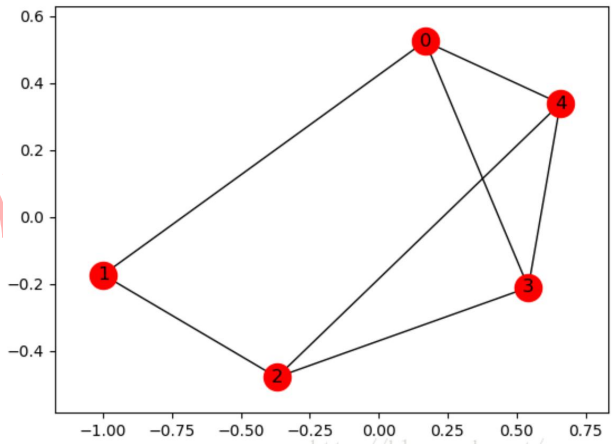
2021 美赛 D 题参考思路

对 influence_data.csv 数据集进行分析发现共有如下 20 种流派：

Avant-Garde’	Latin’
Blues’	New Age’
Children’’s’	Pop/Rock’
Classical’	R&B;’
Comedy/Spoken’	Reggae’
Country’	Religious’
Easy Listening’	Stage & Screen’
Electronic’	Unknown’
Folk’	Vocal’
Jazz’	International’

(1)该问题首先需要使用 Influence_Data 数据集或其中的一部分创建音乐影响力的(多个)定向网络，其中影响者连接到追随者。Influence_Data 中有 influencer_name 影响者有 follower_name 追随者,把影响者和追随者用一条有向边连起来,A 可能影响 B,B 可能影响 C,D.这样把他们都连接起来就可以构造多个定向的网络.通过 python 的 NetworkX 库可以用于创造、操作复杂网络，以及学习复杂网络的结构、动力学及其功能。就可以实现画出类似于如下的网络图：



通过对数据集和之前构建的网络子网进行分析就可以知道各个人的影响力和流派传播的速度以及找出流派的演化方式.

(2)该问题需要使用 FULL_MUSIC_DATA 和/或音乐特征的两个汇总数据集来研究音乐的相似性度量,题目中在 FULL_MUSIC_DATA 的基础上创建了两个汇总数据集,即根据艺术家计算的的平均值“data_by_artist”以及根据年份计算的“data_by_year”。因此我们还可以根据这些艺术家所属的流派进行统计,计算不同流派在各个条目中的平均值,进而通过标准差变异系数等手段来衡量相同流派以及不同流派的离散程度.通过对各个流派的平均值计算相关系数可以得到各

个流派间的关联性.由于变量个数较多，维度较高，分析起来比较麻烦，因此也可以采用降维的手段提取出几个重要的维度，诸如主成分分析（PCA）、随机森林、低方差滤波等等模型。

(3)该问题需要对流派进行分析,通过之前(1)中影响者和追随者之间的关联,就可以推出流派间的动态演化过程,比如A流派产生演化出B流派又演化出CD流派.通过对之前各个流派的数据分析就可以知道在流派演化的过程中它的哪些特性发生了具体的变化,流派的区别是什么,进行详细的阐述及大量的数据分析可视化才是这题的重点.

(4)该问题需要分析有影响力的人真的会影响追随者创作的音乐吗,是某些音乐特征比其他特征更具“感染力”,还是它们在影响某个特定艺术家的音乐方面都扮演着相似的角色,在(2)中分析的是流派内和流派间的相似性度量,那实际上这个问题就在分析(1)中构建的不同网络内部的相似性度量,那结论很显然在同一个网络内部的数据是相似的,在不同网络间的数据离散化程度较高,influencer对follower的创作方式具有很强烈的指导作用.

(5)该问题需要识别哪些艺术家代表变革者,那这个问题其实很简单,只要找到在同一个流派中,哪个人的风格发生了较高的变化,比如说你的网络是A->B,A->C,B->D,B->E,E->F,那你通过结合full_music_data音乐家的特征对数据进行分析发现B的风格和A对比差别较高,而受B影响的D和E它们间的风格是相似的,那么很显然B就是变革者.这里就可以借助聚类定K中“手肘法”的思想，即选取肘部对应的位置，即拐点即为变革者。

(6)该问题需要分析一种流派中随着时间发生的音乐演变的影响过程,在这里只需要详细分析一种流派的情况就可以,这就需要对full_music_data进行统计分析首先找到你要分析的流派的相关数据,然后对这些数据进行统计分析,可以通过线性回归的方式对变量进行拟合,进而观察各个变量随着时间的推进而产生的变化.

(7)可以具体分析各个指标在各个时间点的变化,找到突变位置,结合这个位置所处的时间点的历史事件进行分析,或者在你所构建的网络中去寻找这样的突变点.或者去观察音乐整体的风格的变化,结合现在社会的潮流趋势,就可以得到结论.

而且题目中说道：Note: DATA provided in these files are a subset of larger data sets. These files CONTAIN THE ONLY DATA YOU SHOULD USE FOR THIS PROBLEM.因此只能基于题目所给的数据集之上进行分析.