applying random linear transformation and adding random noise, both generated using the ranges and variances set in the first step. I repeat the entire procedure 5000 times[6], creating simulated datasets

that widely differ in the ranges of simulated variables, variance of individual responses, signal to noise ratio, and many other parameters. A more detailed description of the simulation procedure can be found in Appendix B.

Figure 3 illustrates a single simulation run. The randomly determined parameters from the first step of the run set the schema variance to 0.51, number of schemas to 3, and the maximum error variance shift and scaling factor to 1.02, 1 and 2, respectively. Thus, to make the schemas ρ_1, ρ_2 and ρ_3, three vectors were drawn from the normal distribution with $\mu = 0$ and $\sigma^2 = 0.51$, and rounded to the nearest integer. The resulting schemas are depicted with solid black lines, one per plot. For each schema, the simulation created a set of followers by randomly picking a value of shift b from $\{-1,1\}$, scaling and inversion factor k from $\{-2,-1,1,2\}$, and a noise vector ϵ drawn from the normal distribution with variance of no more than 1.02. A small sample of such respondents is depicted in dashed gray lines behind the appropriate schema.

[Figure 3 about here]

Measuring Accuracy

This simulation thus generalizes and expands the introductory example. Each of the 5000 simulated datasets is based around a different set of taste schemas, and consists of a large population of simulated responses, each produced in a theoretically consistent way from one of those schemas. Thus, as in Goldberg's (2011) introductory example, the true schematic class membership for each simulated respondent is known by design. The simulation's goal is to assess the accuracy with which the group assignments made by the two algorithms correspond to this known membership. If two respondents were generated

from the same schema, they should be assigned to the same group; if they were created from different schemas, they should belong to different groups.

For each run, I measured this classification accuracy with Normalized Mutual Information (NMI):

$$NMI(\Omega, C) = \frac{2 * I(\Omega; C)}{H(\Omega) + H(C)},$$

where vector C contains the true class memberships for every respondent, vector Ω contains the group assignments made by the algorithm, *I* is mutual information, and *H* is Shannon entropy. NMI is an established criterion for measuring the accuracy of network partitioning algorithms (e.g., Danon et al. 2005; Lancichinetti, Fortunato, and Kertész 2009). It achieves its minimum of 0 when the set of memberships estimated by the algorithms is independent with respect to the true memberships, and the maximum of 1 when the estimated memberships perfectly recreate the true classes (Manning, Raghavan, and Schütze 2008).

Simulation Results

The results of these 5000 tests are presented in Table 1 under the heading "Simulation 1". The median accuracy of CCA (0.87) is higher than that of RCA[7] (0.74), a difference that is highly significant statistically (Wilcoxon $W = 8585271$, $p < 0.0001$). The interquartile range (IQR) of CCA's accuracy extends from 0.69 to 0.97,while RCA's extends 0.54 to 0.88. Thus, while CCA's 75th percentile is just shy of a perfect accuracy, RCA's 75th percentile barely surpasses CCA's median. The substantive significance of these differences is clearer when the CCA accuracies (*Y*) are plotted against the RCA

accuracies (*X*) in Figure 4. While the accuracies are strongly associated ($R^2 = 0.79$), CCA is more accurate than RCA in the vast majority of cases (88.1%). In contrast, RCA is more accurate than CCA in only 5.2% of the cases. Thus, when RCA and CCA disagree, which they do in 93.3% of the cases, the odds that CCA's result is more accurate than RCA's exceed 16:1.

[Figure 4 about here]

[Table 1 about here]

To determine if the results point to any classes of data where RCA would be preferable to CCA, I disaggregated them by schema variance and noise variance, which are the parameters most responsible for the difficulty of the classification task. Lower schema variances or higher noise variances result in more challenging signal to noise ratios, which should make the performance of both algorithms poorer. The loess curves demonstrating this effect are presented in Figure 5 below. The two curves closely track each other across both plots, again indicating that the performance of the two algorithms is strongly related. However, CCA's accuracy remains substantially above RCA's throughout the full ranges of both variances. All the classes of data that are challenging to CCA thus appear to be even more challenging for RCA. Moreover, RCA's accuracy does not catch up even in tests with the *least* challenging amounts of noise, which are depicted on the left side of figure 5B. Though CCA's average accuracy in these cases approaches 0.97, RCA's average accuracy never rises above 0.85.

[Figure 5 about here]

These results also allow a comparison of CCA and RCA performance when the schema variance is very low. In such situations, pairwise correlations between responses tend towards zero. On the other hand, their relationalities approach one. Thus, as Goldberg argues, the relationality between low-variance respondents is systematically higher than the correlation. However, Goldberg incorrectly infers that this means that "relationality does a better job at examining relationships between respondents whose responses have relatively low variance" (Goldberg 2011:Appendix A). The fact that relationality produces *higher* values does not imply that it produces *more accurate* values. And indeed, as can be seen on the left side of figure 5A, the opposite appears to be true. As schema variance decreases to its simulation minimum of 0.3, CCA's accuracy drops to 0.38. However, RCA's accuracy drops all the way down to 0.08. Thus, RCA results for low-variance schemas appear to contain almost no information about the true membership structure of the data. This suggests that relationality may have an upward bias for low-variance observations that is substantially more damaging to its performance that correlation's downward bias.

Distributional Assumptions

Relationality also introduces a strong distributional assumption which may further degrade its accuracy when violated. While a correlation of zero always indicates an absence of a linear relationship, the equivalent "null value" of relationality differs from dataset to dataset and is generally skewed above zero (Goldberg 2011:Appendix A). RCA attempts to compensate for this skew by re-centering the matrix of relationalities by its

mean. However, this approach only works under the assumption that the true mean relationality between all the rows in the data is zero, or, equivalently, that the relationality values are distributed symmetrically around their null value. This is generally the case only if the proportion of respondents following a schema without inverting it equals the proportion following its inverse[8]—a quantity I call *inversion probability*. For example, when this probability is 50%, the number of highbrow respondents following the schema "like classical, like opera, dislike rock, dislike country" would equal the number of lowbrow respondents following its inverse, "dislike classical, dislike opera, like rock, like country". However, since in reality the number of highbrow respondents can differ greatly from the number of lowbrow respondents, there is no reason to expect that the inversion probability generally equals 50%. This suggests that RCA's symmetry assumption may be frequently violated by empirical data.

All the simulations presented above have granted RCA's symmetry assumption by keeping the inversion probability fixed at 50%. To examine the performance of both algorithms when this assumption is relaxed, I created a second simulation where the inversion probability is instead drawn from a uniform distribution over its full range, and varies between each of the 5000 simulation runs[9]. The results of this second simulation are reported on the right side of Table 1. As expected, both the median (0.86) and the interquartile range (0.69 to 0.97) of CCA accuracies remain unchanged from the first simulation. On the other hand, the median accuracy of RCA drops to 0.67, significantly lower than its prior median of 0.74 (Wilcoxon $W = 13802245$, $p < 0.0001$), and below CCA's 25th percentile of accuracy. CCA is now more than 3 times as likely as RCA to

produce a nearly perfect answer (NMI > 0.95), while RCA is more than 20 times as likely to produce an almost completely incorrect one (NMI < 0.05). Thus, RCA's accuracy appears to suffer a significant further drop when its assumption of symmetrical distribution is violated, as may generally happen in empirical applications. CCA remains unaffected by this change.

