# The Quest for Ground Truth in Musical Artist Similarity

Daniel P.W. Ellis
Columbia University
New York NY U.S.A.
dpwe@ee.columbia.edu

Brian Whitman
MIT Media Lab
Cambridge MA U.S.A.
bwhitman@media.mit.edu

Adam Berenzweig
Columbia University
New York NY U.S.A.
alb63@columbia.edu

Steve Lawrence
NEC Research Institute
Princeton NJ U.S.A.
lawrence@necmail.com

## ABSTRACT

It would be interesting and valuable to devise an automatic measure of the similarity between two musicians based only on an analysis of their recordings. To develop such a measure, however, presupposes some 'ground truth' training data describing the actual similarity between certain pairs of artists that constitute the desired output of the measure. Since artist similarity is wholly subjective, such data is not easily obtained. In this paper, we describe several attempts to construct a full matrix of similarity measures between a set of some 400 popular artists by regularizing limited subjective judgment data. We also detail our attempts to evaluate these measures by comparison with direct subjective similarity judgments collected via a web-based survey in April 2002. Overall, we find that subjective artist similarities are quite variable between users—casting doubt on the concept of a single 'ground truth'. Our best measure, however, gives reasonable agreement with the subjective data, and forms a useable stand-in. In addition, our evaluation methodology may be useful for comparing other measures of artist similarity.

## 1. INTRODUCTION

There is a strong appeal to the notion that the similarity between two artists can be somehow measured. It seems particularly obvious that the similarity between certain pairs of artists can be judged as greater than between other pairs. Even though the concept of a single numerical similarity score between every pair of a set of artists raises serious epistemological problems, being able to generate such a score would be very useful in music recommendation and organization applications, and several researchers have pursued variations of this idea. A typical goal would be an automatic system that uses examples of the music of two artists to generate a rating of their similarity. This raises the problem of assessing the *quality* of the automatic ratings, and/or choosing the *ideal outcomes* with which to train such a system.

The current paper seeks to address this last problem: can we come up with a quantitative set of similarity scores, for a limited range of artists, which are as close as possible to the 'ground truth' that we would wish for as the output of signal analysis based methods? We want the ground truth values to capture the subjective impressions of the average user, giving a continuously-valued similarity score for a large number of artist pairs, including, crucially, both similar and dissimilar pairs. Assuming this data existed, it could be used to train automatic algorithms by providing a set of 'target' ratings with which to set system parameters, and to assess the accuracy of the automatic systems by measuring how well their scores matched the ideal.

Before considering how such a set of ground truth values might be estimated, we need to examine some of the problems that beset from this idea:



- **Individual variation:** That people have individual tastes and preferences is central to the very idea of music and humanity. By the same token, subjective judgments of the similarity between specific pairs of artists are not consistent between listeners and may vary with an individual's mood or evolve over time. In particular, music that holds no interest for a given subject very frequently 'all sounds the same'.

- **Multiple dimensions:** The question of the similarity between two artists can be answered from multiple perspectives: Music may be similar or distinct in terms of genre, geographical origin, instrumentation, lyric content, historical timeframe, etc. While these dimensions are not independent, it is clear that different emphases will result in different artists. That both Paul Anka and Alanis Morissette are from Canada might be of paramount significance to a Canadian cultural nationalist, although another person might not find them at all similar.

- **Asymmetry:** Defining a single similarity value for a pair of artists suggests that their similarity is symmetric, but, as discussed in [8] and elsewhere, subjective similarity is often asymmetric. We might say that the 90s LA pop musician Jason Falkner is similar to the Beatles, but we would be less likely to say that the Beatles are similar to Jason Falkner, not only because the Beatles recorded most of their music before Falkner was born, but also because the much better known Beatles serve as a *prototype*, in contrast to the specific instance of Falkner. Asymmetry is one of the issues that undermines a geometric (Euclidean) model of similarity, which is nonetheless a widely used assumption in similarity measures.

- **Variability and span:** Few artists are truly a single 'point' in any imaginable stylistic space, but undergo changes through their careers, and may consciously span multiple styles within a single album. Trying to define a single distance between any artist and widely-ranging long-lived musicians such as David Bowie or Prince seems unlikely to yield satisfactory results.

Despite these problems, we believe that there is utility to the idea that an 'average' set of similarity judgments, that would mostly agree with most people, could be constructed. In the remainder of the paper, we pursue this idea. Section 2 briefly reviews related prior work in music similarity. In section 3, we describe our general approach, and define the several different data sources and metrics we have developed for this task. Section 4 explains our evaluation procedure, in which an independent dataset was collected specifically to compare the success of each metric. Finally, in section 5, we discuss the results of our evaluation, and draw conclusions about the best practice for researchers interested in artist similarity.

## 2. PRIOR WORK

Computationally, musical similarity has been studied from the score level, the audio level, and the cultural level. Each type of study informs the next in hypothesis (that music can be modeled statistically and measured against other pieces) but not approach (where models

widely differ.) However, all have the same caveat: that different ideas of computationally- derived similarity cannot be compared to one another, because the methods are as of now lacking a ground truth.

