EMPIRICAL STUDIES OF THE ARTS, 32(2) EOV7, 205-229, 2014

Research Article

COMPOSER SIMILARITIES THROUGH "THE CLASSICAL MUSIC NAVIGATOR": SIMILARITY INFERENCE FROM COMPOSER INFLUENCES

CHARLES H. SMITH
Western Kentucky University, Bowling Green

PATRICK GEORGES
University of Ottawa, Canada

ABSTRACT

Data from "The Classical Music Navigator" (*CMN*) website are used to generate statistics bearing on the relative stylistic similarities of 500 classical music composers. When the *CMN* was compiled, notice was taken of which composers are thought to have influenced which others; it is reasoned here that composers with similar arrays of composer influences should usually themselves be adjudged as similar. Measures of association like the ones more commonly used in fields such as biogeography and biological systematics were applied to pairwise comparisons across the 500 by 500 matrix of composer influences lists, and then examined for the relevant characteristics. Samples of the results using three measures in particular are given.

THE CLASSICAL MUSIC NAVIGATOR

Charles H. Smith created "The Classical Music Navigator" (Smith, 2000; hereafter referred to as *CMN*) as an experiment in music education and reference. To begin, a database was generated listing several main characteristics of each of 444 (now 500) classical composers, including their influences and main works. Indexing was then set up in such a fashion that the user could identify other composers who have produced works that are in one sense or another similar to

doi: http://dx.org/10.2190/EM.32.2.EOV.7
http://baywood.com

any given work. The philosophy behind this arrangement is based on a "points of familiarity" approach: that is, that the novice listener may happen to hear a particular work, and wish to find more of similar type. The word "type," of course, may lead in various directions. At the site itself the process is described thusly: "... translating [one's] interest into productive action (finding more music of the same likable type) implicitly depends on a series of evaluations. Perhaps it's the whole genre of music that holds the appeal; on the other hand, maybe it's only the instrumentation on the particular piece that's set you off. Or perhaps it's the message or style of the performing artist or composer. . . ." It originally had been hoped that the website might also provide streams of music to complement the textual information, but before this could happen other projects with more extensive backing emerged (such as "YouTube" and "Pandora"), and additional efforts were deemed unnecessary.

The original plan, because it was based on statistical analyses of composer attributes rather than personal opinion, remains useful as an online reference tool and "navigator" of composers and their prominent works. In 2007 an extensive review of the more recent literature led to an update, including an expansion in number from 444 to 500 of the composers treated, and a re-ranking.

One of the most time-consuming parts of the construction of the service was the collection of data on composer influences. Originally, more than 1,000 biographical, analytical, and reference sources were examined (and supplemented by information retrieved from reviews of recordings and database searches). For Version 2 the original data were supplemented through new database searches, further investigations of reference works, and a good deal of internet work focusing on online-available dissertations, album liner notes, and concert notes and reviews. At this point, opinions from at least 5,000 sources have contributed to the "influences" compilations. For the more prominent composers, influences were not recorded at the site unless at least five separate sources pointed to them; for lesser composers this standard could not be maintained, though at a minimum at least two sources were required. Unfortunately, some of the composers on the list have been rather little studied, so errors of omission are to be expected.

Nevertheless, a fair number of sources have stated publicly or privately their opinion that *CMN* represents the best (or at the least one of the best) sources of data regarding composer importance and influence. Some investigators (e.g., Georges & Seçkin, 2013; Jacobson, 2011) have made various analytical uses of the composer rankings, or the influences data, or both, whereas others (e.g., Anonymous Prof, 2008; Kurt Jx, 2008; OMRAS2, n.d.) have used them to prepare graphic representations of composer relations for instructional purposes. The authors believe, however, that a lot more can be done with these data than so far has been attempted, and to begin we investigate the *CMN*'s potential for establishing composer similarities on the basis of their common musical influences. In a future work we will examine how much, if any, additional explanation can be added when "ecological measures" (i.e., other composer

characteristics such as time period, school association, instrumentation emphases, etc.) are also taken into account.

COMPOSER SIMILARITIES

A vast and well-cited literature exists on the subject of *musical* similarities and the various means by which particular musical pieces may be related to others. One review categorizes the methods involved into three groups: "methods based on metadata, methods based on analysis of the audio content, and methods based on the study of usage patterns related to a music example (West & Lamere, 2007). A second review further notes that "Music similarity measures rely on one of three types of information: symbolic representations, acoustic properties, and subjective or 'cultural' information" (Logan, Ellis & Berenzweig, 2003). Meanwhile, similarities among the styles of particular classical music *composers* have also been investigated, through a variety of approaches across thousands of papers and books. There are so many such studies that a literature review here is pointless. Almost all of these analyses emphasize characteristics of one or a few individual composers and their likely influences, however, and although a number of investigations (e.g., Collins, 2010; de Carvalho & Batista, 2012; de Leon, 2002; Fazekas, Raimond, Jacobson, & Sandler, 2010; Filippova, Fitzgerald, Kingsford, & Benadon, 2012; Kaliakatsos-Papakostas, Epitropakis, & Vrahatis, 2010; Mostafa & Billor, 2009; Vieira, Fabbri, Travieso, Oliveira, & da Fontoura Costa, 2012) have sought to identify particular stylistic qualities and methodologies that might lead to generalizations, there have been very few if any global attempts to establish similarities solely from compiled lists of influences. The reason for this seems clear: before now there has not been a source of data that could be applied to such work. This article proposes using the composer influence data collected in *CMN* to infer the similarities among composers.

To illustrate, Figure 1 gives a small subset of the network of composers in the *CMN* by focusing on three major composers (J. S. Bach, Mozart, and Beethoven) and all composers flagged as having influenced them. An interesting characteristic of the graph is that the number of shared influences fluctuates among pairs of composers. For example, there are only two common influences between Bach and Beethoven (Palestrina and Handel), and two common influences between Bach and Mozart (Hasse and Handel). However, there are six common influences between Mozart and Beethoven (J. Haydn, Gluck, Fux, C. P. E. Bach, Handel, and J. S. Bach). The larger number of common influences between Mozart and Beethoven suggests a tighter proximity in the musical style of these two composers when compared to the music styles of Bach and Beethoven or of Bach and Mozart.

