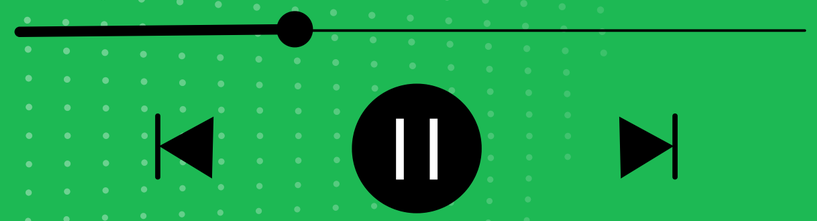
Generational GAP*

Can we still talk about the age of diversity when it comes to music?



API's

*Global Average Pooling

where do we start?

The "**Music Dataset: Lyrics and Metadata from 1950 to 2019**" is fertile ground for our hypotheses as it includes:

- 25K+ songs from 7 different decades
- Metadata for each song (we'll focus on topics and genre)
- full lyrics for each song

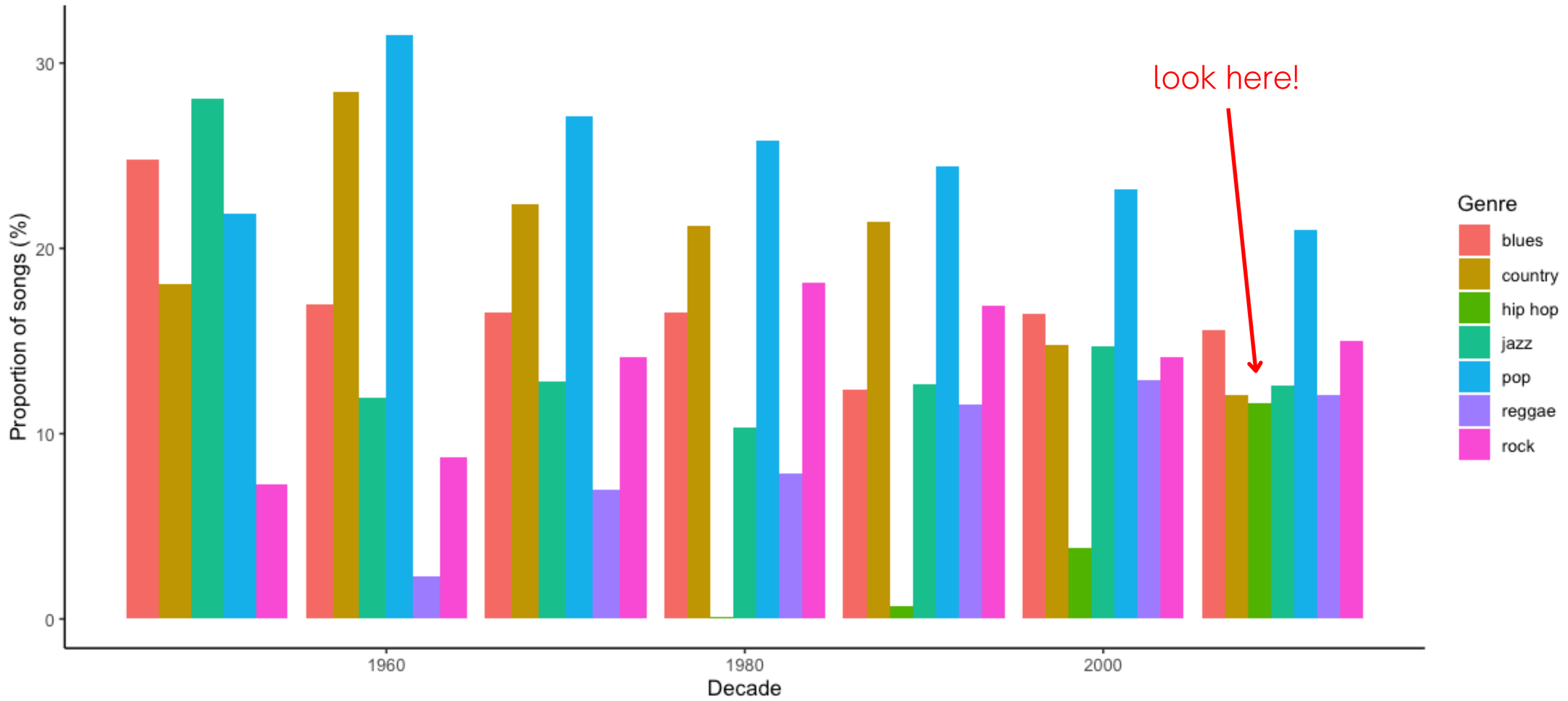
decade	1950	1960	1970	1980	1990	2000	2010
count	1408	3256	3767	4508	4349	4642	5518

what are we looking for?