At the score level (MIDI files, transcribed music or CSound scores) systems can extract style and similarity using the performance characteristics of the piece along with the key and frequently used progressions, where such feature extraction is discretized and definite. Any system trained to do genre or style detection can infer up a level to perform similarity computations by studying the posterior probabilities. In [4], various machine learning classifiers are trained on performance characteristics of the score, and in [2] three types of folk music were separated using a Hidden Markov Model. Recent work in [6] studies the cognitive background of melodic similarity from score data.

When considering the audio domain, spectral information has proven to be instructive but not the only feature necessary to infer acoustic similarity. A system trained on a song identification task (for copyright protection or query-by-example), such as [5], would need only the spectral information, but systems that need to understand what constitutes a similar piece of audio usually need help from higher-level extracted features. In [12], attempts are made to abstract the content from the style of the audio in a manner that could recognize "cover versions" of songs already in a database. Genre identification work undertaken in [9] aims to understand acoustic content enough to classify into a small set of related clusters. The idea of parsing audio with the intent of creating an "eigen-artist" trained to classify future work by the same artist (a specific form of similarity) was first undertaken in [10] and then improved on in [1] with more musical knowledge. Both genre and artist identifiers can claim to compute musical similarity, but both have the inherent advantage of a well-defined ground truth (in genre's case, the record industry's marketing-led genres, and in artists' case, the actual artist.)

Cultural similarity (in which the listener or collection of listeners define the similarity) can benefit from attempting to express innate non-acoustic and non-musical features about a specific piece of music. [11] defines *community metadata* concerning music as a feature vector that changes over time, reflecting the public's perception of an artist. (Their "Klepmit" and "OpenNap" datasets are used as similarity in this article.) Related work in [3] computes music recommendations based on similar artists found together in users' "favorite artist lists."

# 3. APPROACH

The basis for a 'ground truth' artist similarity measure must be the subjective judgments of music listeners, but problems arise when converting subjective opinions into quantitative values, and when extending sparse coverage to give similarity judgments between any pair from a large list of artists. In particular, while we can easily agree that the Backstreet Boys are very similar to N'Sync, judgments about dissimilar artists are less common and more difficult to quantify: how much are Backstreet Boys unlike Velvet Underground? How does that compare to their dissimilarity to Sade?

We have investigated several different basic sources for our subjective information, and several different mechanisms for 'regularizing' that information into a relatively comprehensive matrix of judgments between a large number of artists. Each measure is described in more detail below.

## 3.1 Measures

### *3.1.1 Artist Selection*

We chose 412 artists to be included in our evaluation space. The artists were chosen automatically as the most popular artists on a popular peer-to-peer network as of August, 2001 (see below for a more detailed description of the peer-to-peer data collection component.) Because of their selection criteria, the genre of the artists does not stray far from pop or rock, but has the advantage of being recognizable by almost any arbiter of current culture.

Each similarity measure described defines its output as a similarity matrix on the 412x412 artist space, where $S(a, b)$ is a continuous real-valued function describing the relation of artist $a$ to $b$. Some measures give distances rather than similarity; this distinction is unimportant for simple rankings (providing the correct sense is applied).

### *3.1.2 Erdös*

One promising data source is a published music guide, in which professional editors write brief descriptions for a large number of popular musical artists, often including a list of similar artists. We extracted the similar artist lists from the All Music Guide (www.amg.com), giving for each member of our 412 artist list an average of 5.4 similar artists also within the list (31 of the artists had no neighbors in the set, and were effectively excluded from this measure).

To convert these descriptions of the immediate neighborhood of each artist into a more extensive measure, we adopted the technique used among mathematicians to gauge their relationship to Paul Erdös : those who have co-authored papers with the prolific Hungarian mathematician have an Erdös number of 1; co-authoring with one of those authors will earn you an Erdös number of 2, and so on. (This principle is applied to movie actors in the game known as "Six degrees of Kevin Bacon").

The largest distance in our Erdös matrix is 13, corresponding to the maximally dissimilar pair "Miles Davis" and "Wade Hayes". Our construction of the Erdös measure is symmetric, since links between artists were treated as nondirectional. Erdös measures intrinsically obey the triangle inequality, since the distance between any two points cannot exceed the sum of the distances to a third point - since this sum describes a valid Erdös path.

Erdös distances are of course always integers, meaning that the distance measures are highly quantized. For any given source-target pairing of artists, there will likely be a number of other artists at exactly the same 'distance' from the source. This is clearly an artifact and can be a nuisance, for example when trying to construct a single, canonical ordered list.

### *3.1.3 Resistive Erdös*

An objection to the technique above might be that it is subject to the whims of the human experts who created the original lists of "similar artists". The criteria used to create the lists are not well-defined, and it is likely that no two experts would create the same lists. Furthermore, the expert's decisions about who to include or omit from each list becomes set in stone, because, for example, only the artists $B$ included in artist $A$'s list, or vice-versa, can have Erdös distance $d(A, B) = 1$. But what if there is another artist, $C$, that is very much like artist $A$, but it was overlooked by the expert? Assume further that the expert did note that both artists A and C are similar to several others $(D, E, F)$? In some cases it might seem reasonable that $d(A, C)$ should be even less than $d(A, B)$, because $A$ and $C$ share so many mutual intermediaries and thus must resemble one another.