On the other hand, distinct influences may drive away (i.e., increase the distance between) the musical styles of any pair of composers. For example, we see in Figure 1 that 16 composers have influenced Beethoven and 16 composers have

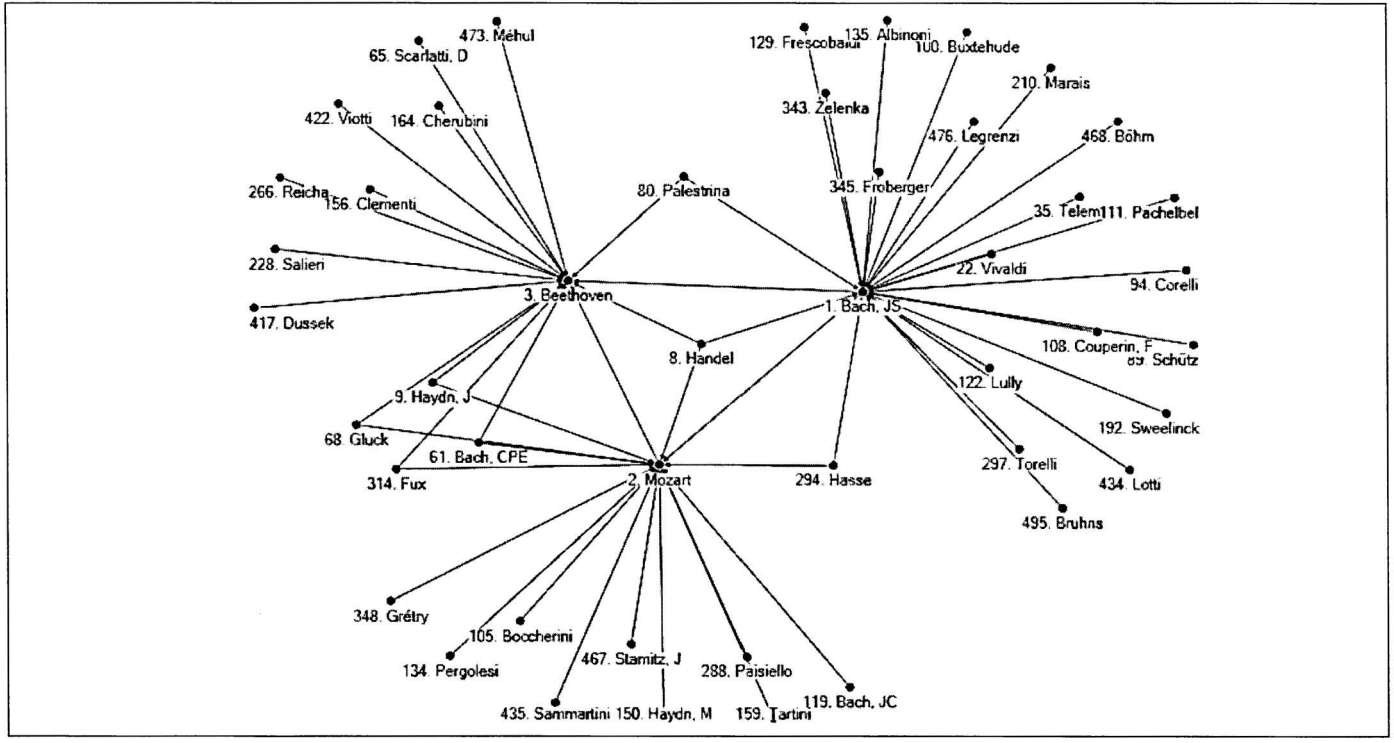

Figure 1. Influences on J. S. Bach, Mozart, and Beethoven.

Notes: The numbers in the graph are the composer rankings in the *CMN*. The length of the arrows linking influential composers to the three subject composers (J. S. Bach, Mozart, and Beethoven) should not be interpreted as a metric of distance, similarity, or influence strength between these pairs of composers.

influenced Mozart. Given that six composers have influenced both Beethoven and Mozart, then ten composers have influenced Beethoven but not Mozart (Mozart, D. Scarlatti, Palestrina, Clementi, Cherubini, Salieri, Reicha, Dussek, Viotti, and Méhul) and 10 composers have influenced Mozart but not Beethoven (Boccherini, J. C. Bach, Pergolesi, M. Haydn, Tartini, Paisiello, Hasse, Grétry, Sammartini, and J. Stamitz). Then there are 20 distinct influences that must have had a tendency to increase the distance between the musical styles of Mozart and Beethoven.

Figure 2 uses a set approach that will be further exploited in the methodology section. It gives the sets of composers that influenced Beethoven and/or Mozart. The intersection gives the set of six composers that influenced both Mozart and Beethoven. In general, the larger the number of composers in the intersection with respect to the number of composers outside the intersection, the more similar will be the style of the pair of composers under consideration. A slightly difficult case arises in the context of Figure 2 due to the direct influence of Mozart on Beethoven. It can be argued that this has driven the musical style of Beethoven towards the style of Mozart. Twisting slightly the interpretation of the intersection set in Figure 2, one might argue that Mozart has influenced both Beethoven and "himself," therefore shifting Mozart as a common influence (shifting Mozart to the intersection set in Figure 2). However, it can equally be argued that part of the difference between Mozart and Beethoven, at least with respect to other pairs of composers, is that Mozart didn't have Beethoven as an influence. Hence we decided to keep this type of occurrence outside the intersection set.[1]

Our objective here is to develop a methodology for gauging the similarity between any pair of composers based on common and distinct influences. This approach is reminiscent of the approaches used in biodiversity analyses. In these analyses, biologists and biogeographers attempt to identify relational patterns that may be useful in explaining observed diversities and distributions. In

[1] Furthermore, it just so happens in this instance that the pair of composers (Mozart, Beethoven) are relatively near to one another in time. However, what should be done for, say, a pair of composers such as J. S. Bach and Busoni, separated by 150 years? Yes, Busoni was influenced by Bach, but all his other influences were from people who lived after Bach's death, and in general it seems un-useful to add Bach as an element of the intersection in a figure such as Figure 2 adapted to the pair of composers (J. S. Bach, Busoni). Such a move might be made in, for example, biological phylogenetic analyses (if it can be assumed a particular character state has evolved out of another one under study), but in another sort of example, a biogeographic one, one usually makes a straight intersection set type of assessment before interpreting what it is about the changing landscape that is causing whole faunas to shift in preferred directions (multi-causality, not linear causality, being involved). Here, we do not have an overt spatial setting, but we do have links among composers we are defining here in terms of their individual influences. Part of the difference between Mozart and Beethoven, at least with respect to other composers, is that Mozart didn't have Beethoven as an influence.