The results of both simulations thus reinforce my earlier suppositions. Though the accuracies of RCA and CCA were highly correlated, they nonetheless differed in almost 95% of the 10000 combined simulation runs. In simulation 1, which obeyed RCA's symmetry assumption, the odds that CCA's result was more accurate exceeded 16:1. In simulation 2, where this assumption was relaxed, these odds further rose to 23:1. CCA was more accurate over the full range of schema and noise variances examined, thus providing no evidence of any category of cases in which RCA would be preferable to CCA. On the other hand, when the schema variance was low, RCA's performance suffered a significant further drop relative to CCA's. Therefore, results from these 10000 simulations point to the same conclusion as the introductory example: the correlation-based approach to identifying schematic classes is more accurate than the one based on relationality.

EMPIRICAL EXAMPLE: MUSICAL TASTES

To compare the results produced by the two methods in an empirical setting, I applied CCA to the 1992 GSS music tastes module previously analyzed with RCA (Goldberg 2011). This dataset contains 1532 respondents' evaluations of 17 musical

genres. Each respondent rated each genre using a five-point Likert scale that ranges from "like very much" to "dislike very much". For comparability, I followed exactly the same coding procedures as Goldberg (2011).

Goldberg's RCA analyses partitioned the survey population into three schematic classes, which he labelled "Omnivore – Univore", "Highbrow – Lowbrow", and "Contemporary – Traditional." For respondents in the "Omnivore – Univore" class, most genre tastes were positively correlated among each other. Goldberg interpreted this as evidence of a culturally omnivorous taste schema, in which no genres are perceived as opposites, but rather a high appraisal of most genres is opposed to a low appraisal of most genres. In the "Highbrow – Lowbrow" class, tastes for "elitist" genres such as opera and classical music were positively correlated among each other, but negatively correlated with most tastes for popular genres. Finally, Goldberg characterizes the third schematic class as "Contemporary – Traditional." Here, a cluster of positively correlated tastes for well-established musical genres including gospel, bluegrass, and country is negatively correlated to tastes for arguably more contemporary genres, including heavy metal, pop and rap, as well as oldies and jazz. The persistence of the Highbrow – Lowbrow taste schema was perhaps Goldberg's most surprising finding, as much contemporary work has argued that omnivorousness has replaced highbrow tastes as a marker of high status in the contemporary United States (Peterson and Kern 1996; Peterson 1997, 2005; see Goldberg 2011 for a more detailed discussion).

The CCA analyses of these data, however, cast doubt on this finding. While RCA identified three classes in these data, CCA identified four, which are presented in Figure

6 below. The first two of these closely resemble those located by RCA. The first class features practically no negative correlations between the genres, suggesting that respondents in this class perceive little opposition between different musical styles. In this population, positive appraisal of any one genre generally "goes with" positive appraisal of any other genre, suggesting an undiscriminating logic of taste that ranges between near-uniformly positive appraisals of all genres on one extreme, and a near-uniformly negative appraisal of all genres on the other. This is the same omnivorous logic as behind the Omnivore – Univore class identified by RCA.

[Figure 6 about here]

The second class located by CCA appears to be defined by an opposition between rock, rap and metal on one extreme, and gospel, country, folk and bluegrass on the other. This suggests a bifurcation of respondents into those who prefer newer musical genres and those who prefer more established ones, which closely resembles the logic of the Contemporary – Traditional class identified by RCA. However, while the RCA analyses had counterintuitively suggested that blues, latin and jazz belong to the contemporary side of this schematic logic, the CCA analyses instead suggest that blues belongs to the traditional side, whereas latin and jazz straddle the two sides without clearly belonging to either. Thus, while substantively similar, the Contemporary – Traditional class identified by CCA appears to have greater face validity.

Goldberg's most controversial claim concerns the remaining group of respondents, who he contends follow a traditional logic which differentiates between "highbrow" genres such as opera and classical music on the one extreme, and popular

"lowbrow" genres on the other. However, the CCA results contain no evidence of such a class. Instead, CCA separates the remaining population into two further schematic classes (see panels C and D of Figure 6). In both, the majority of genres are tied together in a dense cluster of positive correlations, thus suggesting that both are variants of an omnivorous taste schema. These results resemble previous findings that documented the existence of multiple distinct logics of omnivorousness (e.g., Tampubolon 2008).

However, the omnivorousness of respondents in these last two classes features clear exceptions. In the class depicted on the left, a higher appraisal of most genres is generally accompanied by a *lower* appraisal of heavy metal (and frequently also rap music.) The class on the right exhibits a nearly identical structure, except country and gospel music occupy the same position of exclusion as metal and rap did in the class on the left. These patterns closely echo the analyses of Bryson (1996), who famously showed that omnivores may retain a symbolic boundary against genres most closely associated with low education: heavy metal, rap, country and gospel music. Thus, drawing on the title of Bryson's work, I term these latter two classes "Anything (but) Heavy Metal" and "Anything (but) Country".

Table 2 cross-tabulates the group assignments made by the two algorithms. A plurality (though not a majority) of respondents that RCA grouped into its first two classes remain grouped together in the CCA results. However, the respondents that belonged to RCA's third class are completely dispersed. Such a divergence between CCA and RCA class assignments is consistent with the simulation analyses above, where the results of the two methods were correlated but rarely the same, with CCA offering the

more accurate answer in the overwhelming majority of cases. While the simulations showed that this difference is statistically reliable, these GSS analyses illustrate that it can also be substantively important.

[Table 2 about here]

DISCUSSION

The Relational Class Analysis methodology (Goldberg 2011) aims to uncover shared cultural schemas that organize the cultural tastes of distinct groups of survey respondents. The method rests on an eponymous "relationality" measure to quantify the extent that two respondents appear to share one cultural schema. Though rhetorical arguments justifying this measure have been compelling, they have lacked a solid formal underpinning. In this paper, I built on the theoretical reasoning behind RCA to provide a formal definition of shared cultural schemas, the latent construct that relationality is designed to measure. This formal reasoning made it clear that schematic similarity as conceived by Goldberg (2011) should manifest itself in linear dependency between two response vectors—the same measurement task for which Pearson's correlation has long been an established solution.

