A Semantic Map of the last.fm Music Folksonomy

Biberstine, J. [1], Börner, K. [1], Duhon, R. J. [1], Hardy, E. [1], Skupin, A.[2]

[1]School of Library and Information Science, Indiana University, Bloomington, IN 47405

[2]Department of Geography, San Diego State University, San Diego, CA 92040, USA
Email: skupin@mail.sdsu.edu

1. Introduction

What does the world of music look like? We present a visualization derived from a repository of user-generated tags attached to more than one million items within the social music website last.fm (http://www.last.fm/). The site enables users to discover new music based on their listening history and – crucial for our study – users can annotate music-related items such as artists and songs with arbitrary tags, ranging from categories like "rock" or "jazz" to event-related attributes (e.g. "seen live") and affective utterances (e.g., "songs I absolutely love"). Tags also vary in scope, from rather broad categories, like "classical", to finer distinctions, like "britpop" or "female fronted metal".

The map offers viewers a mix of recognition, surprise, and discovery. Viewers have reported appreciation of the coherent patterns of hierarchical relationships among musical styles. The map is also notable for offering opportunities for discovering new musical categories, from the various flavors of "metal" to such niche areas as "shoegaze" or "drone".

The project presented here advances the visual analysis of social media on a number of fronts. Together with a recent study visualizing two million biomedical documents (Boyack et al. 2011), which generated several key technology elements used in our project, this is to our knowledge one of the largest self-organizing maps ever created in a single process (e.g., not hierarchically built, unlike Kohonen et al. 1999), and certainly one of the most comprehensive semantic depictions of the world of music. A project of this scale would not have been possible without parallelization and supercomputing resources.

GIScience concepts are at the core of how the project advances the analysis of social media data, which has traditionally been dominated by a network paradigm. Instead, our use of SOM builds on a view of social media items as existing in a high-dimensional attribute space, which calls for very different conceptual and computational approaches. Such geographic and mapping metaphors as *scale*, *region*, or *base map* feature prominently in the conceptual design of this visualization, while its rendering leverages commercial off-the-shelf GIS technology.

2. Data and Methods

The original data set, collected during the first half of 2009 (Schifanella et al. 2010) contains almost 1.4 million different music-related items, to which over 280,000 different tags were attached. In a vector-space type framework (Salton 1989), this would amount to a space that is extremely high-dimensional and yet occupied by exceedingly sparse vectors. That provides a justification for reducing the dimensionality by removing all but the 1,000 most frequently used tags. Items not annotated with any of those remaining tags are then deleted. This reduces the number of items from 1,393,559 to 1,088,761, with the average item having 6.8 overall tags and 3.8 unique tags associated with it. The top twenty most popular tags were: rock, electronic, seen live, indie, alternative, pop, female vocalists, jazz, classic rock, experimental, ambient, metal, alternative rock, singer-songwriter, 80s, folk, hard rock, progressive rock, indie rock, electronica, punk.

Instead of mapping one million tagged music items per se, they are used to generate a *model* of the 1000-dimensional tag space via a vector-space representation. Since the intent is to recognize not just global patterns existing in that space, but also regional and local patterns, a SOM of relatively high resolution was created (180 x 180 neurons). The confluence of neuron count (32,400), high dimensionality (1,000), and very large numbers of input items (1,088,761) makes the application of standard toolsets for SOM training, such as SOM_PAK (Kohonen et al. 1996), completely unfeasible.

In an earlier study involving two million text documents (Boyack et al. 2011), we succeeded in reducing an estimated serial training time of four years to six days by parallelization (Lawrence et al. 1999) on the supercomputer Big Red at Indiana University. In the same manner, the last.fm data underwent a process of parallelization into 300 processes across 100 compute nodes, for a parallel run time of 13 hours, as compared to the estimated serial runtime of five months. In the process, each of the one million training vectors is cycled fifty times through neural network training.

Post-processing of the trained SOM includes computation of five levels of a neuron label clustering solution (Skupin 2004), which are then stored in an ArcGIS Geodatabase. With more than 37,000 geometric features generated within the five cluster solutions, each with its own label term, automated label placement is a crucial component and is here accomplished via ESRI's Maplex labelling engine.

