Authors' Commentary: Making Good Music by Writing a Top-40 Network Modeling Problem

Chris Arney
State University of New York at Oswego
Oswego, NY

Amanda Beecher
Ramapo College of New Jersey
Mahwah, NJ
abeecher@ramapo.edu

Evangelia (Evelyn) Panagakou
Northeastern University
Boston, MA

Introduction

Music influence is integrated with the way society functions and progresses. The study of music (like most art forms) has the potential to capture societal and cultural values. Additionally, artists creating the music have influence over its message that reflects the current society. The way that a single or group of artists influences other artists propagates through and can transform the musical landscape. So understanding musical changes over time is directly connected to the musical influence of artists.

To study this relationship, it is necessary to take into account different social parameters so as to understand, analyze, and explain music influence. Using networks is definitely an advantageous way of considering and unveiling influence and similarity among artists and songs, and we believed this was an appropriate topic for this year's network science problem.

To understand musical influence on an artist is to study relationships between musical artists. The study of relationships is the basis for the field

The UMAP Journal 42 (3) (2021) 263–269. ©Copyright 2021 by COMAP, Inc. All rights reserved. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice. Abstracting with credit is permitted, but copyrights for components of this work owned by others than COMAP must be honored. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior permission from COMAP.

of network science. So teams working on this ICM problem needed to create an influence network and determine a way to measure musical influence on it. To understand musical change, the data of music itself must be studied. The data given may not share this natural network structure, although it was interesting to read papers that did add this component in their analysis. Therefore, most teams had to balance their ability to understand network measures to see the influence between individual artists, and had to model the music itself to understand how the music changed over time. The problem identified these two components as *musical influence* and *similarity*.

Teams integrated network analysis with data modeling to create information that brings forward ideas of musical influence. We appreciated teams considering how music evolved over time as well as finding revolutionaries who made rapid and significant changes. If we consider how music can mirror culture, and that the two are so intertwined, we have the opportunity to look at a cross-section of culture using network and data modeling. Discussing the value of a network approach is what we had asked the teams to do in the one-page overview to the Integrative Collective Music (ICM) Society.

Problem Formulation

Modeling is a useful tool for problem solving, decision making, understanding phenomena in the real world, and developing insight that can consider many elements of a problem, including context and purpose. Network modeling and data modeling, like music influence, are popular topics with undergraduate teams; so we felt that this was a good theme for an ICM problem.

It was our goal to write a problem that led to a large network (or perhaps several smaller networks) and involved teams in some of the elements of data science. The educational goals were for teams to gain experience in modeling and to learn the advantages of network analysis, that is, to see the modeling problem as a network science problem. We wanted the teams to acquire these skills and understanding by working on a problem that not only provided the learning opportunities we aimed for, but also was so integrated with everyday life that the teams could relate to it, be inspired, and see the added value of their analysis.

We sought to write a problem that, in addition to giving the teams the opportunity to incorporate the elements of modeling, would also lead teams to:

- consider elements of science and create valid measures;
- consider the human element (culture, art, social issues);

- get exposed to the elements of structural complexity (i.e., multiscale, multiple stakeholders, or multiple perspectives);
- encounter data that were relatively clean, available, and substantial;
- be able to make sufficient progress within the time restriction of four days in the contest; and
- have a personal interest in the questions and issues of the problem.

Based on the papers that we read, it seems that teams did achieve these goals.

Data Sets

We realized that the biggest hurdle to our effort was finding the right data set. Although music influence is such an interesting topic, finding appropriate data for an undergraduate contest was not straightforward. Not only did we have to consider the richness of data, but we also had to make sure that the data would support the type of analysis that we were aiming at, and most importantly, be appropriate for use in an educational context.

Once we found and refined the data set, we saw that the teams could investigate the role of music in society, measure the influences of artists on each other, determine music similarity and evolution, determine the characteristics and dynamics of music genres, and define and determine the revolutionaries in music. These tasks enabled the problem to meet our goals for the teams.

