

选择的问题

D

2021  
MCM/ICM  
总结表控制号  
2100112

## 基于网络的音乐影响和演变分析

### 摘要

"音乐就像一个心理医生。它将用人们无法告诉你的事情来回答你。"最著名的流行/摇滚艺术家之一披头士乐队说。在本文中，我们深入探讨了音乐的魔力--它的影响网络、演变路径和对文化的影响。

对于任务1，我们在影响者和追随者之间建立了一个定向网络。基于网络工作理论，我们提出了三个不同的指标--度中心性、加权度中心性和特征中心性。然后，我们开发了这些指标的组合，作为对音乐影响力的全面衡量。之后，我们创建了一个子网络来说明我们的影响力测量。

对于任务2，我们首先使用主成分分析法对数据进行预处理，以减少维度和勾稽关系。然后，我们定义并计算不同曲目之间的距离，以获得艺术家之间的相似度。通过计算平均相似度，我们应用惠特尼测试。结果显示，在62.8%的概率和小于0.001的p值下，一个流派内的艺术家比流派间的艺术家更相似。

对于任务3，根据提议的音乐相似性衡量标准，我们发现，当我们分析不同的流派时，流派内部和流派之间的相似性和影响有很大不同。为了区分一个流派和其他流派，我们建立了一个流派分类树模型。通过分析不同时期的影响者的数量，我们探索了流派的演变路径。基于流派规模的定向网络，我们发现流行/摇滚与R&B、蓝调和民谣有着密切的关系。

对于任务4，我们建立了一个模型，根据曲目中的音乐特征相似性来识别真正的追随者。然后我们应用了一个多变量的双样本平均数测试，并发现没有强有力的证据表明任何音乐特征都比其他特征更具"传染性"。

对于任务5，我们首先分析了流派的兴衰，找到了1950年代的音乐革命。我们提出了一个动态编程算法来检测与革命一致的音乐特征的变化点。我们的结果显示，声学性、能量、舞蹈性和响度可能标志着革命。基于贝叶斯网络，猫王和克里夫-理查德代表了革命者。

对于任务6，为了进一步了解流行/摇滚乐的演变，我们根据整个流派的音乐特征的滞后趋势，提出了一个动态影响者指标。从1960年代到2010年代，有10个动态影响者，每个人都对该流派有其独特的影响。此外，我们还解释了流行/摇滚乐的演变。

对于任务7，根据时间序列分析发现了三个重要的时期，这显示了音乐的文化影响。根据建立的模型，我们确定了社会变化，如反文化运动和技术变化，如互联网的普及。

最后，我们进行了敏感性分析，这表明了我们模型的稳健性。我们还总结了其中的长处和短处，并向ICM社会提供了关于音乐的演变和文化影响的见解。

**关键词：**定向网络；贝叶斯网络；变化点分析定向网络；贝叶斯网络；变化点分析

# 内容

1	介绍	3
2	问题重述与分析	3
3	假设和记号	4
3.1	假设和理由 .....	4
3.2	符号.....。	5
4	模型和解决方案	5
4.1	任务1 .....	5
4.1.1	定向影响者网络.....。	5
4.1.2	音乐影响的衡量标准。	6
4.1.3	解决方案 .....	7
4.2	任务2 .....	7
4.2.1	数据预处理 .....	7
4.2.2	相似性测量和测试.....。	8
4.2.3	解决方案.....。	9
4.3	任务3 .....	10
4.3.1	体裁的相似性和影响。	10
4.3.2	体裁分类树.....。	10
4.3.3	解决方案.....。	10
4.4	任务4 .....	13
4.4.1	相似性贝叶斯网络.....。	13
4.4.2	传染性特征测试.....。	14
4.4.3	解决方案.....。	15
4.5	任务5 .....	16
4.5.1	革命的定义.....。	16
4.5.2	变化点检测（DP算法） .....	16
4.5.3	解决方案.....。	17
4.6	任务6 .....	19
4.6.1	动态影响者指标 .....	19
4.6.2	解决方案 .....	20
4.7	任务7 .....	21
4.7.1	音乐的文化影响 .....	21
4.7.2	在网络内确定的变化 .....	21
5	敏感度分析	22
6	优势和劣势	23
6.1	优势 .....	23
6.2	弱点 .....	23
	参考文献	23
	给ICM协会的一份文件	24