When I applied Pearson's correlation to the same example that Goldberg (2011) used to introduce relationality, I found that correlation yielded substantially more accurate results. To determine whether this difference extends to other situations, I then analyzed 10000 diverse simulated datasets. These results confirmed that Pearson's

correlation is a substantially more accurate measure of schematic similarity than relationality. Across the full range of simulation parameters, the accuracy of correlation-based CCA remained reliably higher than that of relationality-based RCA. In those challenging simulations where schema variances were low, CCA's accuracy dropped to 0.38, but RCA's dropped to 0.08. Conversely, in the near absence of statistical noise, CCA's accuracy approached 0.97, but RCA's never rose above 0.85. When the two methods produced different results, the odds that CCA's answers were more accurate exceeded 16:1 even in simulations that obeyed the strong distributional assumption introduced by relationality. When the simulations violated this assumption, those odds rose to 23:1. A re-analysis of the 1993 GSS musical tastes module confirmed that these differences can lead to substantively different conclusions in empirical settings. Thus, these analytical, computational and empirical results together strongly suggest that there are substantial reasons to prefer the correlation-based CCA methodology proposed here over the relationality-based RCA.

Limitations and Future Directions

Proponents have argued that RCA "does not require any presuppositions" (Goldberg and Baldassari forthcoming:3) about how attitudes are organized in the population. This claim, however, may be misleading. While RCA indeed inductively determines the *specific* cultural schemas and schematic class assignments that describe a given empirical population, its ability to do so rests on implicit theoretical assumptions about the *general* structure, function and distribution of such classes and schemas. In the

present paper, I demonstrated the usefulness of laying out such presuppositions with greater clarity.

While this paper has formally defined the concept of shared cultural schema, however, there is still a second black box that remains to be opened. CCA, like RCA, uses modularity maximization to partition the survey population into schematic classes. As the simulations in this paper demonstrate, it does so with acceptable accuracy. But, while modularity maximization is among the most widely used network analytic techniques across many disciplines (e.g., Neal 2014; Newman 2013; Shwed and Bearman 2010; Porter, Onnela, and Mucha 2009), its use may introduce further assumptions. A substantial literature notes that modularity maximization suffers from a "resolution limit" that biases it against detecting both very small and very large modules in many empirical settings, joining even very different modules together when they are too small in proportion to the whole network, and conversely breaking apart modules that are too large (Fortunato and Barthélemy 2007; Lancichinetti and Fortunato 2011; for a thorough overview, see Good, de Montjoye, and Clauset 2010).

By relying on modularity maximization, RCA and CCA may thus unintentionally introduce an assumption that each member of the survey population belongs to one of a moderate number of schematic classes, and may produce misleading results when this is not the case—e.g., in situations where many respondents follow their own idiosyncratic response patterns, or where the whole population belongs to only one schematic class. Future methodological work should determine whether these methods indeed exhibit this problem, and offer solutions if they do[10]. Before such work takes place, however, it may

be prudent for empirical researchers to seek external evidence that the population under study features the kind of heterogeneity RCA and CCA are designed to seek.

CONCLUSION

In this paper, I demonstrated that Pearson's correlation is a better measure of schematic similarity than the relationality measure introduced by Goldberg (2011). Correlational Class Analysis (CCA) proved reliably more accurate at partitioning survey populations by shared cultural schema than RCA. This change from relationality to correlation may bring a number of further benefits. Relationality is substantially more computationally costly to calculate than correlation, and also requires bias correction and extensive bootstrapping for significance testing. Correlation obviates the need for these further steps. This leaves an algorithm that is clear, fast and easy to implement (see Appendix A). It also clarifies and standardizes the method, thus placing it in conversation with existing methodologies in other disciplines—e.g., the correlation network approaches in bioinformatics which employ very similar analytical steps in a different empirical domain (e.g., Langfelder and Horvath 2007). Future improvements of CCA can draw insights from these existing literatures, thus helping further build on Goldberg's methodological innovations.

APPENDIXES

Appendix A. CCA Algorithm

Correlational class analysis can be implemented in minutes in any programming environment which supports network partitioning by modularity maximization. It consists of only four steps:

1. Create a matrix *G* of absolute row correlations between survey respondents.
2. Set statistically insignificant correlations to 0 to reduce noise (e.g., using *t*--tests[11]).
3. Import *G* into a network analysis package, treating it as an adjacency matrix.
4. Use the existing network partitioning routines to produce the group assignments.

In the *R* statistical environment with the *igraph 0.7* library, this can be implemented as:

```
CCA <- function (dataset, min.significant.row.cor = 0.60)
{
   C <- abs(cor(t(dataset)))                    # 1st step
   C[C < min.significant.row.cor] <- 0          # 2nd step
   G <- graph.adjacency(C, mode="undirected",
             weighted = TRUE, diag = FALSE)     # 3rd step
   leading.eigenvector.community(G)$membership # 4th step

}
```

A more full-featured implementation of the method is available on *CRAN* under the name "*corclass"*.

Since correlation is normalized by the product of variances, it is undefined when the variance of either respondent equals zero. The minimal implementation above requires that such respondents be dropped from the analysis, which may generally be an acceptable solution given the rarity of such respondents in empirical survey data.

However, to keep CCA's results comparable to RCA's, the algorithm I use to analyze the simulated datasets instead sets correlations between two zero-variance respondents to 1, and their correlations with others to 0.

Appendix B. Simulation Procedure

Each of the 5000 simulations is generated in three steps:

1) I first randomly set the maximum ranges of various broad simulation parameters by drawing them from the uniform distribution: schema variance $v \sim U[0.3,3]$, noise variance $\epsilon \sim U[0,3]$, maximum shift $\delta \sim U\{0, \dots, 3\}$ maximum scaling $\gamma \sim U\{1, \dots, 3\}$, and number of schematic classes $c \sim U\{2, \dots, 6\}$.
2) Then, for each of the c classes, I randomly generate a schema vector $\rho = \{\rho_1, \dots \rho_{10}\}$ by drawing from the Normal distribution, $\rho_i \sim N(\mu = 0, \sigma^2 = v)$, and rounding to the nearest integer. Any duplicate vectors are discarded, and new vectors generated in their place, until I have c unique vectors. Then, I randomly set the counts $n_1, \dots, n_k$ of respondents in each schematic class, $n_i \sim U\{100, \dots, 500\}$. The range of the 10 taste variables is then fixed at $\pm[\max(|p_i|) * \gamma + \delta]$.
3) Finally, for each respondent $f \in [1, n_p]$ following schema ρ_p, I simulate the response $X_{pf} = (k_f * \rho_p) + \delta_f + \epsilon_f$ by drawing the values of the vertical shift $\delta_f \sim U\{-\delta, \dots, \delta\}$ and the scaling and inversion factor $k_f \sim U\{[-\gamma, \dots, -1] \cup [1, \dots, \gamma]\}$. I generate each respondent's noise vector ϵ_f by first drawing an

individual noise variance $E_f \sim U(0, \epsilon)$, and then drawing each element, $\epsilon_{fi} \sim N(\mu = 0, \sigma^2 = \epsilon_f)$, rounded to the nearest integer.

I analyzed each simulation run using both CCA and RCA (Goldberg and Zhai 2013), and then compared the results.

ENDNOTES

[1] In Martin's (2002) terminology, such schemas thus underlie the "tightness" rather than the "consensus" of a system of attitudes. If one imagines attitudes as an abstract space where each dimension represents a like/dislike of a given musical genre, such schemas would specify an axis along which culturally valid tastes can be arranged rather than specifying a specific point in space at which tastes should be located.