To investigate musical influence, we needed data that supported the relationship between artists and expressed how they were influenced. Specifically, 143,625 audio files were used from AllMusic.com, and experts' opinions were incorporated as ground truth of artist-to-artist influence. Data for 16,704 artists were obtained from Wenzhe (Harry) Xue, who had used it in his senior thesis [Xue 2018]. We liked this data set because we felt that it could be a natural starting point for the teams to form a directed network of music influence.

We found Spotify data from Kaggle [2020] that allowed teams to explore musical changes over time by analyzing musical characteristics. This database offered several data files to model and measure the influence of artists on one another and determine other factors that affect artists, genres, and songs.

We thought that one important focal point of the analysis should be about the role of evolutionary and revolutionary trends of bands and genres. It is very important to understand how a process evolves and how change is introduced in such a process, what factors interact, what are the dynamics involved, and what is the effect of time. The Spotify data sets

offered enough information for the teams to investigate these phenomena, and we were interested in seeing how the teams could use networks to answer these questions in a meaningful way that also could connect to societal changes.

By using the influence network as a baseline, there are even more opportunities to explore musical influence by overlaying the music and artist characteristics. We didn't want teams mining data during the contest but wanted them still to be able to model musical influence. Thus, if you use the full data sets `original_influence` and `original_by_artist` that are available at this *Journal*'s supplements website

`https://www.comap.com/product/periodicals/supplements.html`

you will see that matching artists in the two files is not easy. Here is a list of the steps that we took to present ready-to-use data for this contest:

- In both those data sets, some artists' names contain explicit language. We decided to remove this language so that the modified data sets would be appropriate for use in an educational context such as ICM. Unfortunately, this meant that we had to exclude many artists—in fact, we removed the entire Rap genre, since doing that was easier than identifying the individual artists' names and songs that contained explicit language. From the other data genres, we had to exclude only a few artists.
- The song titles too contained explicit language. We chose to censor the song titles rather than remove them. We used a pre-built Python package called `better_profanity` [Thanh 2020], which censored more words than was necessary. Many teams (and judges) commented that some song titles were censored even though the titles did not contain inappropriate language. However, the program allowed for a thorough removal in a short time.
- We did artist matching in the original data sets `original_influence` and `original_by_artist` by using the unique IDs given by AllMusic.com. Some exceptions might still exist in cases of artists with the exact same name or if there were misspellings of names.
- We also removed all years prior to 1930 from `original_influence`, and as a consequence also from `original_by_artist`. For the contest data set `full_music_data`, we left in some data between 1921 and 1930 so as to enable retrieving possible influential relationships that are seen in `original_influence` in the 1930s.

The original data sets are at the *UMAP Journal* supplements website noted above, and the final problem version and data sets are available at

`https://www.comap.com/undergraduate/contests/mcm/contests/2021/problems/`

We wanted to ensure that teams would have the opportunity to explore the relationship between influence and music, so we took the steps above to ensure that the data were ready for analysis. The complete database

gives additional opportunities to explore musical influence. It is up to the researchers to decide if they would like to use censored or uncensored data, and which version would serve their goals better.

Classroom Use

For those considering using a problem such as this one for a classroom lesson or projects, we offer a bit of advice. We realize that this is a fun problem (part of our motivation for selecting it) and that there is much more to explore than we cover here. Often in a course project, the tasks are the students' focus, and accomplishment of the tasks is success. For the ICM, we look to the tasks as a baseline to set the context and get the team engaged in the problem so as to create a more-thorough and creative investigation of the topic, and for the team to define and own the purpose of the issue. The point is that *we expect more than the accomplishment of tasks*: We want the teams to show the judges their modeling [Arney et al. 2021] and the power and value of their results. We expect to see flow between the assumptions, the arguments, the analysis, and the conclusion. This way, teams get trained in producing a well-connected scientific paper, a skill that will be useful to them as they progress in their careers. We would love to hear examples of how you furthered this mission with these data sets.