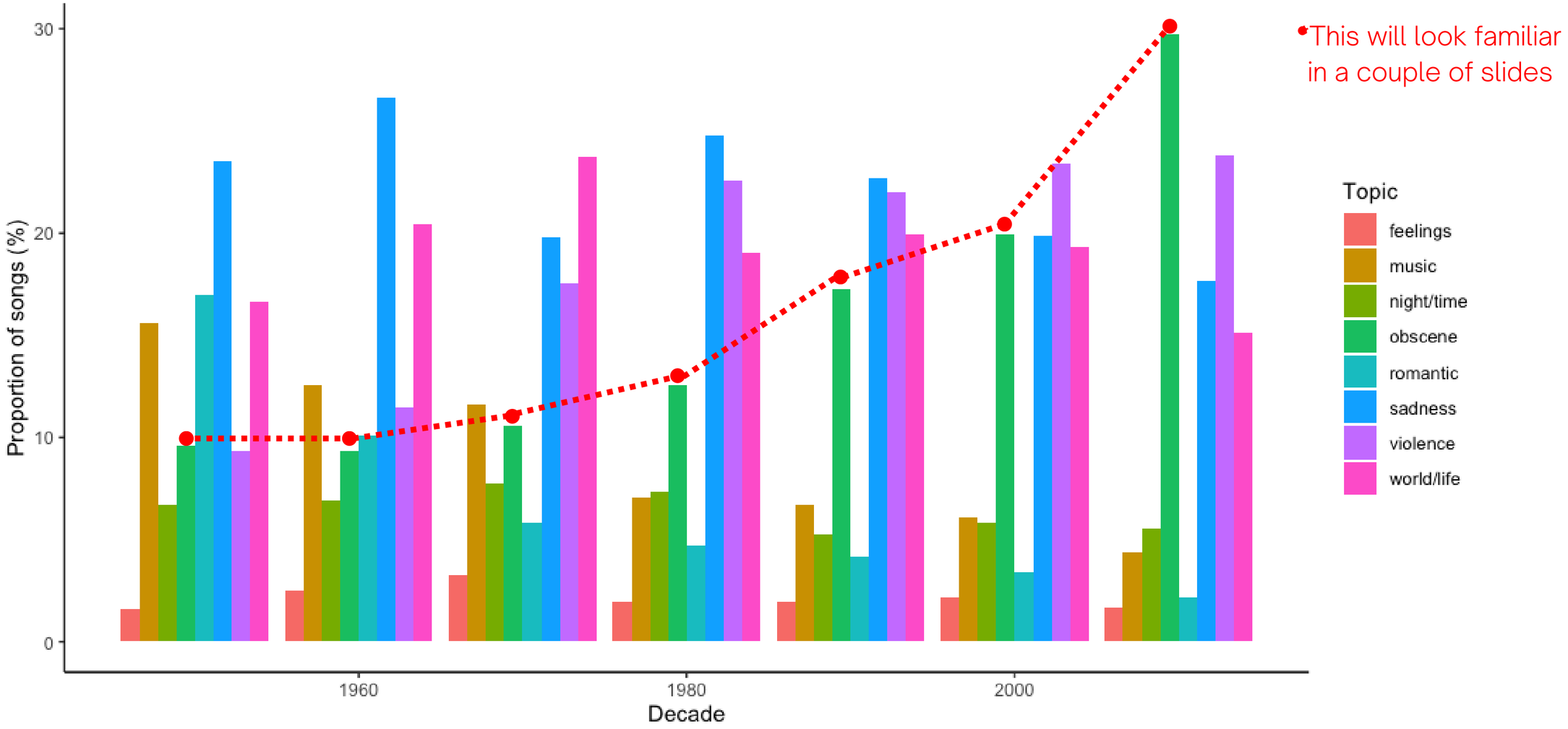
We would like to explore the network of songs to see how songs with similar content tend to aggregate with each other.

- Are there topics that tend to clump together more than others?
- What about genres? And decades?
- What are the most homogeneous periods?
- By what is homogeneity characterized? Genre or Topic?
- Is there a temporal trend in the similarity of lyrics?

warm-up (1): Genres Evolution



warm-up (2): Topics Evolution



This will look familiar in a couple of slides

take home messages

As we look at both graphs, we need to pay attention to some details that will come in handy throughout the analysis:

- Hip-Hop ramped starting from the 90s
- Seems like we are sad and violent mostly of the time, asides from jokes obscenity is the real deal

This insights give us the possibility to asks ourselves:

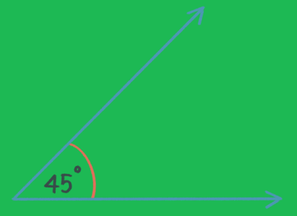
- Is it all about genres and topics? (could be, superficially)
- What type of conclusions can we draw? (spoiler: up to now nothing)

(dig into the graph)

building the graph

To begin with, we need to build the network of similarity between song lyrics. This will allow us to study how semantically and syntactically similar songs tend to aggregate with each other:

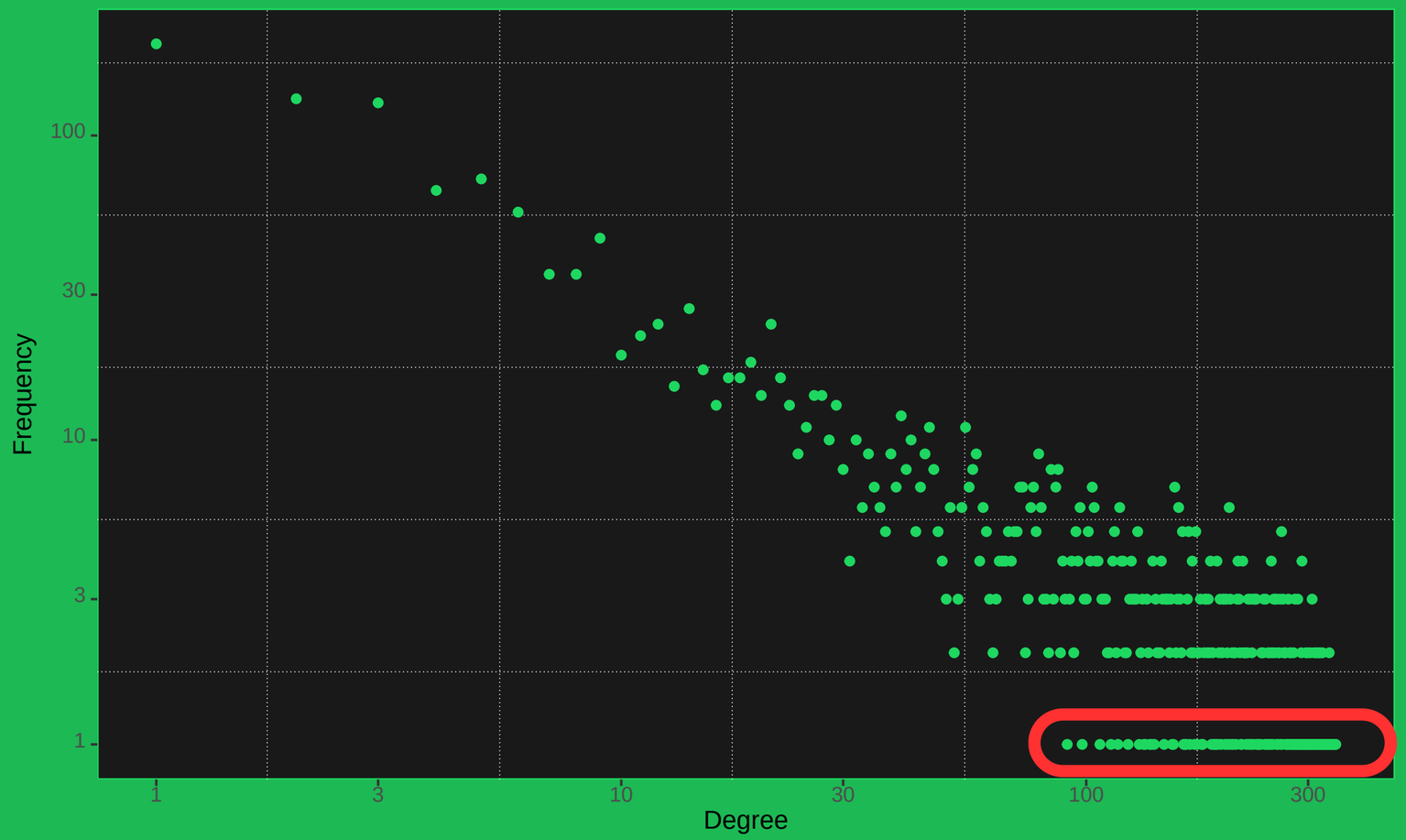
1. Take the lyrics for each song
2. Obtain 384-dimensional sentence embeddings using BERT
3. Compute pair-wise cosine similarities to build the adjacency matrix



