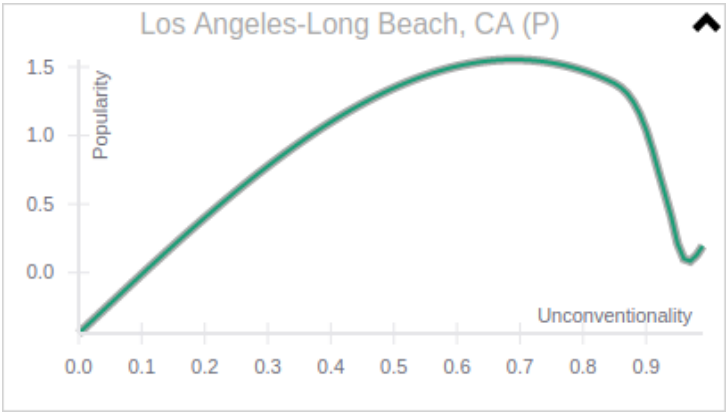
**Method and Interpretation for "Balancing Categorical Conventionality in Music" Visualization Tool**

Cities are breeding grounds of distinct modes of cultural activity. In this study, we examine how musicians define themselves differently depending on their location. In particular, we examine the relationship between bands' degree of conventionality and unconventionality and their popularity. We show that this relationship in general exhibits a non-linear, inverted U pattern: extremely conventional bands are relatively unpopular, somewhat unconventional bands show more popularity, while the most unusual bands are not very popular. This general pattern shifts, however, across musical genres and geography. Some cities show greater receptivity to unconventionality, with more unconventional bands achieving greater popularity, while in others more conventional bands tend to thrive.

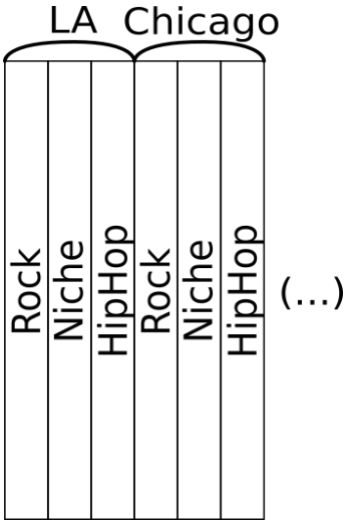
To examine patterns in the relationship between unconventionality and popularity across metro areas, we built an interactive visualization tool. This document describes the method by which we compare metros unconventionality-popularity curves and illustrates how to use the tool.

**Method**

Each pair (metro, world) has a curve that associates conventionality with popularity. To generate these curves, we use generalized additive models (GAM) with a cubic spline. This produces three curves for each metro, with one curve for each musical world. For instance, the curve for 'Rock' in 'Los Angeles - Long Beach' looks like this:




Our objective is to identify trends among these curves. Are there similar patterns between them? To identify those patterns, we use the predicted values from the GAMs. We start by stacking each sets of values leading to a matrix $V_{(n,m)}$:




This new matrix is decomposed using the Non-negative Matrix Factorization (NMF) (Cichocki and Phan 2009; Pedregosa et al. 2011). This factorization finds two new matrices, $W_{(n,k)}$ and $H_{(k,m)}$ such that $V \approx WH$, where k is a parameter of the method, with $k<<m$ and $k<<n$. That means that the matrices W and H can be used to approximate the original matrix V.