[2] Formally, Goldberg (2011) defines the relationality between two respondents *i* and *j* to equal $R_{ij} = \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} (\lambda_{ij}^{kl} * \delta_{ij}^{kl})$, where $\delta_{ij}^{kl} = 1 - |\left|\Delta X_i^{kl}\right| - |\Delta X_j^{kl}|$, and where $\Delta X_i^{kl} = X_i^k - X_i^l$ is the difference between the values of the variables *k* and *l* for respondent *i*, and $\lambda_{ij}^{kl} = 1$ if ΔX_i^{kl} and ΔX_j^{kl} have the same sign, and $\lambda_{ij}^{kl} = -1$ otherwise. The matrix of relationalities is then re-centered by its mean to correct for its bias under the assumption that the true mean value of $R_{ij} = 0$.

[3] Respondents are assigned to classes based on the absolute values of the relationalities $|R_{ij}|$, so I plot the values $|R_{ij}|$ instead of R_{ij} to enable easier visual comparison of their magnitudes. I also omit the Euclidean distances, which are not relevant to the present discussion. These changes make the limited dynamic range of relationality values more visually apparent. Since Goldberg does not present the numerical relationality scores for this example, I reproduce these values by measuring the bar lengths in his figure.

[4] When I ran RCA with default parameters, it partitioned the population into 800 separate classes, thus assigning even identical rows to different classes. This obviously faulty

solution appears to be due to the pseudo-significance testing RCA uses to filter weak relationalities, which is based on strong assumptions about how relationalities are distributed in the data. Disabling it produced the substantially more realistic solution I report above. (As I discuss below, this filter appears to generally decrease the average accuracy of RCA.)

[5] For example, the schema shared by A, B and C in Figure 1 can be specified as $\rho = \{0,0,0,1,1,-1,-1\}$, so that $A = \rho + 4,\ B = \rho + 2$, and $C = -2\rho + 3$.

[6] Because of an apparent bug, the RCA software repeatedly crashed for 69 (1.3%) of these simulated datasets, producing no results. I excluded these cases from the analysis.

[7] RCA software contains a user-configurable filtering step where weak relationalities are dropped prior to partitioning. I examined how filtering affected RCA's performance using 250 simulation runs. Filtering increased accuracy in 56% of the cases but and decreased it in 43%. However, the average decrease (-0.33) was three times greater than the average increase (0.11). Overall, disabling the filter substantially raised RCA's median accuracy, from 0.56 when enabled to 0.69 when disabled. Additionally, in 10% of the cases with filtering, RCA encountered an error and yielded no solution at all (as compared to 1% without filtering). Thus, to increase RCA's accuracy and avoid potential bias from substantial missing results, I disabled the filter for all the simulations reported in this paper.

[8] It can also hold in a number of degenerate or improbable cases—e.g., when there are roughly as many taste schemas as there are respondents.

[9] The changes to the simulation procedure described in Appendix B are as follows. In step 1, I now also draw a random inversion probability: $\zeta \sim U[0,0.5]$. In step 3, I now draw a random inversion factor $z_f \in \{1, -1\}$, with $P(z_f = -1) = \zeta$. Since factor k_f now controls the scaling but not the inversion, I now restrict it to positive values: $k_f \sim U[1, \gamma]$. Each respondent f following schema ρ_p is generated by $X_{pf} = (z_f * k_f * \rho_p) + \delta_f + \epsilon_f$.

[10] For an example of a domain-specific fix to problems stemming from modularity's resolution limit, see Sohn and colleagues (2011).

[11] My exploratory results suggest that more stringent cutoffs may produce more accurate results as long as they are not so extreme as to turn some nodes into isolates. I used $\alpha = 0.05$ as the cutoff for the simulations reported above, and $\alpha = 0.01$ for the GSS analyses. A `min.significant.row.cor` of 0.60 approximates a *t*-test at $\alpha = 0.01$ for rows of 17 variables.

REFERENCES

Bryson, Bethany. 1996. "'Anything But Heavy Metal': Symbolic Exclusion and Musical Dislikes." *American Sociological Review* 61(5):884–99.

Danon, Leon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. "Comparing Community Structure Identification." *Journal of Statistical Mechanics: Theory and Experiment* 2005(09):P09008.

DiMaggio, Paul. 1997. "Culture and Cognition." *Annual Review of Sociology* 23:263–87.

DiMaggio, Paul. 2011. "Cultural Networks." Pp. 286–300 In *The Sage Handbook of Social Network Analysis*.

DiMaggio, Paul, and Amir Goldberg. 2010. "Searching for Homo Economicus: Variation in the Structure of Americans' Moral Evaluations of Markets." Atlanta, GA.

Emirbayer, Mustafa. 1997. "Manifesto for a Relational Sociology." *American Journal of Sociology* 103(2):281–317.

Fortunato, Santo, and Marc Barthélemy. 2007. "Resolution Limit in Community Detection." *Proceedings of the National Academy of Sciences* 104(1):36–41.

Goldberg, Amir. 2011. "Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined." *American Journal of Sociology* 116(5):1397–1436.

Goldberg, Amir, and Delia Baldassarri. Forthcoming. "Neither Ideologues, nor Agnostics: Alternative Voters' Belief System in an Age of Partisan Politics." *American Journal of Sociology*.

Goldberg, Amir, and Jinjian Zhai. 2013. *RCA R Package*.

Good, Benjamin H., Yves-Alexandre de Montjoye, and Aaron Clauset. 2010. "Performance of Modularity Maximization in Practical Contexts." *Physical Review E* 81(4):046106.

Lancichinetti, Andrea, and Santo Fortunato. 2011. "Limits of Modularity Maximization in Community Detection." *Physical Review E* 84(6):066122.

Lancichinetti, Andrea, Santo Fortunato, and János Kertész. 2009. "Detecting the Overlapping and Hierarchical Community Structure in Complex Networks." *New Journal of Physics* 11(3):033015.

Langfelder, Peter, and Steve Horvath. 2007. "Eigengene Networks for Studying the Relationships between Co-Expression Modules." *BMC Systems Biology* 1(1):54.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.

Martin, John Levi. 2002. "Power, Authority, and the Constraint of Belief Systems." *American Journal of Sociology* 107(4):861–904.

Miranda, Shaila, Jama Summers, and Inchan Kim. 2012. "Visions of Social Media: Surfacing Schemas from Firms' Informational Engagements." *ICIS 2012 Proceedings*.

Mohr, John W. 1998. "Measuring Meaning Structures." *Annual Review of Sociology* 24:345–70.

Mohr, John W., and Craig Rawlings. 2012. "Four Ways to Measure Culture: Social Science, Hermeneutics and the Cultural Turn." In *The Oxford Handbook of Cultural Sociology*. Oxford University Press, USA.

Neal, Zachary. 2014. “The Devil Is in the Details: Differences in Air Traffic Networks by Scale, Species, and Season.” *Social Networks* 38:63–73.