The actual similarity graph is obtained by thresholding the edges keeping only those edges which have cosine greater than 0.8.
The decision for such a high similarity threshold has two main reasons:

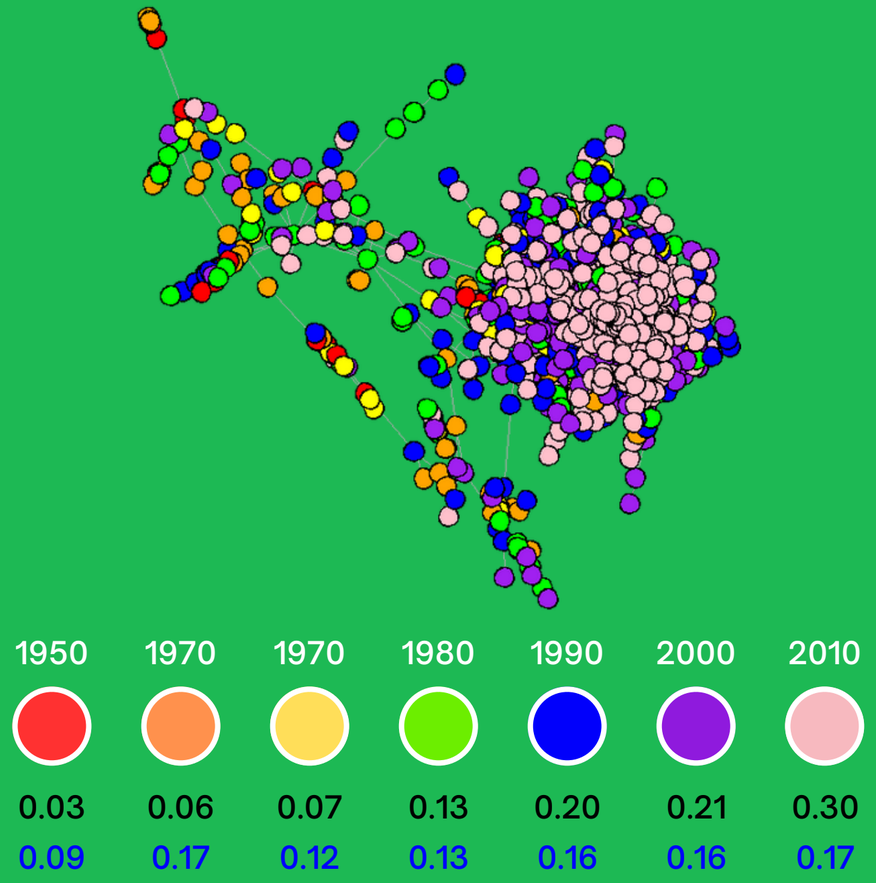
1. Memory constraints: moving the threshold from 0.8 → 0.7 increases drastically the number of nodes and mainly edges
2. Focus the study case on the most representative songs, we would like to say that the language / lyrics are **MOSTLY** similar in a specific period