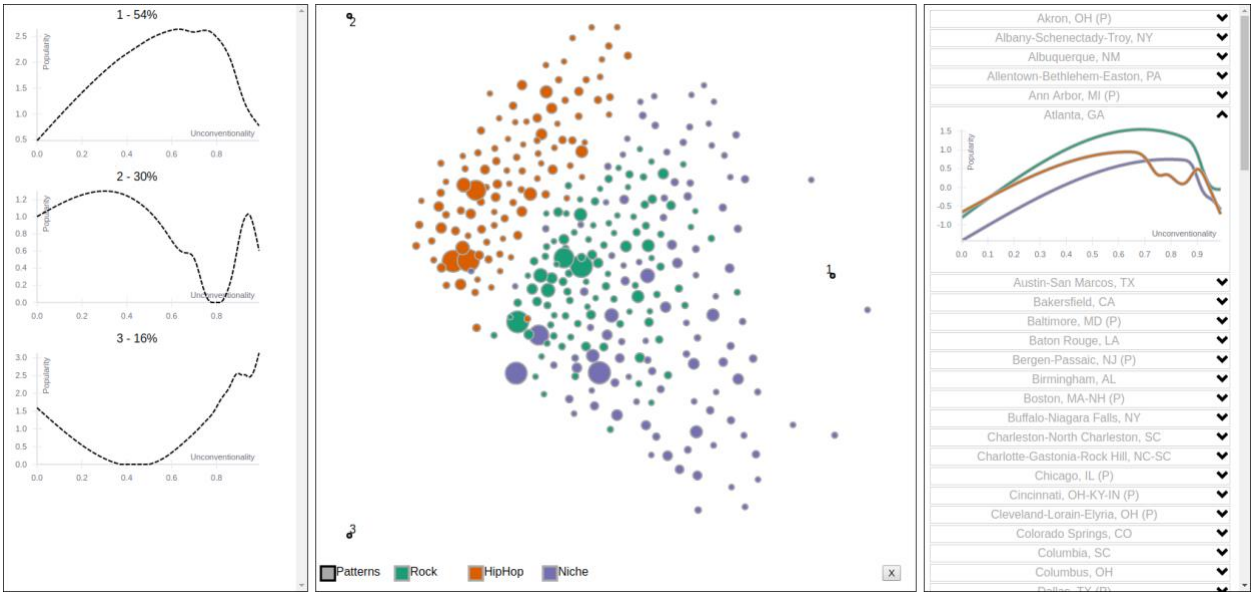
Since in our study k=3, this leads to a matrix W of size (n,3), where n is three times the number of metros in the dataset; and a matrix H of size (3,m), where m is the length of each vector (curve). We sampled each curve 100 times, so m=100.

The number of factors to be used is a parameter of the method, usually application dependent; we selected three factors, seeking a balance between precision and interpretability. With less than three factors, the decomposition was not able to properly differentiate the curves, while higher numbers led to more intricate interdependencies that hindered the interpretation of the results.

**Interpretation of the decomposition:** A more useful way to interpret this decomposition is as *weights* and *patterns*. Each column of V corresponds to a pair (metro,world) and can be approximated by multiplying one row of W by a column of H. In this sense, the columns of H represent composing patterns while the rows of W represent the weights of those patterns. In other words, the non-negative constraint leads to straightforward interpretation, where the factors composing the patterns express the curves. Since the same patterns are used for all curves, they can be classified according to the contribution of each pattern. By using these contributions as weights in a RadViz plot (Albuquerque et al. 2010), each metro-musical world pair can be plotted as a point, such that the points corresponding to similar curves are depicted close to each other.

**Visualization interface**

Our interface illustrates exactly those patterns and weights, as computed by the NMF. It contains three panels:




**The left panel** indicates the three **patterns** present in the matrix H and the average **weights** present in matrix W, if nothing is selected. If a pair (metro, world) is selected, this panel will display the same three patterns, but ordered according to their weights considering only the selection. The weights are also represented as a percentage, so we can easily identify dominant patterns.