## 1 简介

如今，各种音乐已日益成为人类生活中不可缺少的一部分。数以千计的音乐艺术家相互影响，形成了一个复杂的音乐影响网络。一些流派之间表现出极大的相似性，而其他流派的音乐特点则大相径庭。一些艺术家是充满激情的革命者，他们导致了一种新流派的出现或现有流派的重塑。尽管受到文化的影响，艺术家的音乐特征的变化也表明了外部事件，如互联网的普及。为了进一步探索音乐影响网络和音乐对社会所起的作用，有必要对音乐演变进行量化。

## 2 问题重述与分析

- 任务1要求我们根据数据集 "influnce\_data.csv" 在影响者和追随者之间建立一个复杂的网络，并制定指标来捕捉网络中的音乐影响力。这个问题的关键是定义有导向的影响者网络，并提出全面衡量网络中每个影响者的音乐影响力的指标。
- 任务2要求我们利用数据集 "full\_music\_data.csv" 中的各种音乐特征来衡量音乐的相似性，判断同一流派的音乐家是否比不同流派的音乐家更相似。定义艺术家的音乐作品之间的距离是非常重要的，我们可以从中获得艺术家之间的相似性。
- 任务3要求我们比较流派之间和流派内部的相似性和影响。有必要定义体裁之间的距离并衡量它们之间的相似性。我们计划建立一棵分类树，将一种流派与其他流派区分开来。此外，我们还可以用可视化的方式展示流派如何随时间变化以及流派之间的关系。
- 任务4要求我们进一步深入了解影响者和追随者之间的音乐特征的相似性。我们计划建立一个贝叶斯网络来寻找真正的影响者，并使用假设检验来评估一些音乐特征是否比其他的更具有 "传染性"。
- 任务5要求我们确定标志着音乐发展的特征和主要艺术家。为了处理这个问题，我们提出了一个基于DP算法的变化点检测模型，将音乐人物的变化与音乐革命相匹配。然后通过贝叶斯网络找到革命者。
- 任务6要求我们制定指标来分析一个流派中音乐演变的动态影响者和相应的影响过程。在这个问题上，我们将关注流行/摇滚乐，然后讨论几十年来流行/摇滚乐演变的主要贡献者。
- 任务7要求我们确定网络内社会、政治或技术变化的影响，并找到音乐的文化影响。基于时间序列分析，我们打算找到音乐特征的变化与外部事件之间的联系。我们也会在不同的时间或环境中发现音乐的几种文化影响。