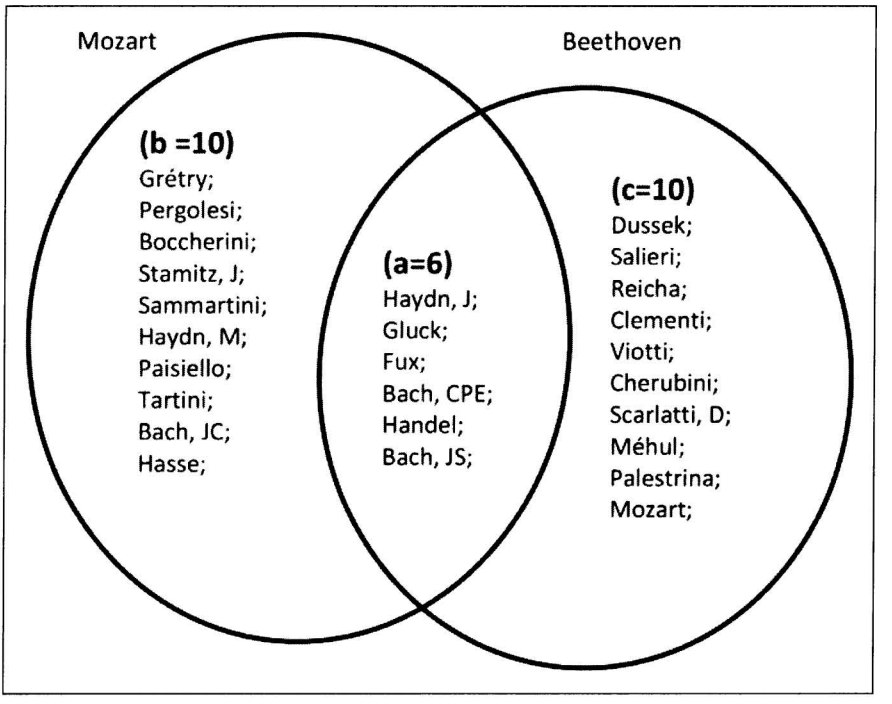

Figure 2. Common and distinct influences of Beethoven and Mozart.
Notes: Subset *a* represents "intersection set" common influences, whereas *b* and *c* represent distinct influences on each composer alone.

biogeography, such interactions are generally examined by means of pairwise comparison of presence-absence of taxa between given areas. See Cheetham and Hazel (1969) and Hayek (1994) for good surveys of these studies and methods used to develop measures of association (also named in a somewhat interchangeable way "similarity," "resemblance," or "matching").

The rest of the article develops the methodology as related to classical music composers, provides a sample of the results, and concludes with suggestions as to possible extensions.

METHOD

The *CMN* establishes, for 500 composers, two types of series of influences:

1. a list of composers who have influenced a subject composer; and
2. a complementary list of composers on which a subject composer had influence.

In this article, we limit our focus to the first type of series of influences, mostly because many subject composers could not be determined to have had significant influence on any later ones (thus presenting severe statistical comparison difficulties). The following set notation is introduced: C is the set of 500 composers in the database, I_i is the set of composers who influenced composer i, and I_j is the set of composers who influenced j. Figure 3 gives a representation of this formalization. For any pair of composers (i,j) for $i, j \in C$, $I_i \cap I_j = CI_{i,j}$ is the set of composers that both influenced i and j; $I_i - I_i \cap I_j = I_{i,-j}$ is the set of composers that influenced i but not j; $I_j - I_i \cap I_j = I_{j,-i}$ is the set of composers that influenced j but not i. Finally, $DI_{i,j} = I_{i,-j} \cup I_{j,-i}$ is the set of composers that influenced either i or j but not both. Note that:

1. $I_{i,-j} \cap I_{j,-i} = \phi$ (the intersection of $I_{i,-j}$ and $I_{j,-i}$ is an empty set); and
2. $DI_{i,j} \cap CI_{i,j} = \phi$ (if a composer influenced either i or j but not both, this composer cannot have influenced both i and j).

An alternative way to organize influence data is by using a count table (e.g., Hayek, 1994), as given in Table 1. The presence/absence dimension with respect to composer i keeps track of whether a specific composer is or is not in the list of composers who influenced i. The same is done with respect to composer j. Then, for example, a in Table 1 represents the number of composers who influenced both composers i and j; b is the number of composers who influenced i but not j; c is the number of composers who influenced j but not i; and finally, d is the number of composers in the database that did not influence i and/or j. Using the cardinal of a set as a notation for the element count in the set, the relation between Table 1 and Figure 3 becomes straightforward. We have that: $a = Card\,(I_i \cap I_j) = Card\,(CI_{i,j})$; $b = Card\,(I_i - I_i \cap I_j) = Card\,(I_{i,-j})$; $c = Card\,(I_j - I_i \cap I_j) = Card\,(I_{j,-i})$; $n = Card\,(c)$; and $d = Card\,(C - I_i \cup I_j) = Card\,(C) - (Card\,(CI_{i,j}) + Card\,(I_{i,-j}) + Card\,(I_{j,-i})) = n - (a + b + c)$. Other obvious links between Table 1 and Figure 3 are as follows: $a + c = Card\,(I_j)$; $a + b = Card\,(I_i)$; $b + d = n - (a + c) = Card\,(C - I_j)$; $c + d = n - (a + b) = Card\,(C - I_i)$.

Table 1. 2 × 2 Frequency Table Using Counts
(see text for explanation)

		Composer j		
		Presence	Absence	Total
Composer i	Presence	a	b	$a + b$
	Absence	c	d	$c + d$
	Total	$a + c$	$b + d$	n

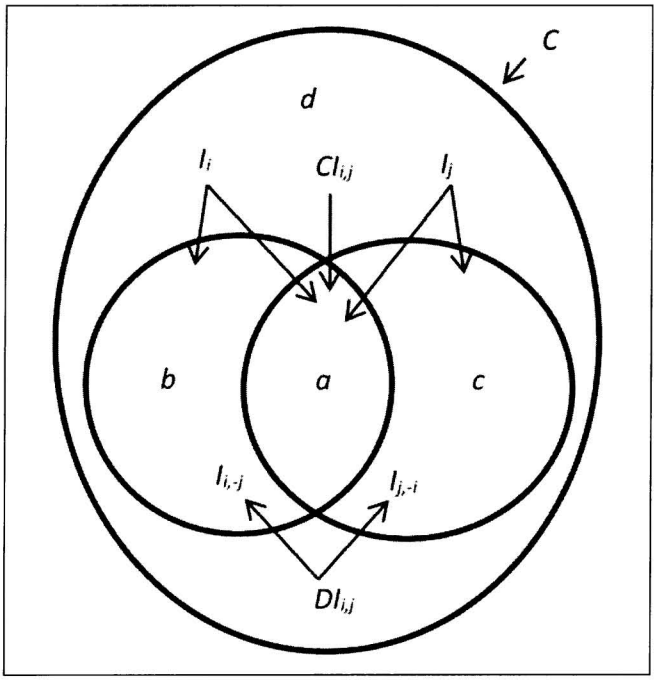

Figure 3. Set diagram (see text for explanation).