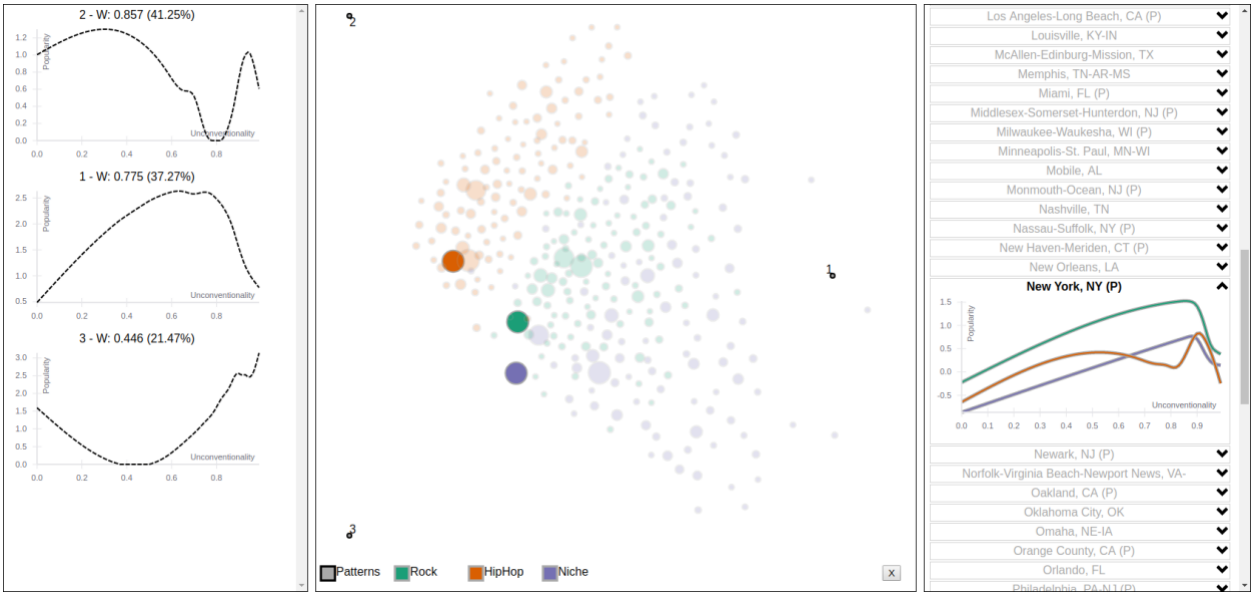
**The middle panel** contains a RadViz plot (Albuquerque et al. 2010) that contains one point for each pattern (grey circles, with an associated number) and points for each pair (metro, world), such that the distance between them is proportional to the weights of that pair. In other words, **points near a pattern are more similar to that pattern**. However, the plot also spreads the points over the space, reducing overlaps and occlusions. For the sake of visual simplicity, visualizations are restricted to metropolitan areas with greater than 500,000 residents. Doing so does not meaningfully change our results.

Accordingly, there are three points for each metro, one point for each musical world. Sizes of points are proportional to the metro population size. Rock curves are green; Hip Hop curves are orange; Niche curves are purple. Hovering over a point highlights all three points for that metro, and reveals the corresponding curves in that metro's heading in the right panel.

**The right panel** contains an expandable list illustrating the **original curves** for each city, for comparison. Clicking on a metro's name on the right panel will also reveal its three curves, and highlight its corresponding points in the central panel. Clicking on the *X* in the lower right of the central panel will reset the visualization.

**Concrete example**

To further illustrate how to interpret the results and the interface, let's consider a concrete example: Hip Hop in New York:




From the middle panel, we infer that the curve corresponding to (NYC, Hip Hop) is more similar to patterns 2 and 3, and less similar to pattern 1. Indeed, pattern 2 contributes with 41% of the weights, followed by pattern 1 with 37% and pattern 3 with 21%. While this may seem counter-intuitive, all points between this one and pattern 1 are more strongly associated to it. Indeed, pattern 1 contributes significantly to all curves.

On the right panel, we can see that the orange curve, corresponding to Hip Hop, increases with unconventionality, with a valley around 0.80, increasing again afterwards. To arrive at this curve using a weighted combination of the patterns depicted on the left panel, pattern 1 contributes with an increasing unconventionality that decreases near the end, pattern 3 includes a quick rise near the end, and pattern 2 provides a valley around 0.80. All three patterns are necessary to compose the original curve, including the increasing behavior (pattern 1), the valley (pattern 2), and the rise at the end (pattern 3).

**WORKS CITED**


Cichocki, Andrzej and Anh-Huy Phan. 2009. "Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations." *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences* E92-A(3):708–21.

Pedregosa, Fabian et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(Oct):2825–30.

Albuquerque, G., M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. 2010. “Improving the Visual Analysis of High-Dimensional Datasets Using Quality Measures.” Pp. 19–26 in *2010 IEEE Symposium on Visual Analytics Science and Technology*.