This intuition is captured by the Resistive-Erdös measure. The desired property, namely that nodes connected by many alternative paths of length $l$ are more similar than nodes connected by only a single path of length $l$, can be modeled by electrical resistance in a network. Resistors connected in parallel add as reciprocals ($R_{eq} = \frac{1}{1/R_1+1/R_2}$), so the equivalent resistance between two nodes connected by multiple paths is less than the resistance of any

single path.

The Resistive-Erdös similarity measure between two artists is defined as the equivalent resistance between the nodes in the Erdös graph if each edge is a resistor of 1 ohm. An all-pairs version of the SPS (Series-Parallel-Star) tree algorithm [7] was used to compute the resistances, written recursively to avoid recomputing intermediate steps when computing resistances between all pairs of nodes in a network.

One problem with the measure is that it is biased towards popular artists (nodes with high degree in the Erdös graph) because the many alternate paths lower the total resistance. Attempts to compensate by using heavier resistance on edges incident to popular artists were not successful, but perhaps improvements can be made in the future.

### *3.1.4 OpenNap Peer-To-Peer Cultural Similarity*

Similarity can be inferred from observation: clusters of music generated from listening patterns are a direct measure of *cultural similarity* and can show relations between artists that could never come out of an edited list or the musical content. We used user preference data (user $i$ has artist $x$ in their collection) to generate a continuous matrix of similarity.

We retrieved user collection data from OpenNap, a popular music sharing service (we did not download any audio files). About 1.6 million user-to-song relations were retrieved, indicating that a user has a particular song in their collection. After processing the data for typos and misspellings, and removing unknown artists, we were left with about 400,000 user-to-song relations covering about 3,000 unique artists.

We define a collection as the set of artists a user had songs by on their shared folder. If two artists frequently occur together in user collections, we consider them similar via this measure of community metadata, since even if users are striving for *variety* in their collections, it is significant if they find variety in the same artists. We also define a collection count $\mathcal{C}(artist)$ which equals the number of users that have $artist$ in their set. $\mathcal{C}(a, b)$, likewise, is the number of users that have both artists $a$ and $b$ in their set.

One problem of this method is that extremely popular artists (such as Madonna) occur in a large percentage of users' collections, which down-weights similarity between lesser-known artists. We developed a scoring metric that attempts to alleviate this problem. Given two artists $a$ and $b$, where $a$ is more popular than $b$ (i.e., $\mathcal{C}(a) > \mathcal{C}(b)$), and a third artist $c$ who is the most popular artist in the set; $a$ and $b$ are considered similar with normalized weight:

$$\mathcal{S}(a, b) = \frac{\mathcal{C}(a, b)}{\mathcal{C}(b)}\left(1 - \frac{|\mathcal{C}(a) - \mathcal{C}(b)|}{\mathcal{C}(c)}\right) \quad (1)$$

The second term is a "popularity cost" which down-weights relationships of artists in which one is very popular and the other is very rare.

### *3.1.5 Community Metadata-derived Similarity*

Another more formal model of cultural similarity is provided by the "Klepmit" system, described in detail in [11]. "Klepmit" provides a continuous measure of cultural similarity by analyzing the community metadata associated with a particular artist (e.g., the text content of web pages returned by a search on the artist's name). This metadata is defined as a feature vector of textual terms (adjectives, unigrams, bigrams, and noun phrases) and similarity is computed by determining a weighted overlap via a Gaussian window over the tf·idf values.

$$\frac{f_t e^{-(log(f_d)-\mu)^2}}{2\sigma^2} \quad (2)$$

Here, $f_d$ is the document frequency of a term, $f_t$ the term frequency of a term, and $\mu$ and $\sigma$ are parameters indicating the mean and deviation of the Gaussian window. See Table 1 for example returned vectors.

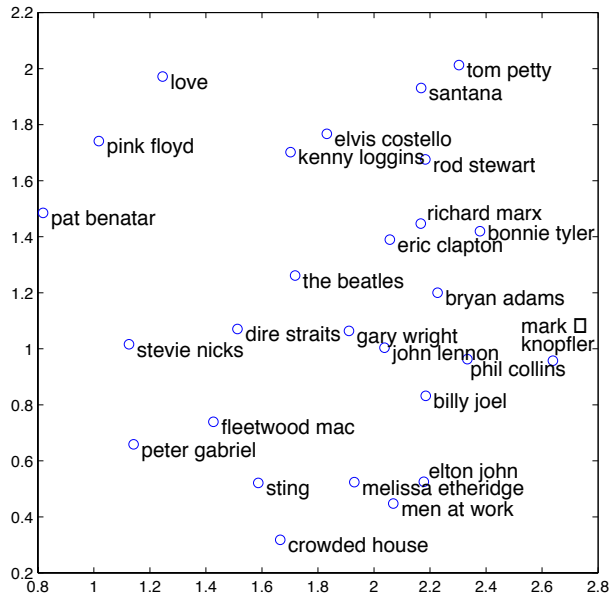


**Figure 1: Artists embedded in a 2-D space. This is a small portion of the full space derived from the Erdös measure.**

This model attempts to measure the popular opinion regarding an artist, and has the valuable property of being time-aware: community-metadata crawled only weeks apart can return widely varying results for a single artist. This data, arranged as a trajectory along time, can uniquely identify similarities of artists at any point in their career, as opposed to other models of similarity that treat artists as static indices in their database.