Dozens of measures of association have been studied in the literature, such as the first and second Kulczynski coeffecients (1927), the Jaccard coefficient (1901), the Dice coefficient (1945), the Simpson coefficient (1943), the binary distance coefficient (Sneath, 1968), and the binomial index of dispersion χ^2 statistic (Potthoff & Whittinghill, 1966). An initial problem faced was which one or ones to use in the present context. After an investigation of some of their relative qualities (see Appendix 1) we decided to concentrate on three specific similarity indices for composers *i* and *j*: Jaccard (1901), Smith (1983), and the binomial index of dispersion (Potthoff & Whittinghill, 1966). Given the notation underlying Table 1 and/or Figure 3 these are given by, respectively:

$$SI_{i,j}|_{Jaccard} = \frac{a}{a+b+c} = \frac{Card\,(CI_{i,j})}{Card\,(CI_{i,j}) + Card\,(I_{i,-j} \cup I_{j,-i})} =$$

$$\frac{Card\,(CI_{i,j})}{Card\,(CI_{i,j}) + Card\,(DI_{i,j})} \qquad (1)$$

$$\begin{aligned} SI_{ij/Smith} &= a-((a+b)-a)-((a+c)-a)=a-(b+c) \\ &= Card(I_i \cap I_j)-Card(I_i - I_i \cap I_j)-Card(I_j - I_i \cap I_j) \\ &= Card(CI_{i,j})-(Card(I_{i,-j}))-(Card(I_{j,-i}))=Card(CI_{i,j})-Card(DI_{i,j}) \end{aligned} \quad (2)$$

$$SI_{i,j}|_{BID} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (3)$$

The numerator in $SI_{i,j/Jaccard}$ ("a"), gives the number of common influences of i and j. For example, given that Mozart and Beethoven were both influenced by Joseph Haydn, then, it is likely that they acquired compositional characteristics, through the influence of Haydn, that increased similarity of their own compositional styles. The denominator includes the number of distinct influences for i and j ($b + c$). Unlike the numerator, distinct influences tend to differentiate composers. If ($b + c$) is large, then both composers were influenced by "non-common" influences, tending to make them less similar. On the other hand, the smaller the term ($b + c$), the smaller the distance between the compositional styles of i and j. Hence, a small denominator and/or a large numerator tend to decrease dissimilarities and increase similarities. In order to normalize the ratio between 0 and 1, the denominator also includes the term $a = Card\ (CI_{i,j})$. The rationale for this term in the denominator is obvious when we think of comparing a composer with himself. In this case, having an index value of 1 is a natural benchmark as both "objects" of comparison are exactly similar. For example, J. S. Bach was influenced by 22 composers (see Figure 1). Then we would get when Bach is compared to himself, that:

$$SI_{Bach,Bach}|_{Jaccard} = \frac{Card\ (CI_{Bach,Bach})}{Card\ (CI_{Bach,Bach})+Card\ (DI_{Bach,Bach})} = \frac{22}{22+0} = 1$$

Two composers i, j with no common influence would get an index value of $SI_{i,j} = 0$ (because the numerator would be zero). In general we have for any pair of composers (i,j) for $i, j \in C$, that $0 \leq SI_{i,j}|_{Jaccard} \leq 1$.

The Jaccard index is, historically, perhaps the most familiar of measures of association. It has been used in hundreds of ecological and biogeographic studies over a more than 100-year period. It may be considered a "typical" proportional measure of similarity.

The Smith index was proposed by Smith (1983) to study the similarities among the mammal faunas depicted within a 10-region classification system. Smith's analysis was based on an entropy maximization procedure, and proportional index data are not usually recommended as input under these circumstances (Daniell, 1991). As applied to the composers problem, in equation (2), we subtract

from the number of composers who both influenced i and j the total number of composers who influenced i but not j and those composers who influenced j but not i. As applied to Mozart and Bach, the similarity score would be given by: $2 - (16 - 2) - (22 - 2) = 2 - 34 = -32$. Note that with the Smith index, when comparing Bach with himself we would get $22 - 0 = 22$ (that is, we get the number of influences on Bach as given in the database). As $Card\ (CI_{i,j})$ is typically much smaller than $Card\ (DI_{i,j})$, almost all numbers we will generate for any pair of i,j where $i \neq j$ will be negative.

Finally, the binomial index of dispersion is based on the χ^2 statistic. In the present instance this is a highly desirable characteristic, as it takes into account varying numbers of composer influences from composer to composer, across the whole data set. Some intuition for the formula in equation (3) may be provided as follows. If no association exists between composers i and j, an equal proportion of influences on composer i in the overall database C should be found among the set of composers who influenced j and the set of composers who did not influence j. That is,

$$\frac{Card\,(I_i)}{Card\,(C)} = \frac{Card\,(I_i \cap I_j)}{Card\,(I_j)} = \frac{Card\,(I_{i,-j})}{Card\,(C) - Card\,(I_j)} \tag{4}$$

Figure 3 may be useful in interpreting the above expression. The first term represents the proportion of the composers who influenced i in the entire database of composers, C. The second term is the proportion of composers who influenced i among all those who influenced j. Finally, the last term gives the proportion of composers who influenced i (but not j) among all those who did not influence j. Assuming for example that i = J. S. Bach and j = Mozart, we say that there is no association between Bach and Mozart in the case where, say, 5% of the total composers in the database (500) have influenced Bach (the first term) and then, when observing influences on Mozart, we find that 5% of the composers who influenced Mozart have also influenced Bach (the second term), and also that 5% of those composers who did not influence Mozart have influenced Bach (the third term). However, a positive association between Bach and Mozart is inferred if we find that the first, second and third terms have values of, say, 5%, 9%, and 1%.

Using the notation in Table 1, we can rewrite equation (4) as: $(a + b)/n = a/(a + c) = b/(b + d)$, so that $a = (a + b)\ (a + c)/n$; that is, when two composers are independent (lack of association), the proportion or frequency of joint influences (a/n) is equivalent to the product of the proportions $(a + b)/n$ and $(a + c)/n$ (that is, the proportion of composers in the database who influenced i and the proportion of composers who influenced j). If the observed frequency is greater than the one expected under independence, then the two composers may be said to

be positively associated. Thus, if composers i and j are associated, then: $a \neq (a + b)(a + c)/n$, and the difference could be written as:

$$D = a - (a + b)(a + c)/n = (a/n)(n - a - b - c) - bc/n = (ad - bc)/n \qquad (5)$$

This term D, or some variation of it, is found in the formula for calculating the usual chi-square statistic (equation 3) and all of its monotonically related statistics.

RESULTS

A routine has been programmed to generate $SI_{i,j}$ for any pair of composers (i,j) for $i, j \in C$. As there are 500 composers in C, 250,000 statistics were calculated for each of the three indices. Selecting the 10 most significant composers (according to the *CMN*), we can report the composers most similar overall to each of these 10 composers. We can also report which composers (among these 10) are most similar within this select group. Tables 2-4 show the top-five most similar composers to these 10 famous composers, according to the Jaccard, Smith, and binomial index of dispersion similarity measures. Tables 5-7 show the parallel scores for the 10 famous composers among themselves, while Tables 8 and 9 give the relevant terms ("a" and "$b + c$") corresponding to Tables 5-7. Figure 4 shows a network representation of these 10 composers with respect to each other.