log-log degree distribution



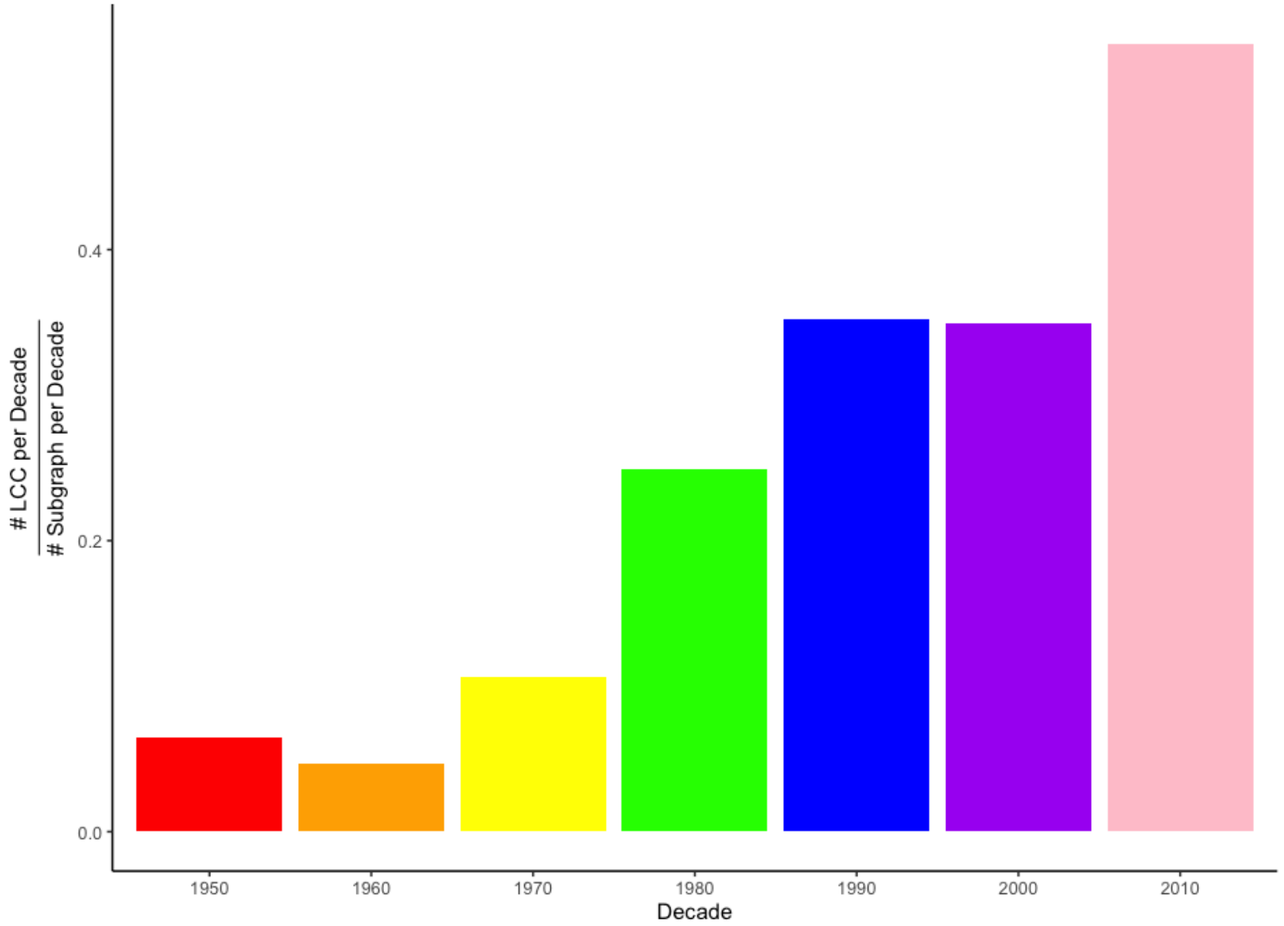
Applying thresholding (t=0.8) we reduced the nodes from 27k to around 7k
In the Top-100 Hubs 99/100 are of topic class "obscene"

LCC



In black the distribution of the decades in the LCC
In blue the same distribution in the graph

how much are same decade songs connected?



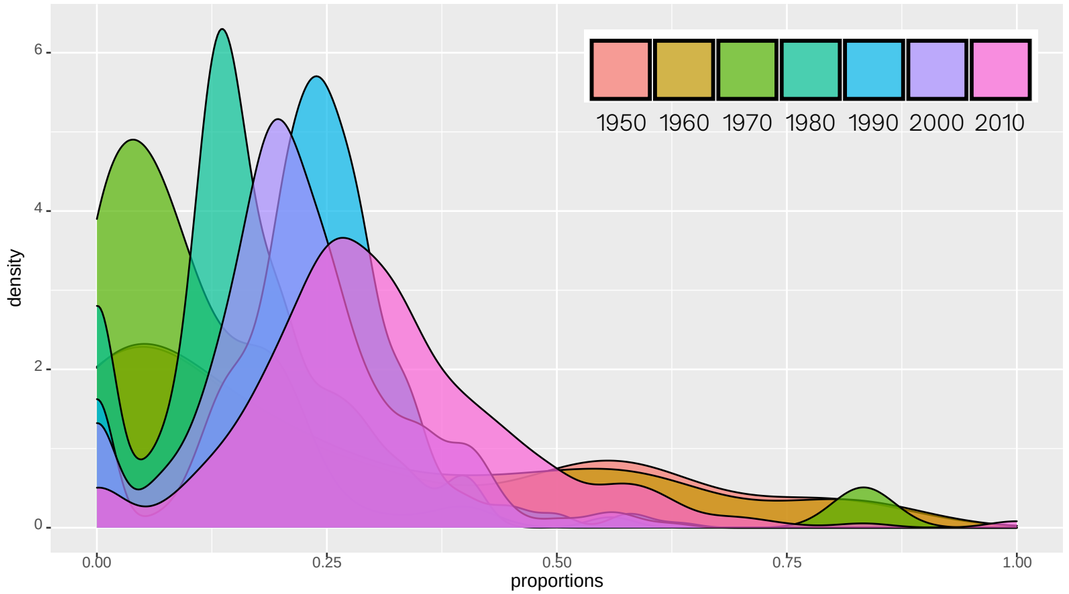
Ideally if the LCC is large wrt the entire subgraph then the songs tend to be very similar between each other.

Top-left:

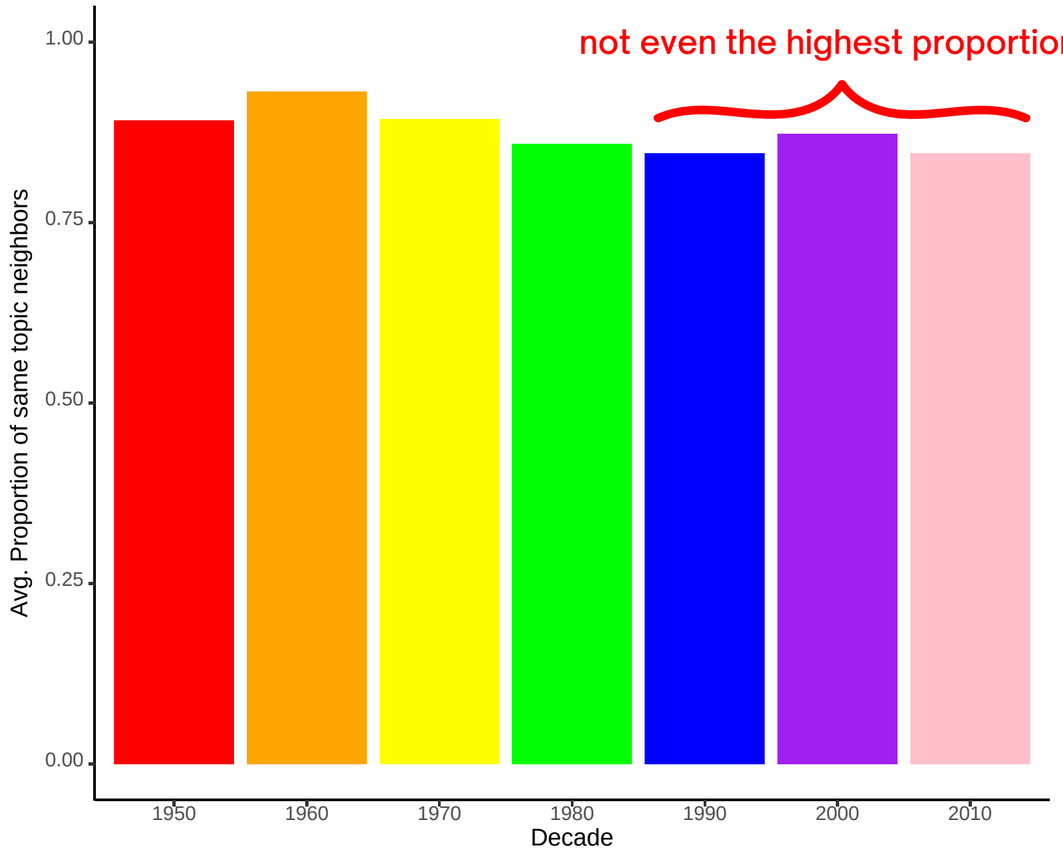
- Take subgraph for each decade
- Compute LCC for each subgraph
- Compare LCC with subgraph

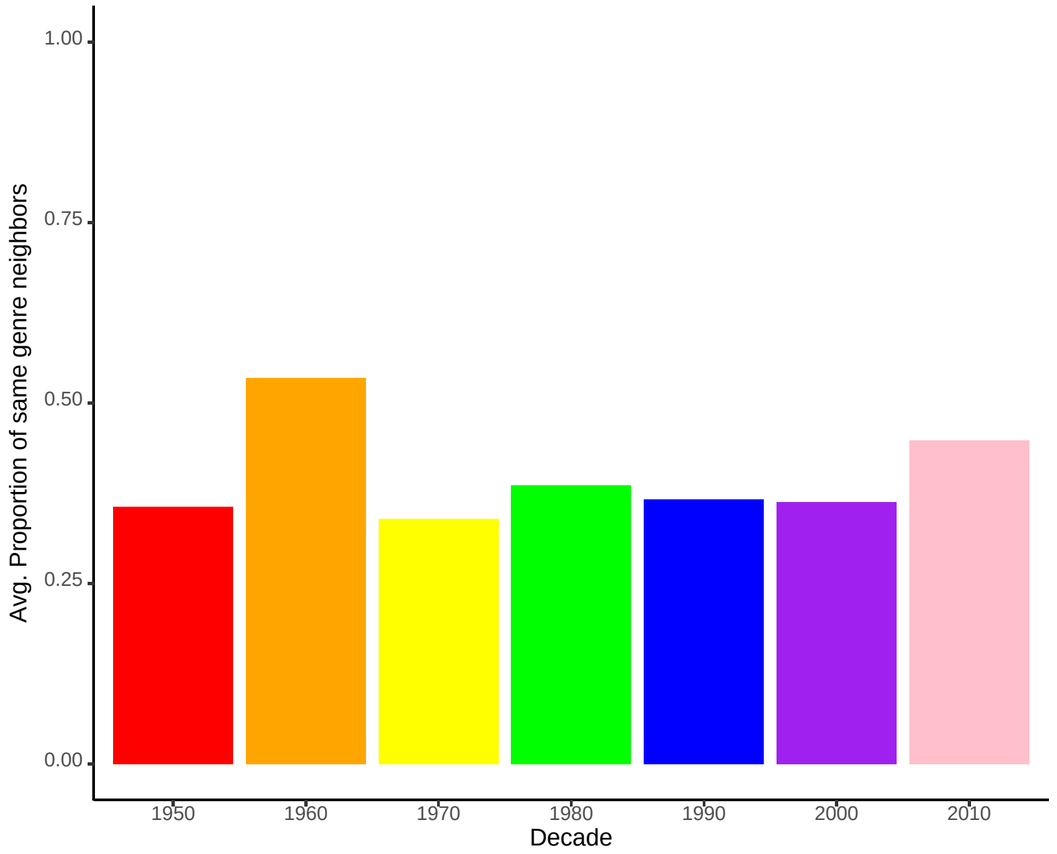
Bottom-right:

- 5-core decomposition (variance, significance)
- Proportion of same decade neighbors



"It has to be Hip Hop's fault" ... Not really!