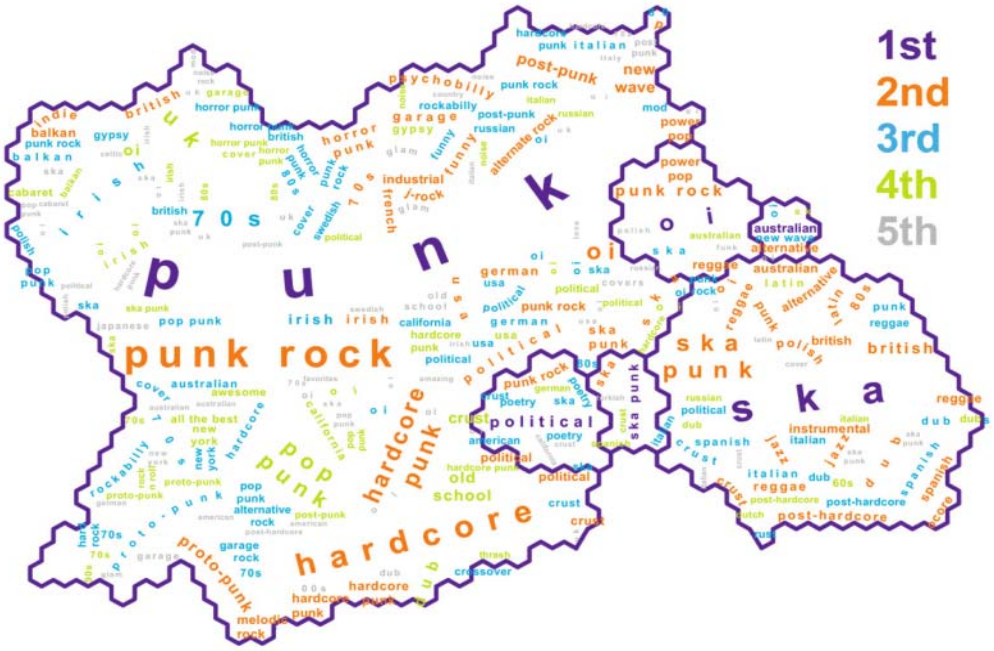

Figure 1. Multi-level delineation and labeling of regional structures in a large SOM trained with over one million tagged music items.

3. Results and Interpretation

With a core mapped area of 66x78 centimetres, the map invites exploration, despite the static output medium. The experiences reported by viewers range from a *recognition* of expected structures to the *discovery* of new styles of music and their relationship to well-known ones. There also can be *surprise* when certain structures in this map of the collective mind of music do not seem to conform to individual expectations. For example, some expect that "blues" should occupy a prominent region in the vicinity of "jazz". Instead, as a top-level term,

"blues" is located near "rock", which actually makes sense in terms of musical genealogy. And yet, on closer inspection it turns out that "blues" in addition plays a second-level role at the boundary region between "jazz" and "soul". Note that we are not performing a mapping of tags as such, but instead a labeling of regions according to the tags that dominate them. Thus, tags can indeed appear in multiple locations, occasionally even at great distance. That occurs when a tag has strong associations with multiple terms that are not in turn strongly associated with each other.