Newman, Mark E. J. 2013. “Spectral Methods for Community Detection and Graph Partitioning.” *Physical Review E* 88(4):042822.

Newman, Mark E. J. 2006. “Modularity and Community Structure in Networks.” *Proceedings of the National Academy of Sciences* 103(23):8577–82.

Peterson, Richard A. 1997. “The Rise and Fall of Highbrow Snobbery as a Status Marker.” *Poetics* 25(2–3):75–92.

Peterson, Richard A. 2005. “Problems in Comparative Research: The Example of Omnivorousness.” *Poetics* 33(5–6):257–82.

Peterson, Richard A., and Roger M. Kern. 1996. “Changing Highbrow Taste: From Snob to Omnivore.” *American Sociological Review* 61(5):900–907.

Porter, Mason, Jukka-Pekka Onnela, and Peter Mucha. 2009. “Communities in Networks.” *Notices of the American Mathematical Society* 56(9).

Rodgers, Joseph Lee, and W. Alan Nicewander. 1988. “Thirteen Ways to Look at the Correlation Coefficient.” *The American Statistician* 42(1):59–66.

Saussure, Ferdinand de. 1916. *Course in General Linguistics*. Columbia University Press.

Shwed, Uri, and Peter S. Bearman. 2010. “The Temporal Structure of Scientific Consensus Formation.” *American Sociological Review* 75(6):817–40.

Sohn, Yunkyu, Myung-Kyu Choi, Yong-Yeol Ahn, Junho Lee, and Jaeseung Jeong. 2011. “Topological Cluster Analysis Reveals the Systemic Organization of the Caenorhabditis Elegans Connectome.” *PLoS Comput Biol* 7(5):e1001139.

Stockburger, David. 2007. *Introductory Statistics: Concepts, Models, and Applications*. Cengage Learning.

Tampubolon, Gindo. 2008. "Revisiting Omnivores in America circa 1990s: The Exclusiveness of Omnivores?" *Poetics* 36(2–3):243–64.

Wu, Angela Xiao. Forthcoming. "Ideological Polarization Over a China-as-Superpower Mindset: An Exploratory Charting of Belief Systems Among Chinese Internet Users, 2008-2011." *International Journal of Communication*.

FIGURES

Figure 1. Musical tastes of four respondents, with evaluations ranging from 1 (strongly dislike) to 5 (strongly like) for each genre. Respondents A, B and C follow the same taste schema. Respondent D does not. This figure recreates the contents of Goldberg's Figure 1A (Goldberg 2011:1405).

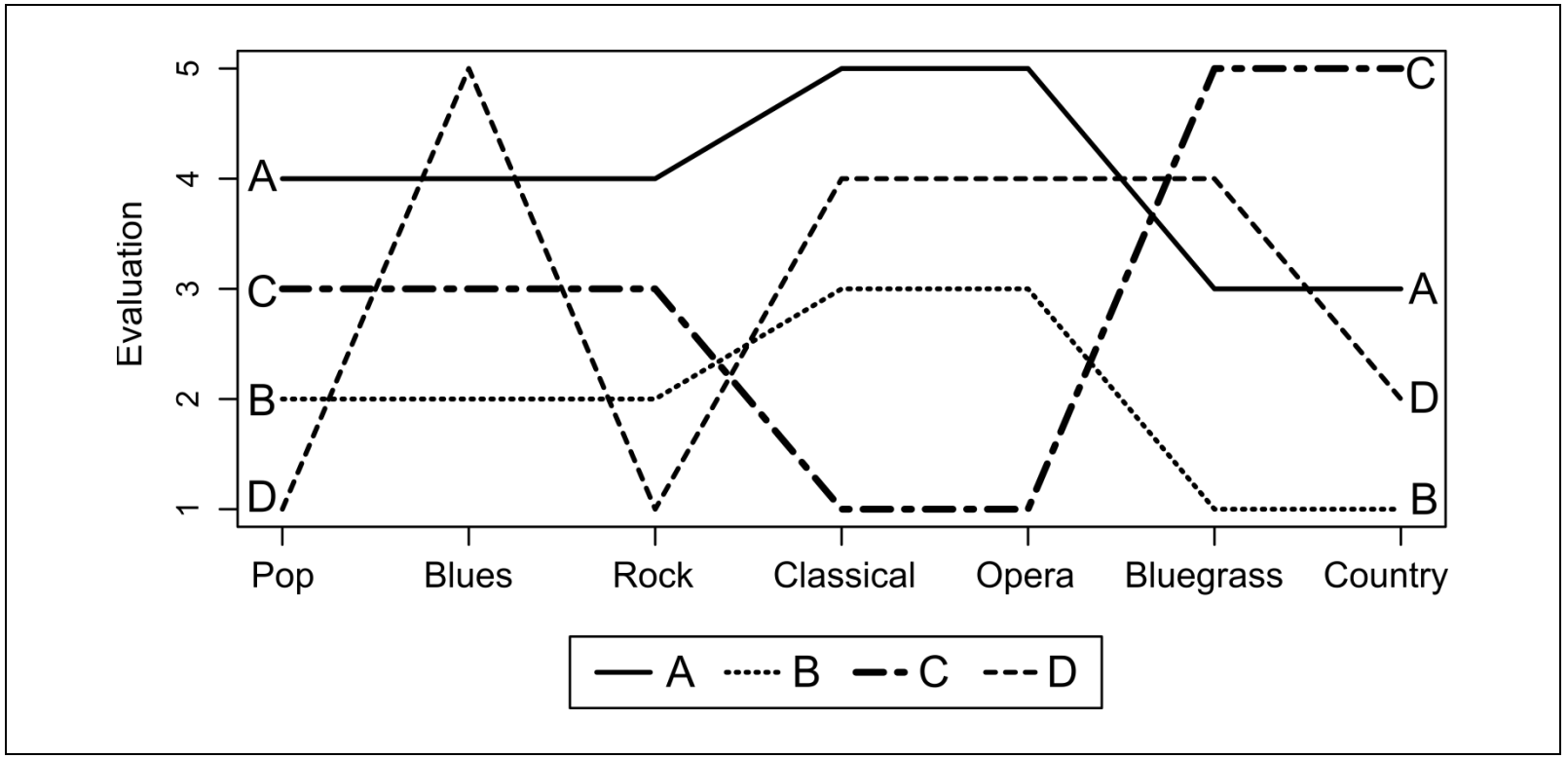


Figure 2. The absolute values of the pairwise relationalities (A) and pairwise correlations (B) of the four patterns depicted in Figure 1. The goal of both measures is to correctly determine that A, B and C belong to one schematic class, but D does not. Relationality results appear to detect the similarity between A and B, but not between A and C or B and C.