本文的工作流程如图1所示。

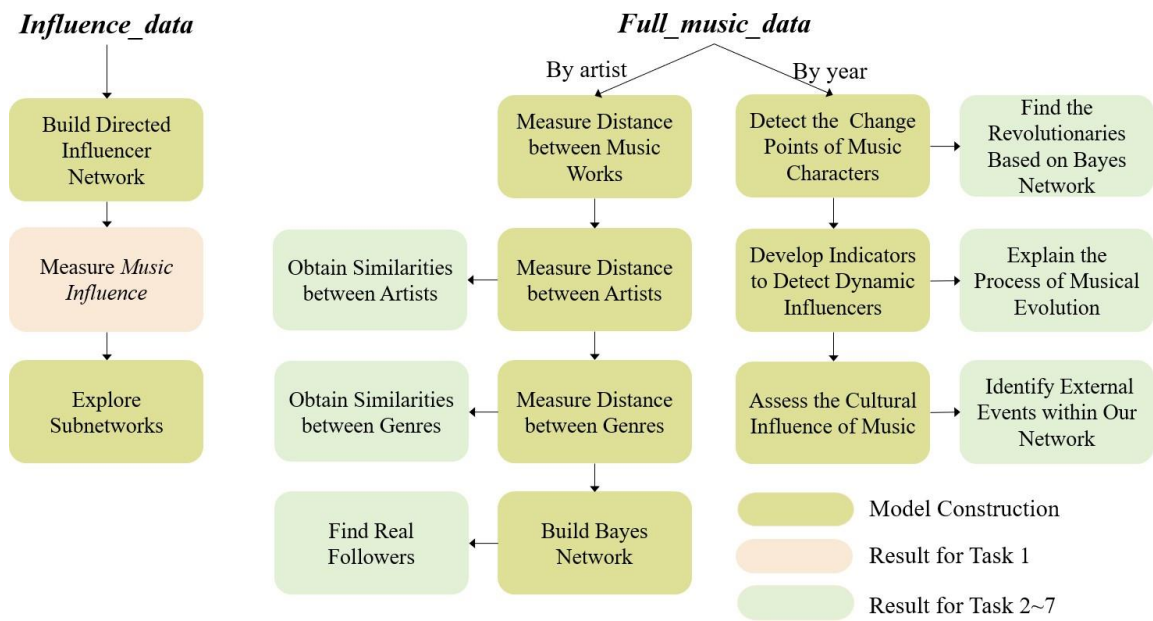


图1：本文的工作流程

### 3 假设和校验

#### 3.1 假设和理由众号

- 艺术家的相似性由其音乐特征的相似性来表示。在所有可用的数据集中，关于艺术家特征的最可靠的信息来源是他/她发布的曲目。因此，使用音乐特征来代表艺术家的特征是合理的。
- 现有流派的再创造和革命可以通过音乐特征的急剧变化来体现。一个流派的革命转变将无限地改变音乐特征，因此我们可以根据这些变化来捕捉革命。
- 在流派发展的不同阶段，音乐角色随着时间的推移而发生线性变化。这个假设是一个合理的简化，使断点识别成为可能。

3.2 记号

主要记号见表1。

表1：记号

符号	定义
$DC_i$	第 <i>i</i> 个影响者的本地影响程度中心性
$WDC_i$	第 <i>i</i> 个影响者的加权重中心性
$EC_i$	第 <i>i</i> 个影响者的特征中心度
$F-Score_i$	第 <i>i</i> 个影响者的综合得分。
$Sim_{i,j}$	曲目 <i>i</i> 和曲目 <i>j</i> 之间的音乐相似度
$Acv_{i,j}$	体裁 <i>i</i> 中音乐特征 <i>j</i> 的绝对变异系数
$\rho_{AB}$	艺术家 <i>A</i> 和艺术家 <i>B</i> 之间的相似度得分
$\mu$	滞后年份的长度

4 模式和解决方案

4.1 任务1

4.1.1 指导影响者网络

我们将影响者和追随者视为节点，并将所有音乐人收集在一起，以使用这个集合。如果艺术家*i*对艺术家*j*有影响，那么从节点*i*到节点*j*的边就是将被生成。所有的边构成了 $E = (e_{ij}^{lm})$ 的集合，节点*i*的边构成了集 $N(i)$ 。基于数据集 "influnce\_data.csv"，我们建立了一个复杂的网络，其中包括  $n=5603$  个节点（艺术家）和  $m=42770$  条边（影响），如图2所示。