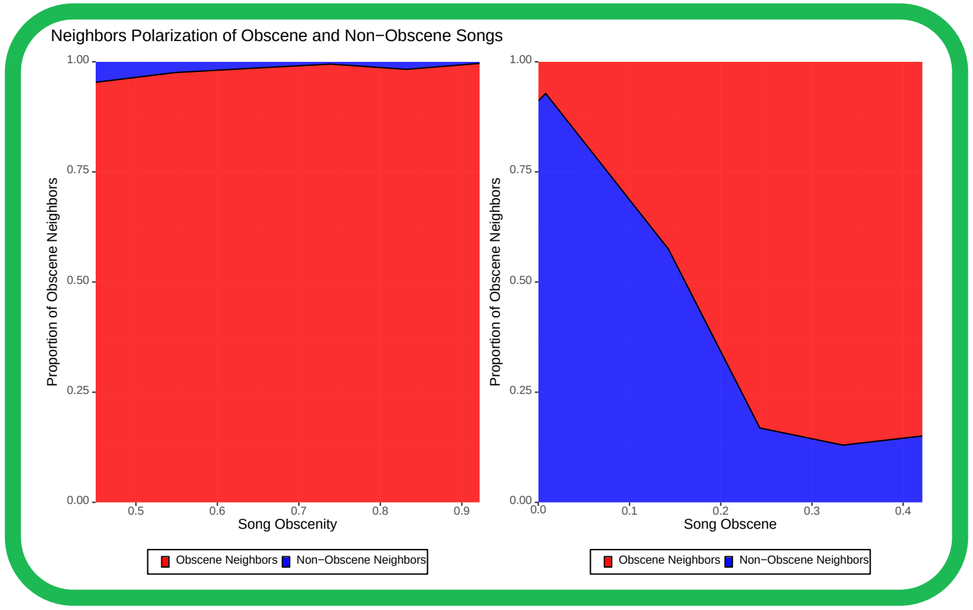
growth of hip hop → obscenity increased → increased similarity of lyrics

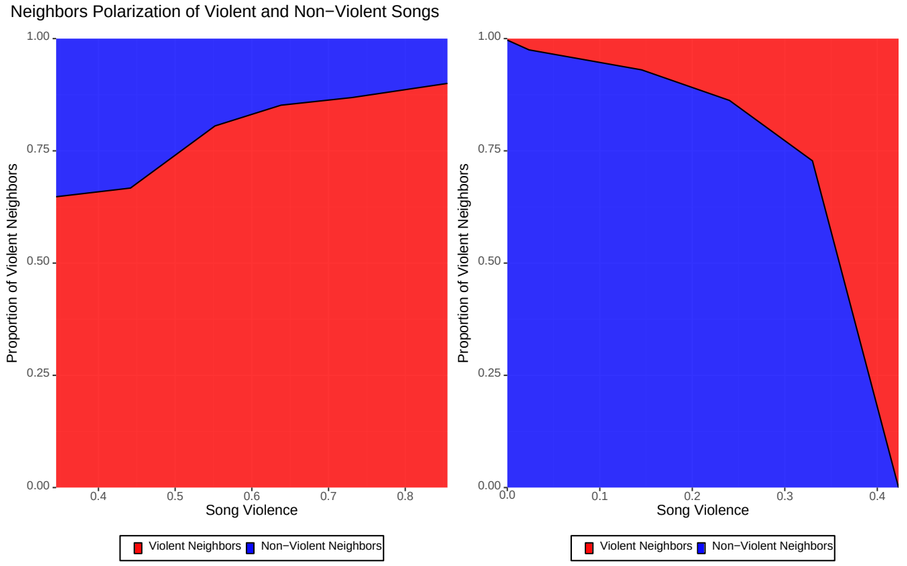
OSS: hip hop is the least represented genre per decade

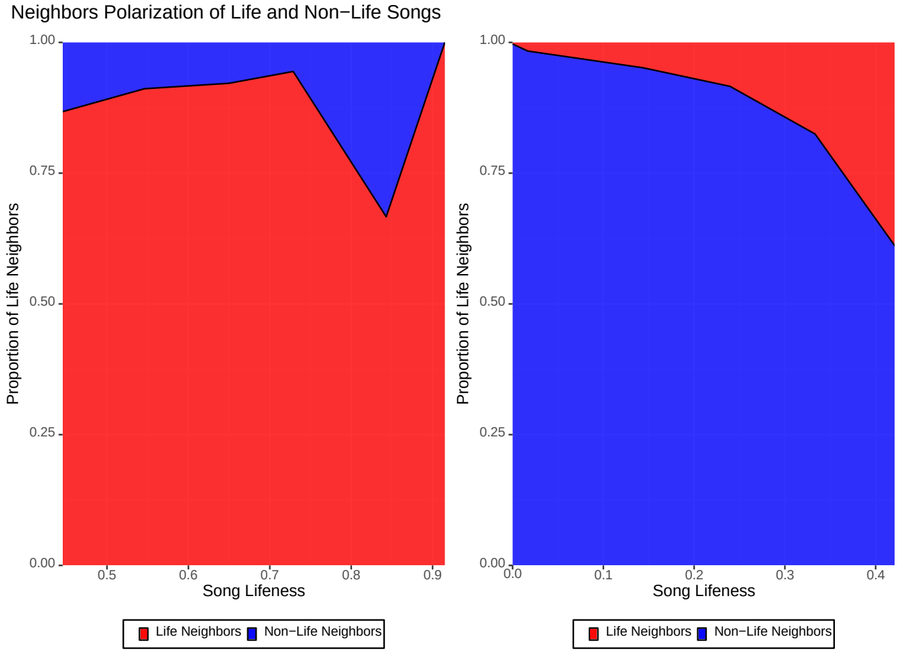
OSS (2): topics (stronger) and genres tend to **cluster together independently from the decade**

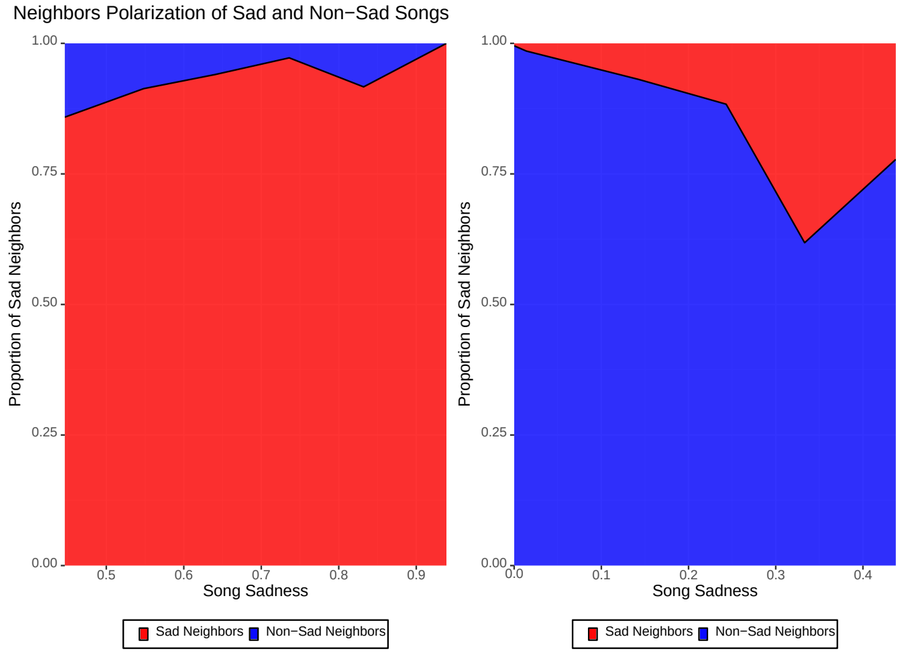
OSS (3): language, consequently (specific) topics, are the cornerstone of this change