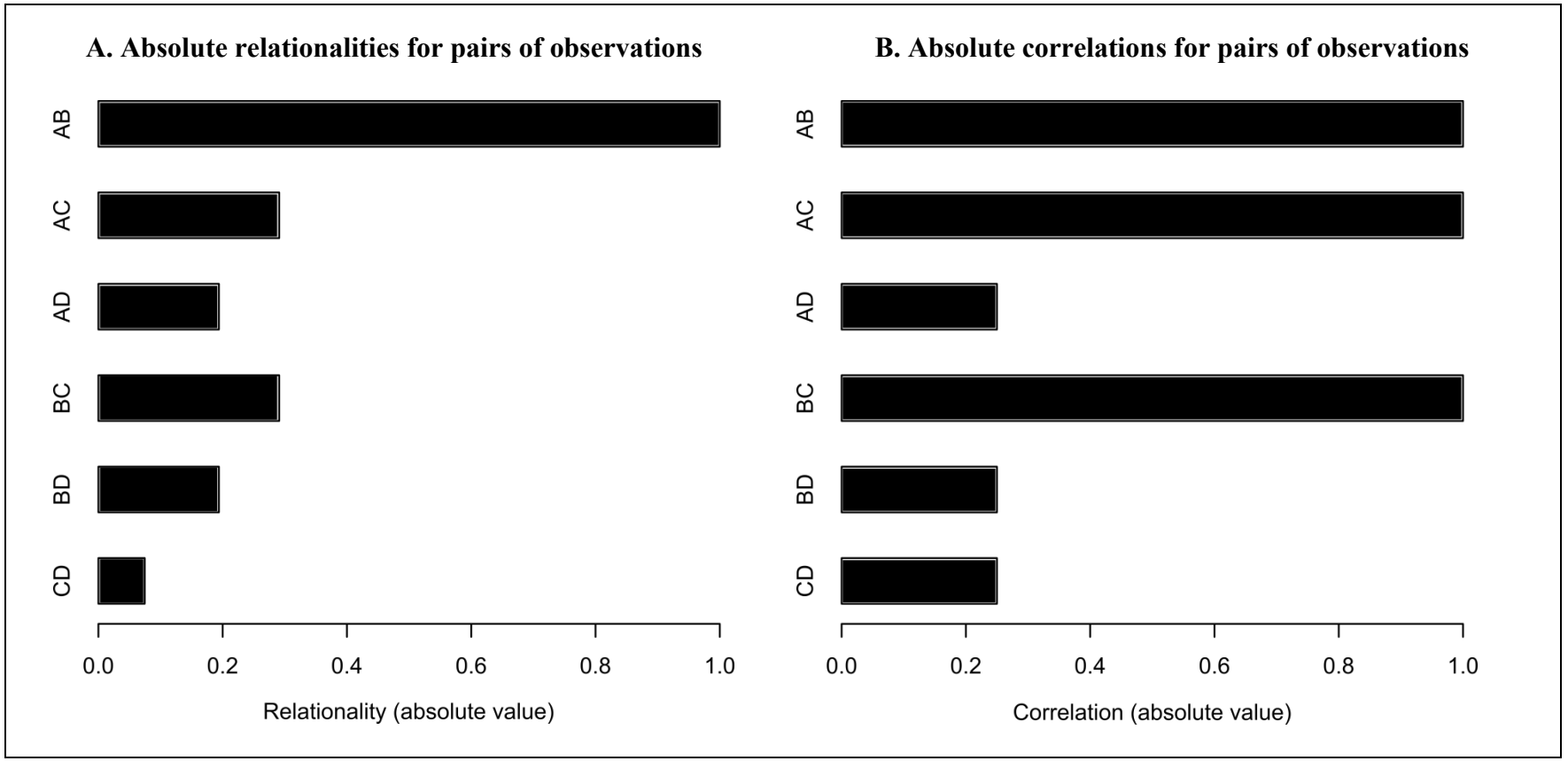


Figure 3. Three simulated schemas for one simulation run. The schemas are plotted in black, and a small sample of responses derived from each schema is plotted in dashed gray in the same plot. In this run, the patterns have a variance of 0.51, and the maximum noise variance used in deriving individual responses is 1.052. This noise variance is moderately high in proportion to the schema variance, creating a relatively difficult classification task.

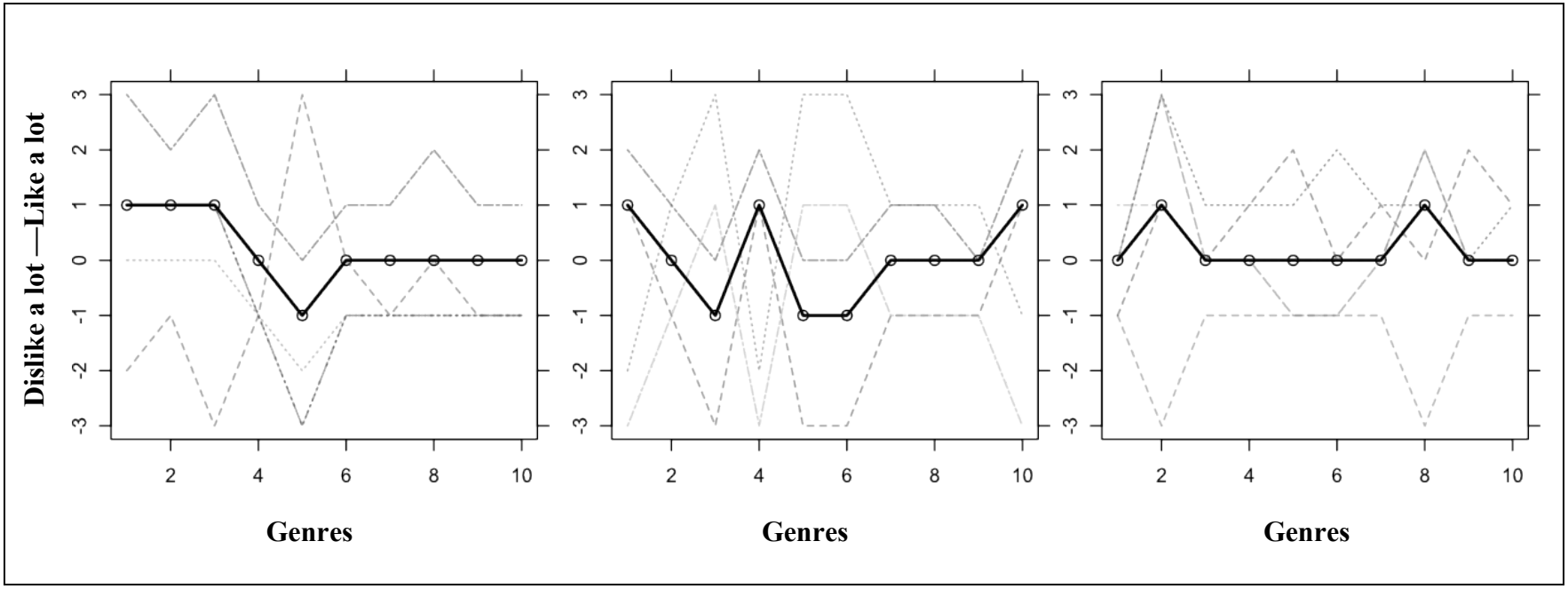


Figure 4. CCA accuracy compared to RCA accuracy for 5000 simulation runs. Each point represents a single simulation run. Runs where CCA produced the more accurate result are above the $Y = X$ diagonal (in gray), while those where RCA was more accurate are below it. Note the absence of points near the bottom-right corner of the plot.