In Table 2, we see for example that the top-five most similar composers (again, with respect to their influences) to J. S. Bach are, in order, Telemann, Zelenka, Fux, Vivaldi, and Handel. We also see that J. Haydn is very similar to Mozart, and that J. C. Bach and Salieri are also very similar to him in this regard. In Table 3, the parallel rankings with respect to J. S. Bach are, in order, Telemann, Fux, Vivaldi, Pachelbel, and Albinoni. Table 4 produces the same order of composers with respect to J. S. Bach as does Table 3. In fact, the rankings produced by the Jaccard index, the Smith index, and the binomial index of dispersion are in general quite similar to one another, as long as one concentrates on the top 20 or 30 rankings (see further discussion in Appendix 1). This is also true of the several other indices we investigated. Of the 10 or so we looked into, the Smith index and the binomial index of dispersion give the most similar results, at least for the highest 20 or 30 rankings.

From column 1, "Bach, J. S." in Tables 5-8, we see that $S_{Handel;Bach} = 0.172$ in Table 5, which is the ratio $a/(a+(b+c)) = 5/(5 + 24)$, where $a = 5$ is given in Table 8 and $b+c = 24$ in Table 9. Observe that (in this column), the composer most similar to J. S. Bach is Handel. The major composer most similar to Mozart is J. Haydn ($SI = 0.350$). Schubert and J. Haydn are equivalently similar to Beethoven according to our index ($SI = 0.286$). Beethoven is the most similar

Table 2. Top-5 Most Similar Composers to 10 Famous Composers (Jaccard Index, Jaccard, 1901)

Most similar	Bach, JS	Most similar	Mozart	Most similar	Beethoven	Most similar	Schubert	Most similar	Brahms
35. Telemann	0.261	9. Haydn, J	0.350	131. Hummel	0.333	212. Mendelssohn-Hensel	0.385	212. Mendelssohn-Hensel	0.353
343. Zelenka	0.192	105. Boccherini	0.294	17. Mendelssohn	0.318	266. Reicha	0.364	254. Loewe	0.313
314. Fux	0.182	68. Gluck	0.263	156. Clementi	0.294	23. Rossini	0.357	13. Schumann, R	0.300
22. Vivaldi	0.174	119. Bach, JC	0.235	4. Schubert	0.286	17. Mendelssohn	0.333	47. Bruckner	0.300
8. Handel	0.172	228. Salieri	0.235	9. Haydn, J	0.286	228. Salieri	0.333	21. Dvorák	0.286

Most similar	Wagner	Most similar	Verdi	Most similar	Handel	Most similar	Haydn, J	Most similar	Chopin
326. Nicolai	0.350	102. Glinka	0.385	35. Telemann	0.267	2. Mozart, WA	0.350	17. Mendelssohn	0.368
102. Glinka	0.333	326. Nicolai	0.308	96. Scarlatti, A	0.267	3. Beethoven	0.286	331. Berwald	0.357
56. Gounod	0.318	133. Boito	0.273	22. Vivaldi	0.214	68. Gluck	0.267	13. Schumann, R	0.353
99. Meyerbeer	0.300	335. Mercadante	0.273	108. Couperin, F	0.200	119. Bach, JC	0.231	326. Nicolai	0.333
29. Berlioz	0.273	6. Wagner	0.250	1. Bach, JS	0.172	228. Salieri	0.231	102. Glinka	0.313

Table 3. Top-5 Most Similar Composers to 10 Famous Composers (Smith Index, Smith, 1983)

Most similar	Bach, JS	Most similar	Mozart	Most similar	Beethoven	Most similar	Schubert	Most similar	Brahms
35. Telemann	−11.000	9. Haydn, J	−6.000	131. Hummel	−6.000	212. Mendelssohn-Hensel	−3.000	212. Mendelssohn-Hensel	−5.000
314. Fux	−14.000	105. Boccherini	−7.000	156. Clementi	−7.000	266. Reicha	−3.000	254. Loewe	−6.000
22. Vivaldi	−15.000	68. Gluck	−9.000	17. Mendelssohn	−8.000	23. Rossini	−4.000	13. Schumann, R	−8.000
111. Pachelbel	−16.000	119. Bach, JC	−9.000	417. Dussek	−8.000	228. Salieri	−4.000	47. Bruckner	−8.000
135. Albinoni	−16.000	228. Salieri	−9.000	4. Schubert	−9.000	367. Taneyev	−4.000	226. Rheinberger	−8.000

Most similar	Wagner	Most similar	Verdi	Most similar	Handel	Most similar	Haydn, J	Most similar	Chopin
326. Nicolai	−6.000	102. Glinka	−3.000	35. Telemann	−7.000	2. Mozart, WA	−6.000	331. Berwald	−4.000
102. Glinka	−7.000	133. Boito	−5.000	96. Scarlatti, A	−7.000	68. Gluck	−7.000	13. Schumann, R	−5.000
56. Gounod	−8.000	326. Nicolai	−5.000	22. Vivaldi	−8.000	119. Bach, JC	−7.000	17. Mendelssohn	−5.000
99. Meyerbeer	−8.000	335. Mercadante	−5.000	65. Scarlatti, D	−8.000	228. Salieri	−7.000	326. Nicolai	−5.000
29. Berlioz	−10.000	142. Delibes	−6.000	94. Corelli	−8.000	428. Mayr	−7.000	102. Glinka	−6.000

Table 4. Top-5 Most Similar Composers to 10 Famous Composers
(Chi-Square Statistics from the Binomial Index of Dispersion, Potthoff & Whittinghill, 1966)

Most similar	Bach, JS	Most similar	Mozart	Most similar	Beethoven	Most similar	Schubert	Most similar	Brahms
35. Telemann	111.595	9. Haydn, J	132.625	131. Hummel	135.306	266. Reicha	179.252	212. Mendelssohn-Hensel	156.047
314. Fux	87.610	105. Boccherini	125.891	156. Clementi	125.891	212. Mendelssohn-Hensel	158.136	254. Loewe	152.778
22. Vivaldi	68.623	119. Bach, JC	96.168	417. Dussek	121.976	228. Salieri	142.080	451. Franz	121.976
111. Pachelbel	65.575	228. Salieri	96.168	17. Mendelssohn	110.523	367. Taneyev	142.080	226. Rheinberger	106.691
135. Albinoni	65.575	68. Gluck	92.295	228. Salieri	96.168	23. Rossini	137.394	13. Schumann, R	106.278