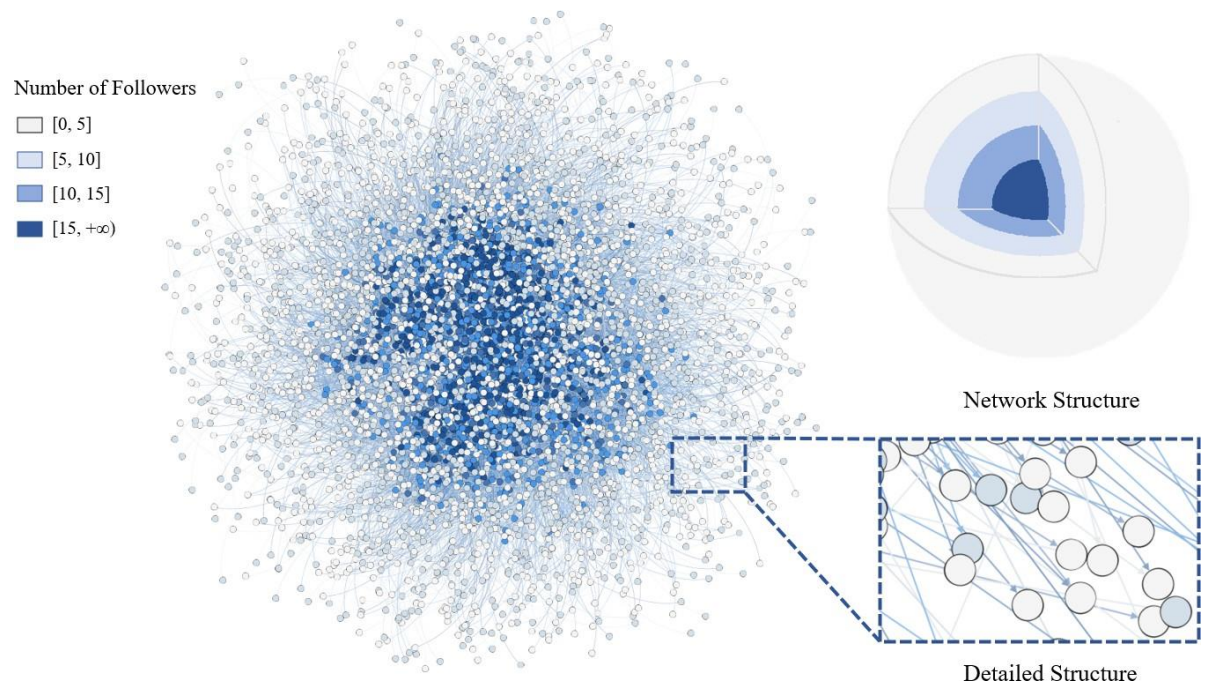


图2：三维定向影响者网络



### 4.1.2 音乐影响的衡量标准

我们现在定义一些基本的指标来衡量网络中的音乐影响力。

**学位中心性。**度是网络理论中的一个重要概念。在有向图中，节点 $v$ 的出度代表了来自节点 $v$ 的边的数量[1]。

$$\text{过度度}_v = \#N(v) \quad (1)$$

衡量节点的本地重要性的一个天真的想法是节点的出场度。换句话说，我们用追随者的数量来衡量一个音乐人的本地影响力。我们把

本地影响的度中心，并定义了以下的度中心 **影响者I** 如下：

$$DC_i = \text{outdegree}_i / n \quad (2)$$

其中 $n$ 是网络中的节点数。

某些流派之间的互动密切，而其他流派似乎没有什么联系。因此，我们不能平等对待所有的影响，网络的边应该分享不同的权重。如果一个音乐家对其他流派的艺术家有影响，这意味着该音乐家有广泛的影响力。同样，如果一个音乐家在几十年后对未来的音乐世代产生了影响，这表明这种影响持续了很长一段时间。在上述情况下，我们对准更大的权重。

我们将音乐人 $i$ 的流派定义为 $G(i)$ 。追随者和影响者之间的年差被定义为他们职业生涯开始时的时间差。当年差超过阈值时，我们称其为长期影响，否则为短期影响。这里我们设定阈值=20。

权重矩阵 $W = (W_{ij})$ 是通过整合影响范围和影响期限来定义的，如下所示。

$$W_{ij} = \begin{cases} \frac{1}{3} * (1 + I_{\{S(i) \neq S(j)\}} + I_{\{yeargap > threshold\}}) & j \in N(i) \\ 0, & \text{否} \end{cases} \quad (3)$$

然后，我们提出了加权度中心论（WDC）来修改上述度中心论。

$$wdc_i = \frac{1}{n} * \langle w_i, 1_n \rangle \quad (4)$$

其中， $w_i$ 代表矩阵 $W$ 的第 $i$ 行， $1_n$ 是所有条目为1的列向量。DC和WDC都是衡量一个影响者的局部影响力。

特征中心性的基本思想是把一个节点的影响力看作是其相邻节点的局部影响力的函数。换句话说，艺术家的追随者对其他人的影响越大，艺术家本人的特征中心度就越大。

特征中心度的定义如下。

$$ec_i = \frac{1}{n} * \langle w_i, od \rangle. \quad (5)$$

其中 $OD = (\text{outdegree}_1, \text{outdegree}_2, \dots, \text{outdegree}_n)^T$ 和 $W_i = (W_{i1}, W_{i2}, \dots, W_{in})^T$ 都去注意一个列向量。直观地说，Eigen Centrality将相邻节点的Degree Centrality按比例分配给所有节点，这似乎可以“分散”Degree Centrality。

**综合F-Score。**上述三种不同的程度从不同的方面衡量了网络中艺术家的音乐影响力。为了得到各方面的综合衡量

影响力，我们使用三个程度的加权和。为了衡量相对值，在加权总和之前，每个程度都要除以相应的最大值。

$$F\text{-Score}_i = w_1 * \frac{DC_i}{\max_k(DC)_k} + w_2 * \frac{WDC_i}{\max_k(WDC)_k} + w_3 * \frac{EC_i}{\max_k(EC)_k} \quad (6)$$

F-score，作为所有三个指标的组合，全面衡量复杂网络中每个影响者的音乐影响力。

### 4.1.3 解决方案

根据音乐影响力测量的F分数（这里我们设定  $w_1 = w_2 = w_3 = 1$ ），我们从原始网络中提取了10个有向影响者子网络，如图3所示。这10个影响者的所有指标如表2所示。每个小的子网络中核心的大小表示顶级艺术家的音乐影响力。很明显，子网络显示了一个辐射结构，以一个伟大的艺术家为中心，连接到他的追随者。

