Cultural evolution of emotions in 50 years of pop song lyrics

A. Acerbi, C. Brand, A. Mesoudi

Data preparation

Dataset "mxm"

We downloaded the bag-of-words data (N=237,662 songs) from https://labrosa.ee.columbia.edu/millionsong/musixmatch . We used the mxm track ID to retrieve from musixmatch.com the name of the artist(s), the year of first release, and the genre for each song.

We eliminated from the original dataset songs for which was not possible to retrieve the year or the artist, as well as songs likely to not be in English language. To assess if a song was in English we checked whether the word "the" was present at least once. The new dataset contains N=163,551 songs.

Artists' names were further processed using the cluster function in R library *refinr* (https://github.com/ChrisMuir/refinr) to cluster and merge similar names (e.g. "madonna", "Madonna", "MADONNA", etc.). To disambiguate collaborations we looked for standard separators in artist names ("featuring","feat.","feat","and","AND","&","with",","). We considered artists where no separators were find as single artists. We then run the strings where separators were find (e.g. X and Y) through the list of single artists. If X was find in this list, than X and Y were considered separately as single artists (e.g. Eminem and Dr. Dre), if not, they were considered a collaboration (e.g. Simon and Garfunkel).

Dataset "billboard"

We downloaded the data from https://github.com/walkerkq/musiclyrics (N=4,913 songs). Since the lyrics in the "mxm" dataset are stemmed (a common practice in digital text analysis: words are reduced to their stems, roughly analogous to their morphological roots, e.g. "happily", "happy", and "happiness" are all reduced to the stemmed form: "happi"), we processed in the same way the "billboard" dataset, re-adapting the script: https://github.com/tbertinmahieux/MSongsDB/blob/master/Tasks_Demos/Lyrics/lyrics_to_bow.py

Artist were further processed as for the "mxm" dataset. Notice the "genre" entry is not present in this dataset, but, importantly, the "rank" is. "Rank" is the position, from 1 to 100, in the yearly Billboard top-list.

Sentiment analysis tool

We used the "positive emotions" and "negative emotions" categories of the text analysis application Linguistic Inquiry and Word Count (LIWC, see Pennebaker et al. 2007). These categories are "virtually unchanged" in respect to the most recent (2015) version of LIWC

(Pennebaker, personal communication to AA). The words were stemmed as mentioned above, and we analysed the lyrics with N=267 negative emotion stems and N=169 positive emotion stems.