Most similar	Wagner	Most similar	Verdi	Most similar	Handel	Most similar	Haydn, J	Most similar	Chopin
326. Nicolai	170.385	102. Glinka	151.827	35. Telemann	90.827	2. Mozart, WA	132.625	331. Berwald	156.303
102. Glinka	147.617	326. Nicolai	110.140	96. Scarlatti, A	90.827	3. Beethoven	95.727	17. Mendelssohn	138.412
99. Meyerbeer	145.749	133. Boito	109.632	65. Scarlatti, D	81.660	68. Gluck	86.335	326. Nicolai	132.809
56. Gounod	129.912	335. Mercadante	109.632	94. Corelli	81.660	119. Bach, JC	78.420	13. Schumann, R	132.759
29. Berlioz	106.728	289. Giuliani	98.394	329. Geminiani	81.660	228. Salieri	78.420	324. Kuhlau	119.477

Table 5. Composer Similarity Index S_{ij}, (Jaccard, 1901) between Pairs of Composers (i,j)

i / j	1. Bach, JS	2. Mozart	3. Beethoven	4. Schubert	5. Brahms	6. Wagner	7. Verdi	8. Handel	9. Haydn, J	10. Chopin
1. Bach, JS	—									
2. Mozart	0.056	—								
3. Beethoven	0.056	0.231	—							
4. Schubert	0.031	0.227	0.286	—						
5. Brahms	0.118	0.103	0.231	0.227	—					
6. Wagner	0.024	0.059	0.161	0.240	0.200	—				
7. Verdi	0.032	0.000	0.040	0.105	0.040	0.250	—			
8. Handel	0.172	0.000	0.000	0.000	0.037	0.000	0.000	—		
9. Haydn, J	0.065	0.350	0.286	0.158	0.125	0.069	0.000	0.045	—	
10. Chcpin	0.000	0.036	0.160	0.263	0.208	0.269	0.211	0.000	0.091	—

Table 6. Composer Similarity Index S_{ij} (Smith, 1983) between Pairs of Composers (i,j)

i / j	1. Bach, JS	2. Mozart	3. Beethoven	4. Schubert	5. Brahms	6. Wagner	7. Verdi	8. Handel	9. Haydn, J	10. Chopin
1. Bach, JS	—									
2. Mozart	–32	—								
3. Beethoven	–32	–14	—							
4. Schubert	–30	–12	–9	—						
5. Brahms	–26	–23	–14	–12	—					
6. Wagner	–39	–30	–21	–13	–18	—				
7. Verdi	–29	–26	–23	–15	–23	–12	—			
8. Handel	–19	–28	–28	–23	–25	–32	–22	—		
9. Haydn, J	–27	–6	–9	–13	–18	–25	–21	–20	—	
10. Chopin	–35	–26	–17	–9	–14	–12	–11	–25	–18	—

Table 7. Composer Similarity Index S_{ij}, (Chi-Square Statistics from the Binomial Index of Dispersion, Potthoff & Whittinghill, 1966) between Pairs of Composers (*i,j*)

i / j	1. Bach, JS	2. Mozart	3. Beethoven	4. Schubert	5. Brahms	6. Wagner	7. Verdi	8. Handel	9. Haydn, J	10. Chopin
1. Bach, JS	—									
2. Mozart	2.58	—								
3. Beethoven	2.58	62.78	—							
4. Schubert	0.59	64.83	95.73	—						
5. Brahms	16.68	12.90	62.78	64.83	—					
6. Wagner	0.02	3.11	31.96	74.83	48.31	—				
7. Verdi	0.76	0.34	1.52	15.03	1.52	83.33	—			
8. Handel	40.59	0.41	0.41	0.28	1.05	0.51	0.25	—		
9. Haydn, J	5.08	132.62	95.73	32.86	21.04	5.89	0.23	2.15	—	
10. Chopin	0.61	0.87	32.75	81.57	53.57	86.36	56.36	0.33	10.78	—

Table 8. Number of Common Influences ("*a*") between Composer Pair (*i,j*)
(Number of Influences on a Composer *i* on the Diagonal)

i / j	1. Bach, JS	2. Mozart	3. Beethoven	4. Schubert	5. Brahms	6. Wagner	7. Verdi	8. Handel	9. Haydn, J	10. Chopin
1. Bach, JS	22	2	2	1	4	1	1	5	2	0
2. Mozart	2	16	6	5	3	2	0	0	7	1
3. Beethoven	2	6	16	6	6	5	1	0	6	4
4. Schubert	1	5	6	11	5	6	2	0	3	5
5. Brahms	4	3	6	5	16	6	1	1	3	5
6. Wagner	1	2	5	6	6	20	6	0	2	7
7. Verdi	1	0	1	2	1	6	10	0	0	4
8. Handel	5	0	0	0	1	0	0	12	1	0
9. Haydn, J	2	7	6	3	3	2	0	1	11	2
10. Chopin	0	1	4	5	5	7	4	0	2	13

Table 9. Number of Distinct Influences ("$b + c$") between Pairs of Composers (i,j)

i / j	1. Bach, JS	2. Mozart	3. Beethoven	4. Schubert	5. Brahms	6. Wagner	7. Verdi	8. Handel	9. Haydn, J	10. Chopin
1. Bach, JS	0	34	34	31	30	40	30	24	29	35
2. Mozart	34	0	20	17	26	32	26	28	13	27
3. Beethoven	34	20	0	15	20	26	24	28	15	21
4. Schubert	31	17	15	0	17	19	17	23	16	14
5. Brahms	30	26	20	17	0	24	24	26	21	19
6. Wagner	40	32	26	19	24	0	18	32	27	19
7. Verdi	30	26	24	17	24	18	0	22	21	15
8. Handel	24	28	28	23	26	32	22	0	21	25
9. Haydn, J	29	13	15	16	21	27	21	21	0	20
10. Chopin	35	27	21	14	19	19	15	25	20	0