For the purposes of this experiment, we generated a matrix of similarities comparing each artist in our set with each other, along each of the different term types computed in the community metadata feature space.

## 3.2 Geometric Embedding

In addition to extending the coverage of a metric beyond directly-specified subjective comparisons, regularization may be required to give a particular metric properties such as symmetry and transitivity (i.e. the triangle inequality); one extreme way to ensure these properties is to convert a set of distance judgments into a set of points in a Euclidean space such that the Euclidean distances between the points do the best job of approximating the original distances. These points may be found via a straightforward gradient descent in a procedure often known as Multidimensional Scaling (MDS). A typical choice for the global error to be minimized is the root-mean-square (RMS) 'stress' along all links, i.e. the proportional difference between ideal and actual lengths. The final stress is also a measure of how successful MDS was in fitting the original distance measures. Points can be embedded in a space of arbitrary dimensionality; more dimensions afford more degrees of freedom and hence a lower stress. 2 and 3 dimensional embeddings have the attraction of permitting visualization of the dataset's geometric configuration; a small portion of a 2-D embedding of the Erdös distance is shown in Figure 1. For our artist similarity data, a 3D space provides for reasonably low-stress embedding, and we saw a plateau in RMS stress at 4 dimensions; using higher order spaces gave negligible improvements in fit.

Embedding can be applied to any of the measures. Where a similarity between 0 and 1 is provided (as with the OpenNap and

**Table 1: Top 10 terms for various community metadata vectors of the group Portishead. Here, the noun phrase and adjective terms seem to give the best descriptions and are imperative identifiers for uncovering cultural similarities.**

| n1 Term | Score | n2 Term | Score | np Term | Score | adj Term | Score |
|---|---|---|---|---|---|---|---|
| gibbons | 0.0774 | beth gibbons | 0.1310 | beth gibbons | 0.1648 | cynical | 0.2997 |
| dummy | 0.0576 | sour times | 0.0954 | trip hop | 0.1581 | produced | 0.1143 |
| displeasure | 0.0498 | blue lines | 0.0718 | dummy | 0.1153 | smooth | 0.0792 |
| nader | 0.0490 | 17 feb | 0.0675 | goosebumps | 0.0756 | dark | 0.0583 |
| tablets | 0.0479 | lumped into | 0.0665 | soulful melodies | 0.0608 | particular | 0.0571 |
| godrich | 0.0479 | which come | 0.0635 | rounder records | 0.0499 | loud | 0.0558 |
| irks | 0.0467 | mellow sound | 0.0573 | dante | 0.0499 | amazing | 0.0457 |
| corvair | 0.0465 | in together | 0.0519 | may 1997 | 0.0499 | vocal | 0.0391 |
| durban | 0.0461 | musicians will | 0.0494 | sbk | 0.0499 | unique | 0.0362 |
| farfisa | 0.0459 | enough like | 0.0494 | grace | 0.0499 | simple | 0.0354 |

Klepmit measures), it can be converted to a distance via $dist = (-\log(sim))^k$. Here, $k$ implements an arbitrary power-law monotonic transformation of every distance; in all cases, we searched over a range of such transformations ($k$ between 0.1 and 3.0) to find the one giving the lowest stress solution, since the relations of the measures to Euclidean distances are only specified up to a monotonic transformation.

# 4. EVALUATION

Having produced various alternative candidate ground-truth measures, we are faced with the problem of trying to compare their quality. Again, this needs to be related to true subjective judgments, but to use the same information as was the basis for one or more of the measures would be circular and misleading. Therefore, we collected a completely separate set of judgments for the specific purpose of evaluating our measures. First, we will describe the data collection, then how we used it to evaluate the measures.

## 4.1 Evaluation Collection Web Site

For the purposes of collecting large-scale evaluation data, we developed a web-based game and survey termed 'MusicSeer' (which is currently available at `http://musicseer.com/`). Using the 412 artists in our set, MusicSeer collects subjective human responses about artist to artist relationships. The system has two modes (freely selectable by the informant), both with their own specific purpose.

### *4.1.1 Artist Survey*

In the more direct route, we can ask informants "given an artist $x$, who is the most similar?" This is the approach of the artist survey mode, but with a few twists to make the data more valuable.

- **Pre-selected Choices:** The survey automatically selects a source artist and 10 target artists from the list of 412 artists. The source artist is selected from amongst popular artists, or artists that the user is familiar with (see below), while the target artists are randomly selected from the top 10 most similar artists according to the following three similarity metrics: OpenNap, Klepmit noun phrases, and Erdös.

- **Triplet Encoding:** Along with the pair of *source artist, target artist* that each judgment contains, we also store the remaining artists that were *not* selected. This allows us to understand a certain hierarchical ordering (over many judgments) from a particular source artist. For each selected artist, then, we actually store nine 'triplets' *source artist, target artist* (is more similar to source than...), *unselected artist*.