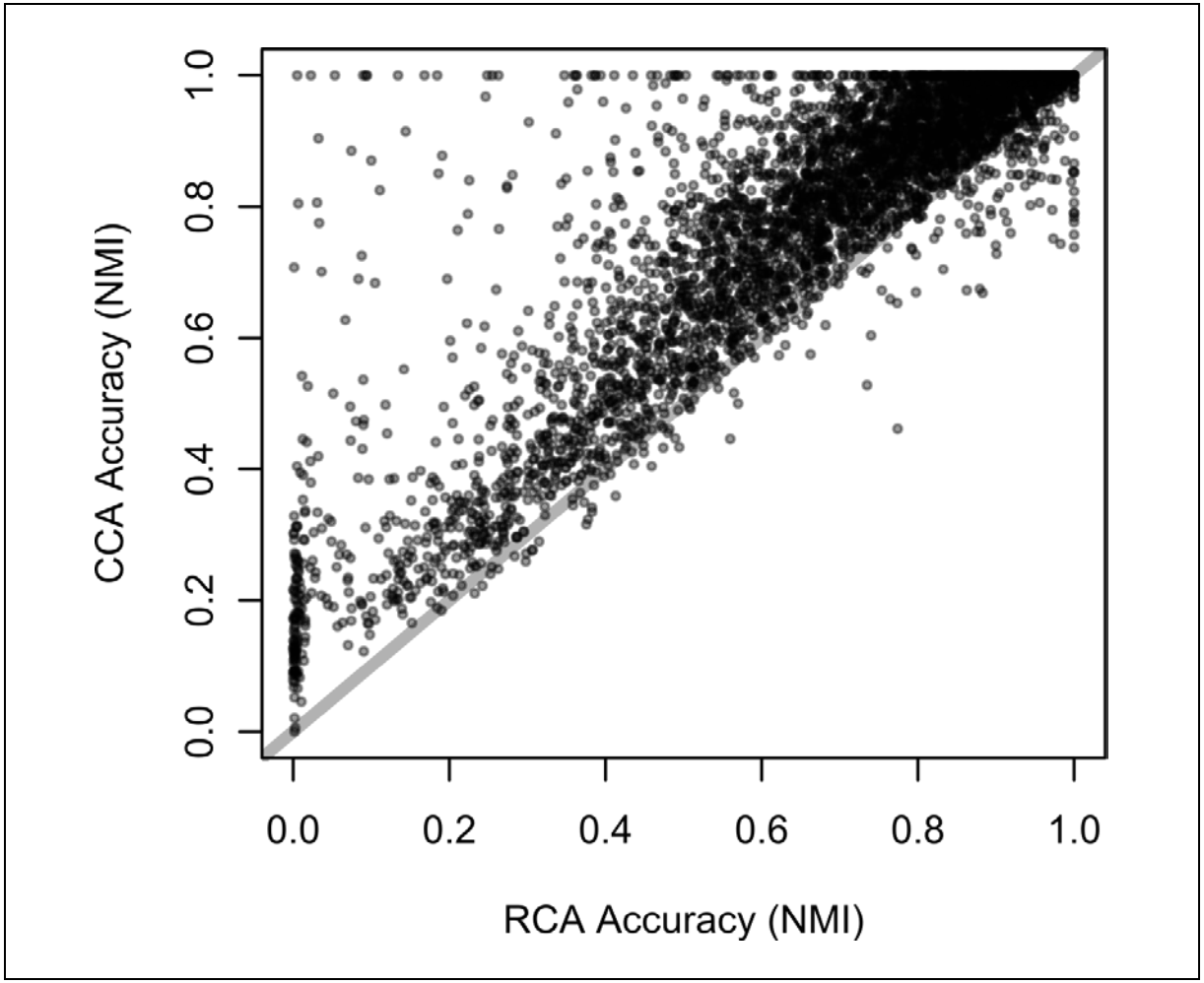


Figure 5. Loess curves comparing RCA and CCA accuracy by schema variance (left), and by noise variance (right), based on 5000 simulation runs.