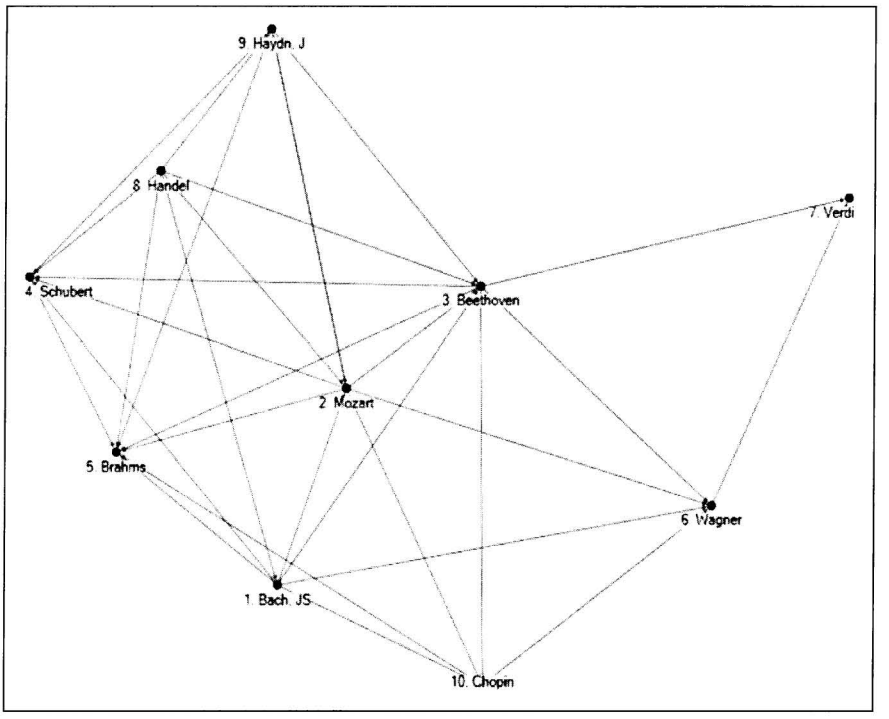

Figure 4. Network of the similarities of the 10 most important composers among themselves.

to Brahms ($SI = 0.231$). The ranking generated by the Smith index leads again to a ranking generally consistent with the Jaccard index and the binomial index of dispersion.[2] Finally, note as expected that the similarity matrices in Tables 5-7 are symmetric ($SI_{i,j} = SI_{j,i}$).

We believe that even a casual look at the results here produces some confidence in this approach. It is true that it lacks an independent measure

[2] A note on the chi-square statistical interpretation of the values given in Table 7 is useful. As shown in this table, the chi-square similarity statistic for the pair of composers Bach and Handel is 40.59. In the dual outcome of presence/absence in Table 1, the degree of freedom is 1 and the critical value at a 5% significance level is thus 3.84. Because $40.59 > 3.84$, we reject the null hypothesis of no association between both composers in favor of the alternative that Bach and Handel are statistically significantly similar. For Bach and Mozart, however, the similarity index is 2.58; thus, we cannot reject the null hypothesis of no association between these two composers.

of reliability (apart from what hundreds of music historians and biographers have concluded more subjectively); however, a future analysis (as alluded to in the Conclusion section) will attempt to confirm its results by applying a second series of measures of similarity that are independent of the influences data.

Two caveats: the approach apparently works well, as long as two conditions are met:

1. the composers being compared have some minimum number of identified influences, perhaps four or five; and
2. attention is restricted to the top 10 or 20 most similar composers identified in any given instance.

Both problems will be reduced as more becomes known about the more obscure composers and their influences.

CONCLUSION

In this article we have addressed similarity of style through the study of common influences. As stated before, we have avoided the matter of what to do about cumulative influence. For example, because Grétry has influenced Cherubini, who himself has influenced Beethoven, then, to some extent, Beethoven might have been (indirectly) influenced by Grétry himself (through Cherubini). If Grétry is a direct influence of Mozart and an indirect influence of Beethoven, then Grétry might be in the "*a*" count/list instead of the "*b*" list, as he is now. Then should we also consider the fact that Gluck himself has both influenced Grétry and Beethoven?

There are various ways this matter could be studied. One which was initially considered here was to extend the compilations on a "second order" basis. Thus, instead of merely counting the "first order" influences as listed at the *CMN*, one could extend the count to capture all instances in which the influences of the subject composer's influences included names on that subject composer's list of influences. A more direct "lineage" approach could also be calculated by following ahead, or backwards, individual lines of influence. This type of thinking leads us to the concept of "genealogic" influence trees. This is left for further research.

In a future work we will examine how much, if any, additional explanation can be added when "ecological measures" (i.e., other composer characteristics such as time period, school association, instrumentation emphases, etc.) and general influences (e.g., "jazz" or "folk music") are also taken into account. Hence, the results described here should best be viewed as interim figures.

APPENDIX 1

Because, as mentioned, dozens of measures of association have been formulated, an immediate question became which one or several should be applied here. We investigated nine initially, and found that the results, at least as pertaining to the calculation of the most similar figures to each subject composer, were rather similar.[3] In this instance, however, we had a piece of information missing from other such analyses: the composer rankings.

All other things being equal, there is no reason why those composers showing the closest affinities of influence to a subject composer should be of higher or lower rank than the mean value of 250 for the whole lot of 500 composers. Thus, any consistent bias of this type represents a flag that there is something about the index that over- or under-estimates certain subsets of the entire list. This may be because the number of noted influences for certain subject composers is either very high or very low, or involves a small or large intersection set of common influences with other particular composers. Thankfully, such biases show up more at the negative end (most dissimilar composers in terms of common influences) than they do at the positive end, but there does turn out to be an observable bias even at the latter.

As a further complication, the mean rank of the list's composers active at different particular times through history changes. For example, the 25 composers on the list active in 1800 carry a mean rank of 292.1, whereas in 1880 this number drops to 196.2 for the 75 composers on the list active at that time. This problem can, however, be minimized to a significant extent by studying patterns of bias based on a random sample of composers across the several centuries represented.

Two quick studies were made to check for possible systematic biases. In the first, the top 25 influences on 20 major composers were listed out in rank order, and for each of the three indices (Jaccard, Smith, binomial). The composer ranks in the main *CMN* site were then substituted for the similarity ranks in each of the 60 lists. For each such list a mean value was taken. Remember, it was anticipated above that an unbiased placement of composers, and their ranks, should produce a mean value of 250. Here, the problem was complicated by a large number of ties, but it can be said, roughly, that the Jaccard index produced a mean value of less than 200, the Smith index just under 250, and the binomial index a bit over 210.

The second study sampled 15 composers from the main *CMN* rankings, starting with number 10 (Chopin), and continuing with each following 20th composer

[3] These measures of association are the first and second Kulczynski coefficients (1927), the Jaccard coefficient (1901), the Dice coefficient (1945), the Smith coefficient (1983), and a transformation, the Simpson coefficient (1943), the Sneath binary distance coefficient (1968), and the χ^2 statistic or binomial index of dispersion (Potthoff & Whittinghill, 1966).

(including number 30, Gershwin, number 50, Bernstein, and so forth on to number 290, Duparc). For each of these composers (and each index) the similarities ranks were sampled (each 20th rank), and again connected to the main composer ranks on the *CMN* site. Sixty Pearson *r* correlation coefficients between the similarities ranks and the respective composer ranks were then calculated, under the null hypothesis that a fully unbiased data set should produce *r* values of zero in all instances. Again, ties and sampling problems caused some difficulties and valid *r* values for the Jaccard data could not be calculated. The mean of the *r* values for the Smith index data was roughly –.400, and for the binomial index .332. Thus there is a tendency for high-ranking composers to be lowly placed in the similarities listings through the Smith index, and just the opposite in the binomial index.