- **Bad Judgment Detection:** Peppered throughout the survey are a small amount of randomly generated 'fake band names.' We developed a set of statistically average artist name grammars and ran the terms used in current band names through them. Informants that select such red herrings as "Sleeplessness Explosive" or "Blonde and Bipolar" are treated with suspicion in later processing.

- **Unknown Artists:** The survey has an option to skip responding if the user is unfamiliar with the source artist, or with most of the target artists.

- **Adaptive Artist Selection:** The survey keeps track of artists that the user knows (the source or the selected target from prior responses) and does not know (the survey assumes the source artist is unknown when the "unknown" option is selected). Source artists are initially chosen from the most popular artists. After 5 responses, source artists are chosen from amongst the known artists 80% of the time, and from popular artists 20% of the time. Artists that we know the user is unfamiliar with are never chosen as source artists.

At the time of writing, the survey has generated over 6,200 responses (roughly 56,000 triplets.)

### *4.1.2 Erdös Game*

The Erdös Game (also known as 'poperdos' or 'the rabbit game') came about from the uniqueness of the Erdös distance measure extracted from the All Music similarities. Links between relatively distant artists were exciting to study (how could you get from Marilyn Manson to ABBA?) and we felt that a game founded on this data could attract attention to our data collection effort.

In the game, the informant is asked to select a target artist to go with a randomly chosen source artist, and is immediately presented with the pre-computed Erdös distance. The informant is then asked to match or beat that distance by moving along a chain of similar artists. Some pressure is added by the compelling back story of a lost rabbit trying to flee the clutches of an evil record store owner, who is curiously bent on denying the rabbit his favorite carrots and raisins.

At each 'hop,' the informant is presented with a list of immediate neighbors, from whom the artist most similar to the desired ultimate destination should be chosen. For example, at each hop in a Marilyn Manson to ABBA game, the user must select the closest artist to ABBA among the present similarity list. The list of possible artists is based on our existing metric set, slightly augmented from the basic All Music data, so that it is sometimes possible to beat the Erdös distance.

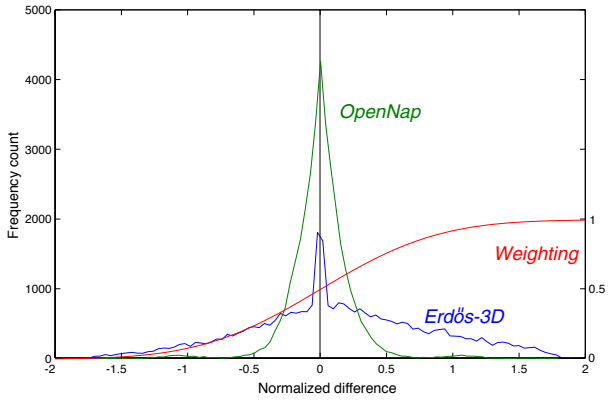


**Figure 2: Histograms of the scores when the evaluation triplets are converted into the difference in distance between selected and unselected targets, and normalized by the magnitudes of each distance according to the internal noise model. Note the slight bias visible in the distribution of the Erdös -3D data towards the positive (agreement) side. Superimposed is the $erf$ sigmoid weighting used to weight these histograms before integrating to give the overall weighted agreement.**

From our own experience, we realized that informants' judgments vary in nature and quality depending on the stage in the game. In earlier steps, judgments are for artists who may be very dissimilar, and while this is unique and valuable data, we also record the position within the game in our database in case we should wish to filter on this attribute at a later stage.

The Erdös Game has currently attracted 7,400 selections (over 82,000 triplets). Figure 3 shows sample screenshots of the web interface to both the survey and game modes.

## 4.2 Evaluation Measures

The web site has collected over 13,000 total selections, giving some 138,000 (source, target, unselected) relative similarity triplets with which to test our metrics. We use this data in two ways:

- **Average ranking:** For each selection, we use the metric under test to sort the list, then record the ranking of the actual item selected by the informant. Each ranking is normalized to a scale of 1 to 10 (for lists that contain greater or fewer than ten items), then averaged across all the judgments. A metric that perfectly predicted informant responses would give an average ranking of 1; random orderings should give a ranking around 5.5.

- **Average unweighted/weighted agreement:** A simple way to use the data triplets is to count the cases in which the inferred subjective judgment (that the source is more similar to the target than to the unselected alternative) agree with the distances given by the metric. This measure, the average unweighted agreement, has the disadvantage that it makes no distinction between a disagreement over artists of approximately equal similarity to the source (which is not serious), and the more significant situation in which an informant chooses a target that the metric rated as vastly inferior.

  This leads to the weighted agreement measure: We can model the informant's judgment as the comparison of 'true' similarity measures that have been corrupted by an internal noise source. If we assume the noise has a standard deviation in proportion to the magnitude of the similarities, then the significance of each triple becomes a function of the difference between their metric distances divided by the expected error margin i.e. $(d(S,T) - d(S,U))/\sqrt{(}d(S,T)^2 + d(S,U)^2)$. When $d$ is a distance, values less than zero indicate agreement between informant and metric. Positive or negative values close to zero are relatively insignificant, since the internal noise could easily cause an error in this range. A histogram of this normalized difference over the entire evaluation set gives a quick summary of the metric's performance, showing the extent to which it is biased to the 'agreement' side. Figure 2 shows examples for the OpenNap measure and the distances measured from the embedding of the Erdös measure in a 3-D space.