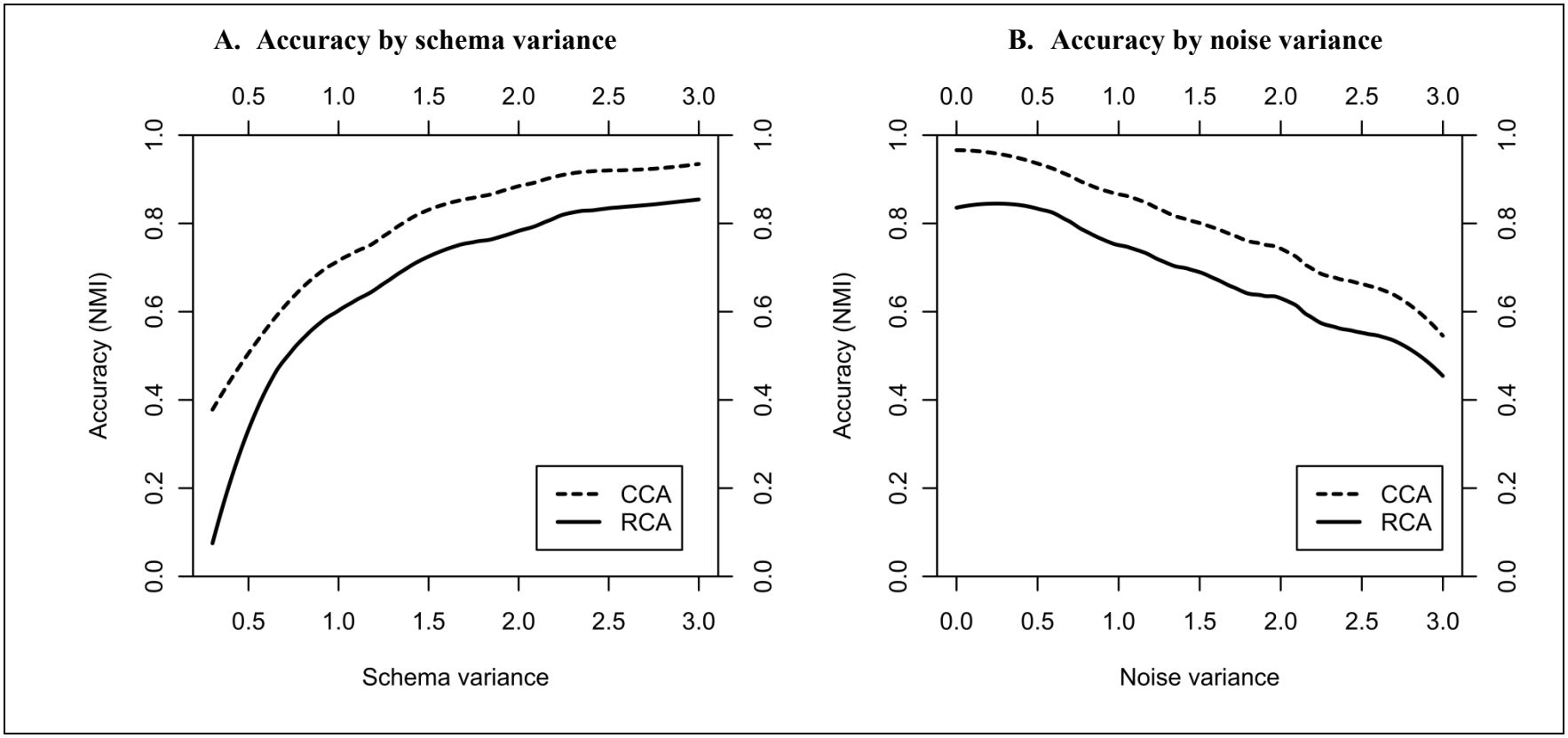
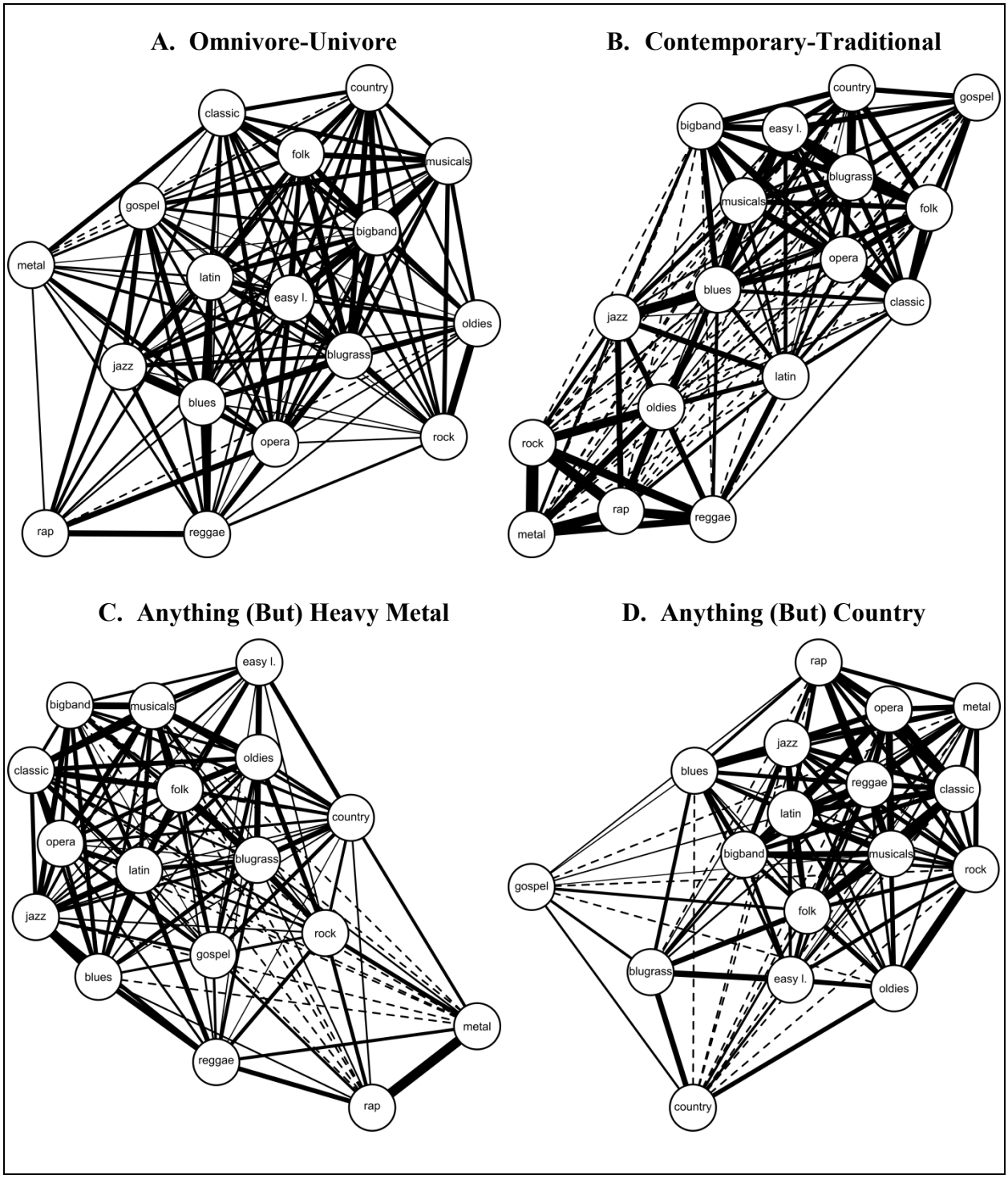


Figure 6. Networks illustrating the four correlational classes present in the data. Dashed lines indicate negative correlations. Weak correlations ($|r| < 0.05$) were not plotted.



TABLES

Table 1. Comparison of RCA and CCA Accuracy in 10,000 Simulation Runs

Measure	Simulation 1 (5000 runs)		Simulation 2 (5000 runs)	
	Relationality (RCA)	Correlation (CCA)	Relationality (RCA)	Correlation (CCA)
Overall accuracy (median NMI)	0.74	0.87	0.67	0.87
Accuracy, interquartile range (25% to 75%)	(0.54, 0.88)	(0.69, 0.97)	(0.46, 0.84)	(0.69, 0.97)
Runs with near-perfect accuracy (NMI > 0.95)	13.2%	30.5%	9.2%	30.3%
Runs with near-complete inaccuracy (NMI < 0.05)	2.8%	0.1%	3.1%	0.1%
Runs with higher accuracy than other method	5.2%	88.1%	3.8%	91.6%
Approximate odds of higher CCA accuracy in a given run	1 : 16.9		1 : 23.8	

Note. Results from simulation runs with randomly varying schema variances, noise amounts, ranges of linear transformation constants, and other simulation parameters. In Simulation 2, the inversion odds (proportions of positive and negative multipliers used in the linear transformations) also randomly varied.

Table 2. Cross-tabulation of Estimated Schematic Class Memberships in 1993 GSS Music Tastes Data

	CCA class			
RCA class	Omnivore – Univore	Contemporary – Traditional	Country – Anything But	Heavy Metal – Anything But
Omnivore – Univore (673)	281	92	167	133
Contemporary – Traditional (394)	60	241	35	58
Highbrow – Lowbrow (461)	142	36	159	124
Total (N=1528)	483	369	361	315

Note. Cross-tabulation of class memberships estimated by CCA (columns) and RCA (rows). RCA group memberships are indicated in parentheses. Four out of 1532 respondents had no response variance and were omitted from this analysis.