Going into the reasons for these discrepancies here in any detail would draw us off-subject, but in general it can be said that most of the problems lie with the similarities rankings at the low end; as mentioned earlier the top 10- or 20-ranked similarities scores remain fairly consistent from one index to the next. For example, the Smith index tends to rank composer pairings of individuals with many noted influences, but none in common, at the bottom of the similarities listings; thus the high negative mean correlation.

Our reasons for treating these three particular indices have to do both with seeking appropriate indicators, and with providing scores that can be used in further studies. For those interested in a best overall *ranking* of similarly-influenced composers, the binomial index probably serves best, being a chi-square statistic that accounts for varying size of sample. The Smith index provides non-proportionalized scores suitable for further processing through techniques such as entropy maximization or multidimensional scaling. The proportionalized scores of the Jaccard index may be directly compared to analogous values within the same data set, or outside of it.

REFERENCES

Anonymous Prof. (2008). 3D tour of classical music history. Retrieved June 26, 2013 from http://www.visualcomplexity.com/vc/project.cfm?id=570

Cheetham, A. H., & Hazel, J. E. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology, 43*, 1130-1136. doi: 10.2307/1302424

Collins, N. (2010). Computational analysis of musical influence: A musicological case study using MIR tools. *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (pp. 177-181). [s.l.]: International Society for Music Information. Retrieved June 26, 2013 from http://www.mirlab.org/conference_papers/International_Conference/ISMIR%202010/ISMIR_2010_papers/ismir2010-32.pdf

Daniell, G. J. (1991). Of maps and monkeys: An introduction to the maximum entropy method. In B. Buck & V. A. Macaulay (Eds.), *Maximum entropy in action* (pp. 1-18). Oxford, UK: Oxford University Press.

de Carvalho, A. D., & Batista L. V. (2012). Composer classification in symbolic data using PPM. In *11th International Conference on Machine Learning and Applications (ICMLA), 2012* (pp. 345-350). Los Alamitos, CA: IEEE Computer Society. doi: 10.1109/ICMLA.2012.176

de Leon, P. P. (2002). Musical style identification using self-organising maps. In C. Busch (Ed.), *Proceedings, Second International Conference on WEB Delivering of Music* (pp. 82-89). Los Alamitos, CA: IEEE.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*, 297-302.

Fazekas, G., Raimond, Y., Jacobson, K., & Sandler, M. (2010). An overview of semantic web activities in the OMRAS2 project. *Journal of New Music Research, 39*, 295-311. doi: 10.1080/09298215.2010.536555

Filippova, D., Fitzgerald, M., Kingsford, C., & Benadon, F. (2012). Dynamic exploration of recording sessions between jazz musicians over time. In *Privacy, Security, Risk and Trust (PASSAT 2012), and 2012 International Conference on Social Computing (SocialCom)* (pp. 368-376). Los Alamitos, CA: IEEE Computer Society Conference Publishing Services. doi: 10.1109/SocialCom-PASSAT.2012.78

Georges, P., & Seçkin, A. (2013). Black notes and white noise: A hedonic approach to auction prices of classical music manuscripts. *Journal of Cultural Economics, 37,* 33-60.

Hayek, L.-A. C. (1994). Analysis of amphibian biodiversity data. In W. R. Heyer, M. A. Donnelly, R. W. McDiarmid, L.-A. C. Hayek, & M. S. Foster (Eds.), *Measuring and monitoring biological diversity: Standard methods for amphibians* (pp. 207-269). Washington, DC: Smithsonian Books.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. de la Societé Vaudoise de la Science Naturelle, 37*, 547-579.

Jacobson, K. (2011). *Connections in music*. Thesis, Centre for Digital Music, Queen Mary University of London. Retrieved June 26, 2013 from http://kurtisrandom.com/files/kurtj-thesis.pdf

Kaliakatsos-Papakostas, M. A., Epitropakis, M. G., & Vrahatis, M. N. (2010). Musical composer identification through probabilistic and feedforward neural networks. In C. Di Chio (Ed.), *EvoApplications 2010, Part II, LNCS 6025* (pp. 411-420). Berlin, Germany: Springer. doi: 10.1007/978-3-642-12242-2_42

Kulczynski, S. (1927). Zespoly roslin w Pieninach. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres, ser. B: Sciences Naturelles, 1*(suppl. 2), 57-203.

Kurt Jx (2008). Wagner is center of the Universe? Retrieved June 26, 2013 from http://kurtisrandom.blogspot.com/2008/02/wagner-is-center-of-universe.html

Logan, B., Ellis, D. P. W., & Berenzweig, A. (2003). Toward evaluation techniques for music similarity. *SIGIR 2003: Workshop on the Evaluation of Music Information Retrieval Systems, 1 August 2003, Toronto, Canada*. Retrieved November 14, 2013 from http://academiccommons.columbia.edu/catalog/ac:148882

Mostafa, M. M., & Billor, N. (2009). Recognition of western style musical genres using machine learning techniques. *Expert Systems with Applications, 36,* 11378-11389. doi: 10.1016/j.eswa.2009.03.050

OMRAS2, n.d. *Classical music universe.* Retrieved June 26, 2013 from http://www.omras2.org/ClassicalMusicUniverse

Potthoff, R. F., &Whittinghill, M. (1966). Testing for homogeneity. I. The binomial and multinomial distributions. *Biometrika, 53,* 167-182. doi: 10.1093/biomet/53.1-2.167

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science, 241*, 1-31. doi: 10.2475/ajs.241.1.1

Smith, C. H. (1983). A system of world mammal faunal regions. I. Logical and statistical derivation of the regions. *Journal of Biogeography, 10*, 455-466.

Smith, C. H. (2000). *The classical music navigator*. Retrieved June 26, 2013 from http://people.wku.edu/charles.smith/music/

Sneath, P. H. A. (1968). Vigour and pattern in taxonomy. *Journal of General Microbiology, 54,* 1-11. doi: 10.1099/00221287-54-1-1

Vieira, V., Fabbri, R., Travieso, G., Oliveira, O. N. Jr., & da Fontoura Costa, L. (2012). A quantitative approach to evolution of music and philosophy. *Journal of Statistical Mechanics: Theory and Experiment, 2012,* P08010. doi: 10.1088/1742-5468/2012/08/P08010

West, K., & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing, Volume 2007*, Article ID 24602, 10 pages. doi: 10.1155/2007/24602

Direct reprint requests to:

Charles H. Smith
University Libraries
1906 College Heights Blvd.
Western Kentucky University
Bowling Green, KY 42101
e-mail: charles.smith@wku.edu