  To convert the histogram to a single score, we can sum the histogram bins, individually weighted to indicate their correctness and significance. The sigmoid function shown overlaid on the histogram provides such a weighting; judgments clearly reversed from the metric's predictions score 0, highly consistent judgments score 1, and ambiguous judgments land up in the middle of the histogram and have a weight of around 0.5. The width of the sigmoid transition corresponds to an assumption of the magnitude of the internal noise, i.e. over what range the choice between similar distances should be discounted. Arbitrarily, we used the large value illustrated in the figure, where the unweighted agreement would correspond to a zero transition widht.

  Averaging the weighted or unweighted counts over all the known-artist evaluation triplets gives an indication of how strongly the metric agreed (or disagreed, for a score below 50%) with the subjective data.

One issue that arose in using the evaluation website was that in many cases some of the artists on a list may be unknown to the informant. In this case, the selection cannot be accurately interpreted as meaning that the informant judged the selected target as more similar to the source than the unknown, unselected alternative. We devised a conservative procedure for ensuring that our data excluded such invalid triplets: Over the entire history of selections made by a particular informant (tracked via an anonymous web cookie), a list of 'known' artists is constructed as all the artists ever selected, on the assumption that informants would never select artists with whom they were not familiar. Then the triplets are filtered to retain only those in which both target and unselected alternate are affirmatively known by the informant. This removes about two thirds of the data triplets.

## 4.3 Results

Table 2 lists the results of our evaluation schemes. Average rankings are reported for each measure over four subsets of the evaluation data, broken down into the two modes (survey and game) and into all results, or known artists only. Restricting the ranking to the smaller set of artists known to each informant greatly reduces the effective list length and tends to increase average rankings. This may be because the unknown artists are more likely to be dissimilar to the known source artist, and hence we are removing items primarily from the bottom of the list before renormalizing to the 1-10 scale.

The ranking numbers are unfamiliar and we have been unable to calculate an *a priori* significance bound. However, some feeling for the stability of this data can be gained by looking at the variation in the ranking score of the random measure across the different subsets of the evaluation data. We expect the average score to be 5.5 (the average of values uniformly distributed in the range 1-10); there appears to be a slight negative bias, but the ranking values appear to be reliable at least to the first decimal place. We have adopted an average ranking difference of 0.1 as our significance threshold for this data.