let's take a look at the polarization of specific topics

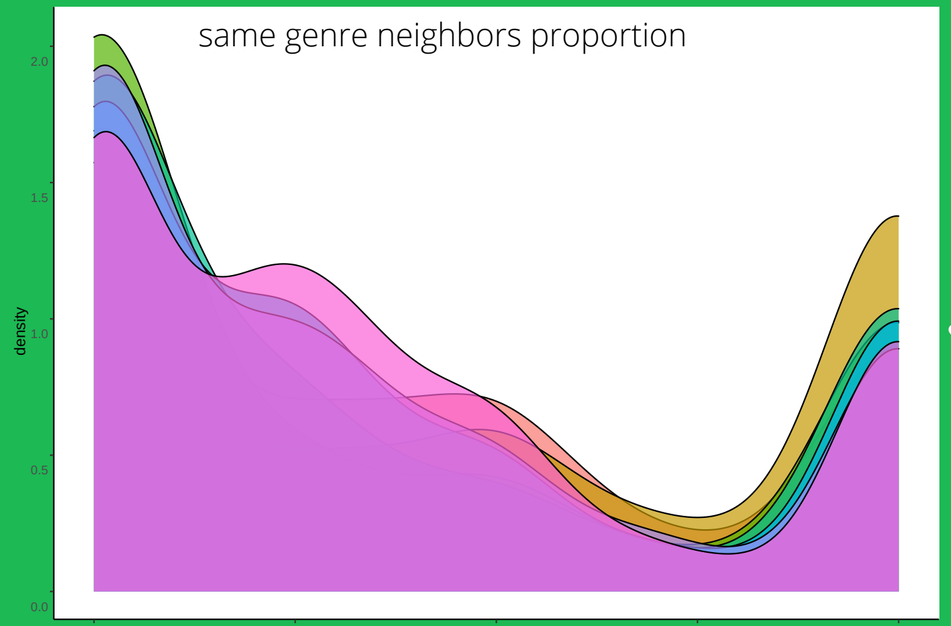






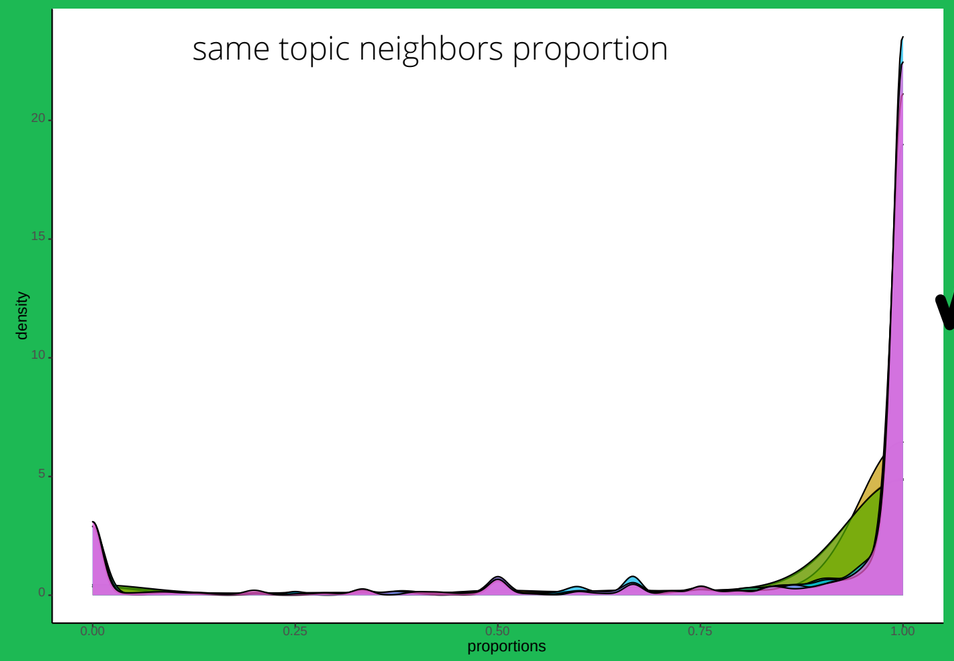


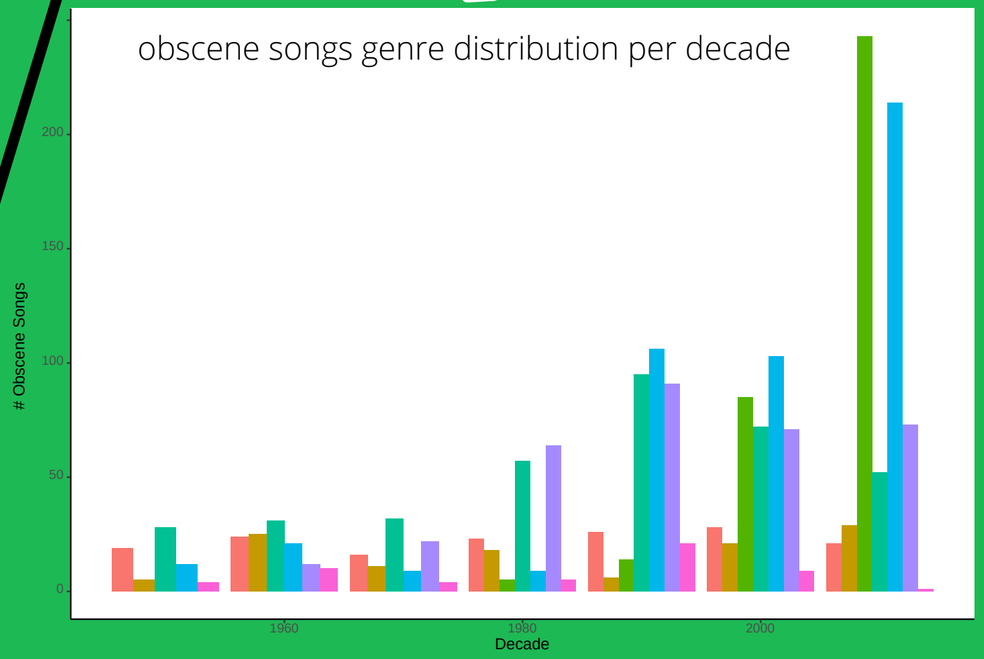
let's ask the neighbors



1950 1960 1970 1980 1990 2000 2010

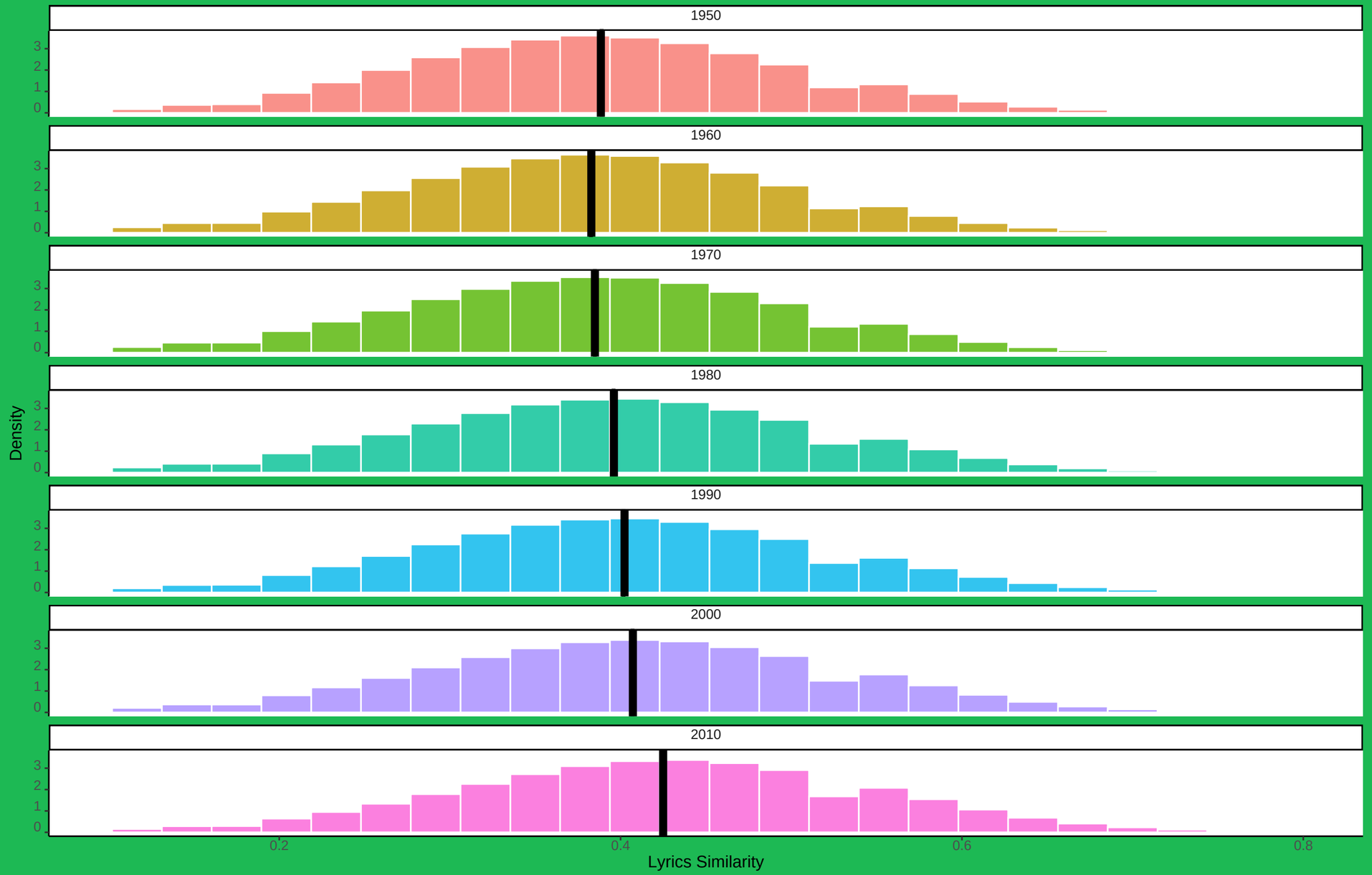
As time increases the **topic polarization** increases
but it **embraces more (relevant) genres**





blues
country
hip hop
jazz
pop
reggae
rock

cosine similarity over decades



μ = 0.388	σ = 0.111
μ = 0.382	σ = 0.111
μ = 0.385	σ = 0.113
μ = 0.396	σ = 0.117
μ = 0.402	σ = 0.117
μ = 0.407	σ = 0.119
μ = 0.425	σ = 0.120

conclusions

Analytical conclusions:

In the last graph it was clearly seen that on average songs are flattening out toward a more similar language and embracing more genres, this along with the increased production of songs could explain the slight increase in variability.

Provocative conclusions:

At a time when language has been cleared through customs we are moving toward a decay of the latter that results in texts that are repetitively vulgar and self-celebrating, not to mention devoid of a message.

Philosophical conclusions:

freedom of language allows artists more exposure to important issues and universal messages that as such address similar issues with similar language.
Let's think for example at the cultural background of Hip Hop.