We've been pleased that many instructors, advisors, and teams have already reached out to us about using the full data set. We have a few cautions for those who would like to use it in other academic pursuits or classroom tools.

- We removed the rap genre because many of the artists' names and or song titles include explicit language; if you use the full data set, you should be aware of that fact. On the other hand, if you analyze the full list, you'll realize that there are much richer data and that the influence network is a single connected component (due to how Harry Xue obtained the data, it must be a connected network).
- We cleaned the data to match artists in the two data sets. If you consider the full original data sets, you must link them together to match the artists who are in both lists and realize that not every artist appears on both lists. We matched the artists by the ID given by AllMusic.com.
- As many teams noted in their solutions, many of the summative data of artists and years given in the auxiliary files do not match the actual means of the data within the full song data. This is because we removed artists from certain years who don't appear on both lists but we kept their data so that you could see annual trends from the full set of data.

Acknowledgments

The authors would like to sincerely thank Harry Xue for his generosity in sharing this data set with us and agreeing to make it available for educational purposes. For those wishing to use the data for their own purposes, please cite in your reference list the items by Xue and COMAP listed below.

References

Arney, Chris, Amanda Beecher, and Jessica Libertini. 2021. What is math modeling? It's a noun! It's a verb! It's ... a mindset??? *Consortium* 120 (Spring/Summer) 1–2. https://www.comap.com/pdf/1733/file.pdf.

COMAP. 2021. Interdisciplinary Contest in Modeling 2021: Problem D: The Influence of Music. https://www.comap.com/undergraduate/contests/mcm/contests/2021/problems/2021_ICM_Problem_D.pdf. Data at http://www.mathmodels.org/Problems/2021/ICM-D/2021_ICM_Problem_D_Data.zip.

Kaggle. 2021. Spotify Dataset 1922–2021, 600k Tracks. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks.

Thanh, Son Nguyen. 2020. `better-profanity`. Version 0.7.0, MIT, 2020. https://pypi.org/project/better-profanity/.

Xue, Wenzhe. 2018. Modeling musical influence through data. Senior thesis, Harvard University. https://dash.harvard.edu/bitstream/handle/1/38811527/XUE-SENIORTHESIS-2018.pdf?sequence=3&isAllowed=y.

About the Authors

Chris Arney has a Ph.D. in mathematics from Rensselaer Polytechnic Institute. He graduated from the US Military Academy and served in the US Army for 30 years, including teaching mathematics and network science for 29 years at the Academy. He is the founding director of the ICM and served as the director of the contest for 21 years.

Amanda Beecher is an Associate Professor of Mathematics at Ramapo College of New Jersey. She earned her Ph.D. from the University at Albany, SUNY. Before arriving at Ramapo, she had a three-year postdoc at the US Military Academy, where she first began advising for the MCM and judging for the ICM. Her research interests include combinatorics, graph theory, commutative algebra, and environmental modeling. Her educational endeavors include bringing interdisciplinary opportunities to her colleagues and students, including establishing an interdisciplinary data science program at Ramapo College.

Evangelia (Evelyn) Panagakou is the Education, Outreach, and Diversity Coordinator of the Network Science Institute at Northeastern University. Evelyn holds a B.S. and a Ph.D. in physics from the University of Athens in Greece, an M.S. in applied mathematics from the University of Massachusetts at Amherst, and an M.S. in applied developmental and educational psychology from Boston College. Her research interests include reaction-diffusion systems, computational epidemiology, and network science, as well as learning science, particularly cognitive, contextual, and cultural processes that support math and science learning. In 2021, Evelyn was one of the authors of Problem D, the ICM Problem D triage coordinator for individual judges, and one of the triage and final judges for the ICM network problem.