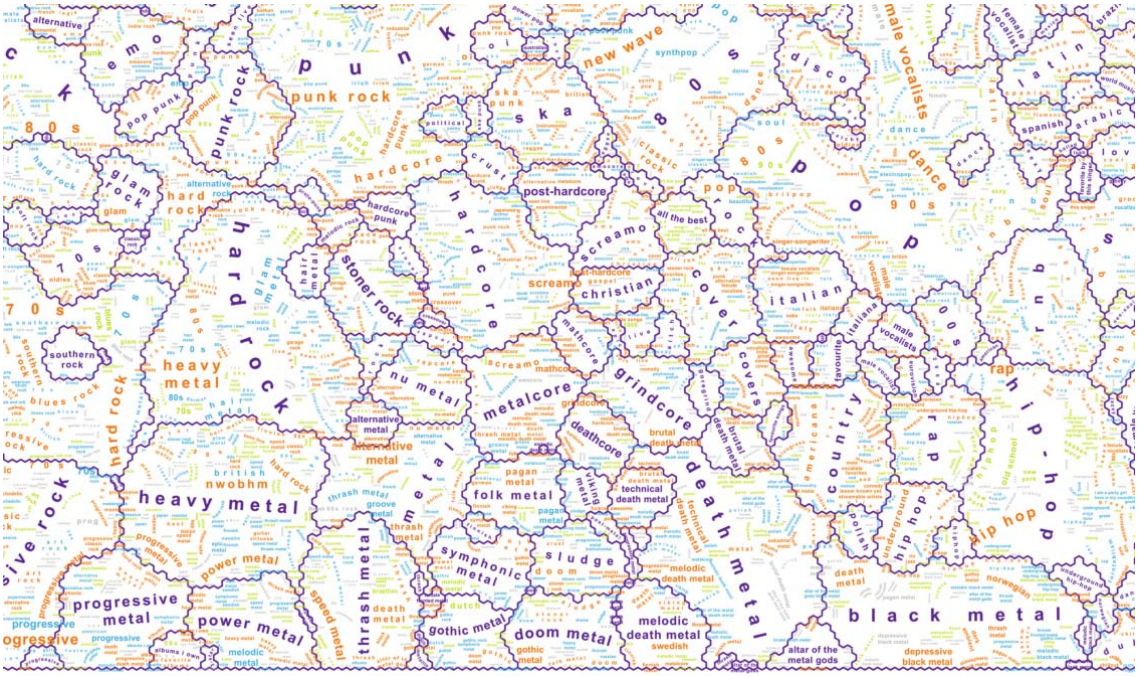

Figure 2. Small snippet (35%) of the finished last.fm visualization.

4. Future Plans

One key direction in this work is the development of solutions for interactive exploration of highly detailed maps of social media artefacts. One project currently underway examines the use of contemporary Web GIS technology to enhance the experience of these non-geographic information visualizations.

Among specific modifications to the computational techniques, the pre-processing of very high-dimensional data stands out, such as the use of latent Dirichlet allocation (Blei et al. 2003) to preprocess the original input vectors prior to SOM training. The result would be a SOM input data set that reflects a semantically richer, more nuanced representation – as compared to the minimum term frequency threshold used here – and yet in a form that involves fewer dimensions.

Given the computational cost, it would currently not be feasible to have a map like this – with high resolution and dimensionality – evolve in real-time, such as in response to dynamic social media content. However, its use as a *base map* is a different story, since we are not mapping similarities between tagged music artefacts, but rather construct a high-dimensional model with low-dimensional topological constraints. That model allows speedy projection of newly arriving artefacts onto the base map.

Efforts like the one described here can benefit from further integration of GIScience concepts and principles. An example is the introduction of a continuous field view of social media items and activities, as an alternative to current conceptualizations that seem predominantly informed by an information retrieval tradition (Skupin 2009).

Acknowledgements

This work was funded by the Cyberinfrastructure for Network Science Center at Indiana University, the James S. McDonnell Foundation, and the National Science Foundation under grant SBE-0738111. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also gratefully acknowledge the feedback received from three anonymous reviewers whose comments helped to improve this extended abstract.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003, Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, pp. 993-1022.

Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., Schijvenaars, B., Skupin, A., Ma, N.A.L. and Börner, K., 2011, Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. In *PLoS ONE*, p. e18029.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, T., Paatero, V. and Saarela, A., 1999, Self Organization of a Massive Text Document Collection. In *Kohonen Maps*, E. Oja and S. Kaski (Eds.), pp. 171-182 (Amsterdam: Elsevier).

Kohonen, T., Hynninen, J., Kangas, J. and Laaksonen, J., 1996, *SOM_PAK: The Self-Organizing Map Program Package* (Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science).

Lawrence, R.D., Almasi, G.S. and Rushmeier, H.E., 1999, A scalable parallel algorithm for self-organizing maps with applications to sparse data mining problems. *Data Mining and Knowledge Discovery*, 3, pp. 171-195.

Salton, G., 1989, *Automated Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Reading MA: Addison-Wesley Publishing Company).

Schifanella, R., Barrat, A., Cattuto, C., Markines, B. and Menczer, F., 2010, Folks in Folksonomies: Social Link Prediction from Shared Metadata. In *Proc. 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pp. 271-280 (New York, NY, USA: ACM).

Skupin, A., 2004, The World of Geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101, pp. 5274-5278.

Skupin, A., 2009, Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics*, 3, pp. 233-245.