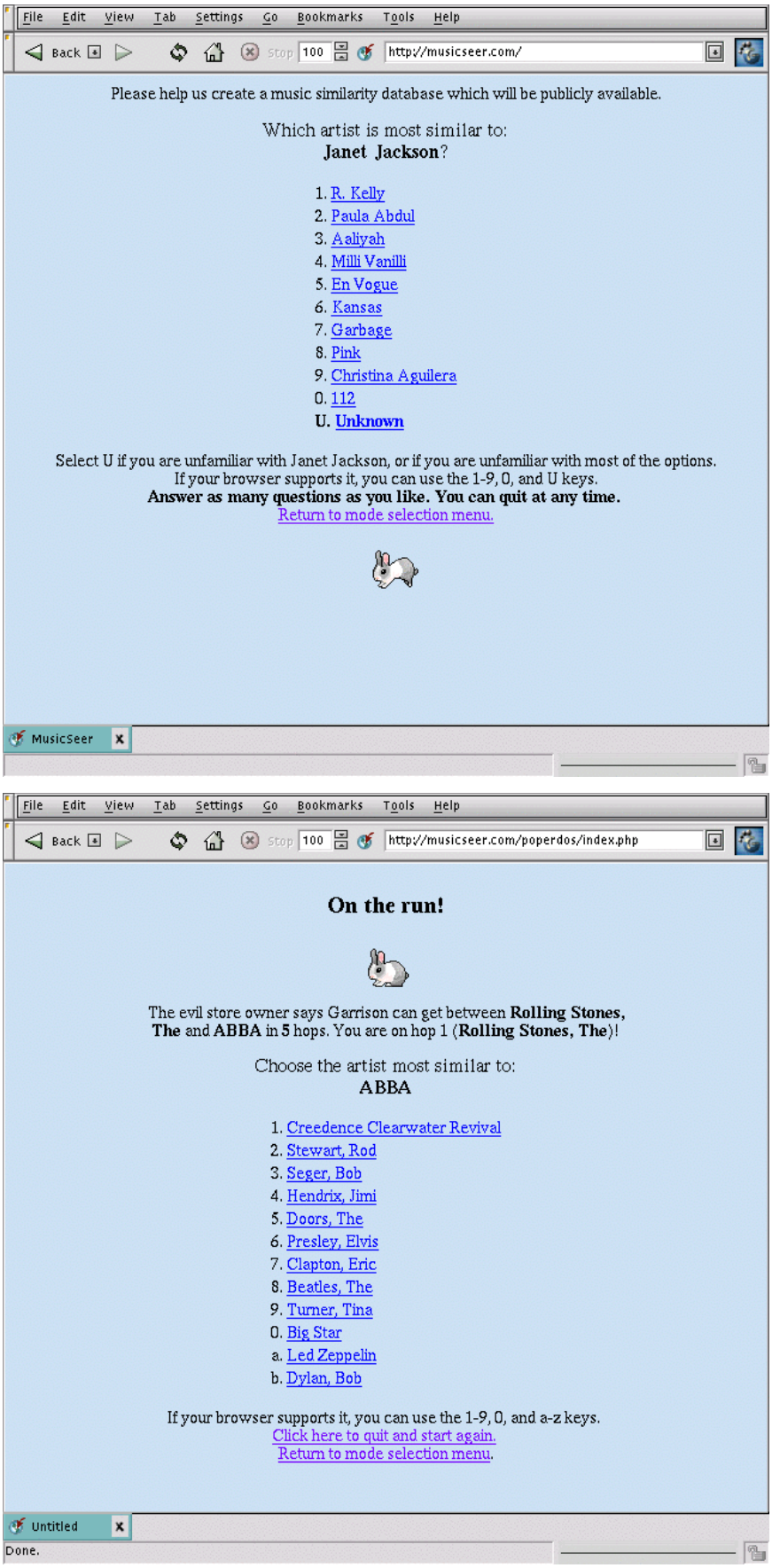


**Figure 3: Screenshots of the web interfaces used to collect the evaluation data. Top pane: Survey mode. Bottom pane: The Erdös game.**

**Table 2: Evaluation results. Each column describes a different metric, being: opt - the 'optimized' measure derived from the survey data; cmb - similarities from Erdös and OpenNap measures combined by simple averaging; erd - the plain Erdös distance; e2d - Erdös distance embedded in 2-dimensional space, then converted back into a similarity matrix based on the actual Euclidean distances; e3d - the same for a 3D space; e4d - the same for a 4D space; Rer - the "resistive" Erdös extension; onp - the OpenNap measure; kn1 - unigram features from the Klepmit data; kn2 - Klepmit bigram features; knp - Klepmit noun-phrase features; kaj - Klepmit adjectives; rnd - a random similarity matrix included for comparison. (Since rankings are normalized to fall between 1 and 10, we expect random choices to average out to around 5.5, as observed). Each row presents a different quality index for the metrics; the first four rows present average rankings of the user selection under each metric, broken up according to the collection mode (survey or game), and both with (all) and without (known) ratings involving artists that the informant may not know. 3D embedding stress is the final stress when the metric is embedded in a 3D space, and is of course zero for the metrics derived from Euclidean spaces of that size or smaller (e2d and e3d); the low embedding stress of the 'opt' measure arises because it defines only a small proportion of all the possible distances. Average unweighted agreement gives the proportion of collected judgment triplets that agree with the metric; average weighted agreement weights this value to discount errors where the artists in question are almost equivalent, as described in the text. In both cases, random agreement should score 50%.**

| Mode | opt | cmb | erd | e2d | e3d | e4d | Rer | onp | kn1 | kn2 | knp | kaj | rnd |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Survey, all (6177 resp, 8.97 av.choices) | 1.98 | 3.52 | 3.83 | 4.26 | 4.08 | 4.05 | 4.14 | 4.06 | 4.53 | 4.55 | 5.20 | 4.72 | 5.42 |
| Survey, known (4802 resp, 3.59 av.choices) | 2.24 | 4.26 | 4.07 | 4.50 | 4.26 | 4.22 | 4.92 | 5.14 | 4.62 | 4.46 | 4.66 | 4.96 | 5.44 |
| Game, all (7421 resp, 11.10 av.choices) | 1.91 | 4.41 | 4.50 | 4.64 | 4.56 | 4.54 | 4.77 | 4.65 | 5.44 | 5.37 | 5.39 | 5.57 | 5.49 |
| Game, known (6515 resp, 4.72 av.choices) | 2.68 | 5.02 | 4.87 | 4.94 | 4.88 | 4.90 | 5.31 | 5.35 | 5.42 | 5.36 | 5.35 | 5.57 | 5.45 |
| 3D embedding stress (%) | 15.3 | 23.7 | 20.8 | 0.0 | 0.0 | 13.5 | 20.3 | 27.9 | 34.1 | 34.5 | 34.7 | 36.2 | 35.8 |
| Average unweighted agreement (%) | 85.0 | 56.5 | 58.5 | 57.4 | 58.9 | 59.1 | 52.3 | 50.6 | 56.6 | 57.4 | 56.0 | 54.2 | 50.6 |
| Average weighted agreement (%) | 71.6 | 52.6 | 56.6 | 55.9 | 56.7 | 56.6 | 51.4 | 49.8 | 51.3 | 51.6 | 51.4 | 51.1 | 50.2 |

There is a question over the internal consistency of the survey data: in view of the introductory discussion, is it even possible for a single similarity measure to have good agreement with the judgments from more than 1,100 informants logged by the site? To answer this, we developed an optimal 'cheating' metric, constructed to have the best possible agreement with the survey data. For each source artist, we searched for an optimal ordering of the remaining artists by testing each referenced target artist at every point in the list and calculating the resulting agreement with all the judgments related to that source. This gave the "optimal" metric shown in the tables, which agrees with 88.2% of the collected judgments; we conclude that there is a good degree of consistency within the ratings. Note, however, that this cheating metric fares poorly by our original standards - it has no transitivity or symmetry (there is no effort to relate $d(A, B)$ to $d(B, A)$), and it specifies relations for each source artist only for the other artists with comparisons in the evaluation data - an average of 83.4 artists each, or about 20% of the total similarity matrix.

# 5. DISCUSSION AND CONCLUSIONS

The results show that on both the average ranking and the weighted agreement measures, the plain Erdös score performs the best among the various base measures we have proposed. Geometric embeddings of Erdös become increasingly similar to the plain measure as the dimensionality increases to 4 (and have the advantage of being true metrics, reflected in their low 3D embedding stresses). Resistive Erdös appears inferior to plain Erdös , although as discussed above there may be other forms of this measure that will perform better.

The OpenNap measure performs quite well on the rankings but not on the weighted agreement; as seen in Figure 2, this reflects the tight bunching of the length differences around zero for this measure. (The poor correlation between the weighted agreement and the average rankings in this case seems to imply that more sophisticated normalization is required within the weighted agreement calculation.) The various Klepmit similarities seem less promising than OpenNap. Notice that the embedding stress of these metrics is similar to the value for the random similarity matrix, implying that geometric embedding is not at all appropriate for this data, at least as we have implemented it.

Apart from the 'optimal' measure (which cannot be fairly compared, since it uses prior knowledge of the evaluation data to optimize its score), the best rankings are obtained by the combined measure that averages similarities from the Erdös and OpenNap sets. It seems logical that a combination should be able to outperform either measure alone, since the combined measure draws on the pooled knowledge represented by the subjective judgments underlying each measure. Our combination scheme, however, is very simple. It seems likely that a more sophisticated and better-performing combination measure could be found.

Differences between the survey and the game in the absolute values of the average ranking scores are to be expected because the cohorts from which user choices are made are very different: Game choices are made among a set of similar artists (the neighbors of the current 'position'), whereas survey sets come from a broader range. Thus, we expect non-cheating measures to do worse on the more closely-bunched game choices.

Returning to our original goal of constructing a full matrix of similarities among a given set of artists that could be used to train an automatic measure of artist similarity, the combined measure is at least a usable starting point. It may be, however, that the evaluation methodology and the judgments collected though the web site are equally useful; in our own current work developing signal-based music similarity measures, this evaluation procedure has turned out to be very valuable as a way to judge progress and refine our algorithms.

## 5.1 Summary and Conclusions

We have investigated the feasibility of deriving the 'ground truth' that underlies subjective assessments of artist similarities. This task is daunting, not only because such values defy direct measurement, but also because several considerations imply that a single metric cannot exist.

Nevertheless, we were able to coerce relatively modest amounts of subjective rating data from various sources into full similarity matrices with varying properties. In order to evaluate the different metrics, we collected a new dataset consisting of direct judgments of artist similarity. Under the various indices we devised to rate

our metrics against this evaluation data we found that several metrics performed quite well, and a simple combination of the metrics performed still better.

The motivation of this work was to define consistent measures over a large set of artists to be used as training data for automatic similarity measures based on audio data. We feel that the results of our best-performing combined metric is suitable for this task, although the evaluation methodolgy and data may turn out to be the more useful contribution. We plan to make the data from this metric, as well as the raw data used in our evaluation, freely available as a resource for the research community.

## 6. ACKNOWLEDGMENTS

This work was extensively supported in part by the NEC Research Institute, whose contribution is gratefully acknowledged.

## 7. REFERENCES

[1] A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *AES 22nd International Conference*, Espoo, Finland, June 15–17 2002.

[2] W. Chai and B. Vercoe. Folk music classification using hidden Markov models. In *Proceedings of International Conference on Artificial Intelligence*, 2001.

[3] W. W. Cohen and W. Fan. Web-collaborative filtering: recommending music by crawling the web. *WWW9 / Computer Networks*, 33(1-6):685–698, 2000.

[4] R. B. Dannenberg, B. Thom, and D. Watson. A machine learning approach to musical style recognition. In *In Proceedings of the 1997 International Computer Music Conference*, pages 344–347. International Computer Music Association., 1997.

[5] J. Herre, E. Allamance, and O. Hellmuth. Robust matching of audio signals using spectral flatness features. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130, Mohonk, New York, 2001.

[6] L. Hofmann-Engl. Towards a cognitive model of melodic similarity. In *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval*, pages 143–151, Bloomington, Indiana, 2001.

[7] J. Mauss and B. Neumann. Qualitative reasoning about electrical circuits using series-parallel-star trees. In *1st International Workshop on Model-based Systems and Qualitative Reasoning, ECAI'96 Workshop W23*, Budapest, 1996.

[8] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[9] G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *Proc. Int. Symposium on Music Inform. Retriev. (ISMIR)*, pages 205–210, October 2001.

[10] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with Minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568, Falmouth, Massachusetts, September 10–12 2001.

[11] B. Whitman and S. Lawrence. Inferring descriptions and similarity for music from community metadata. 2002. in preparation.

[12] C. Yang. Music database retrieval based on spectral similarity. In *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval*, pages 37–38, Bloomington, Indiana, 2001.