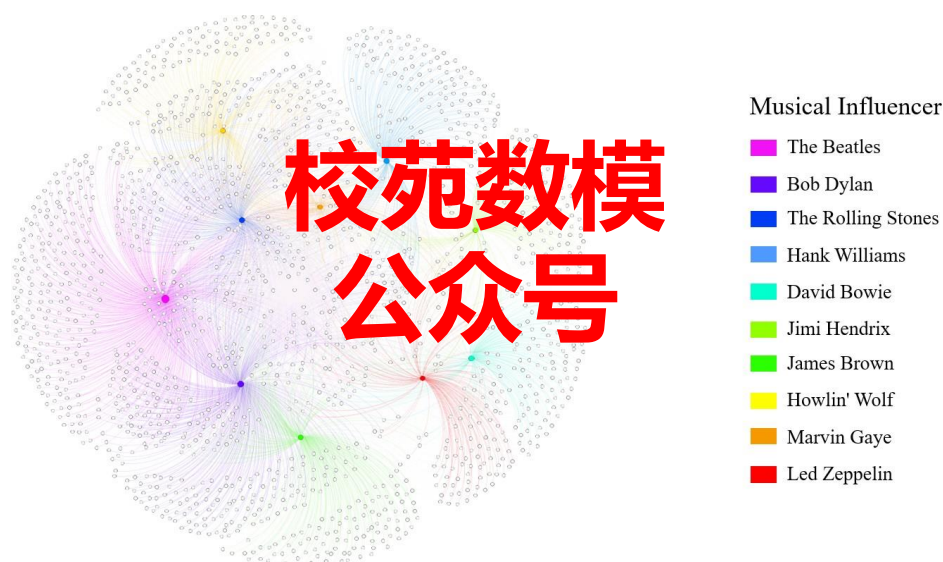


图3：定向影响者子网络

一般来说，音乐影响力大的艺术家有更多的追随者。然而，Howlin' Wolf的追随者比Marvin Gaye少，而他的音乐影响力更大。这是因为Howlin' Wolf的追随者更有名，更有影响力。

## 4.2 任务2

### 4.2.1 数据预处理

首先，我们对数据集 "full\_music\_data.csv" 进行预处理。数据集中包括14个与音乐有关的特征，包括舞蹈性、能量、响度等。通过数据可视化，我们发现布尔变量 "Explicit" 是无效的。在98430首歌曲中，只有3647首歌曲被标记为1，这意味着只有不到4%的歌曲有明确的歌词。由于它非常罕见，我们删除了它。之后，我们对所有的连续变量进行标准化。

表2：前10名艺术家的衡量标准

同上	命名	类型	直流电	WDC	EC	F-Score
754032	披头士乐队	流行/摇滚	0.11	0.14	2.3	1.00
66915	鲍勃-迪伦	流行/摇滚	0.07	0.1	1.65	0.68
894465	滚石乐队	流行/摇滚	0.06	0.07	1.24	0.52
549797	汉克-威廉斯	国家	0.03	0.07	1.57	0.49
531986	大卫-鲍伊	流行/摇滚	0.04	0.06	0.71	0.37
354105	Jimi Hendrix	流行/摇滚	0.04	0.05	0.94	0.37
128099	詹姆斯-布朗	R&B	0.03	0.05	1.11	0.36
276085	豪林狼	蓝调	0.02	0.04	1.3	0.35
316834	马文-盖伊	R&B	0.03	0.06	0.72	0.34
139026	齐柏林飞船	流行/摇滚	0.04	0.05	0.65	0.34

一些音乐特征具有相似的含义，例如，“能量”和“响度”都反映了曲目的强度和活动。为了减少串联性的影响，在计算相似性时，我们使用PCA (主成分分析)来缩小在尽可能多地保留数据的变异的情况下，对数据进行分割。经过计算，累积方差贡献率如表3所示。

表3：累积差异贡献率

主要成分的数量	1	2	模3	4	5	6	7	8
累积差异 缴款率	0.23	0.37	0.49	0.59	0.67	0.74	0.81	0.85

根据PCA的结果，我们选择前七个主成分，而忽略其余的。这七个新变量保持了原始数据80%以上的信息。

4.2.2 相似性测量和测试

受欧氏距离的启发，我们将轨道*i*和轨道*j*之间的音乐相似度定义如下。

模拟=

$$ij = \frac{1}{\sum_{t=1}^m x_{it} - \bar{x}_i} \frac{1}{\sum_{t=1}^m x_{jt} - \bar{x}_j} \quad i, j = 1, 2, \dots, m \tag{7}$$

与Mahalanobis距离等其他测量方法相比，这个简单的测量方法不需要对数据进行任何假设，而且很容易计算。

探讨同一流派内的艺术家更相似。我们构建曼-惠特尼假设检验。与传统的检验方法如双样本t检验相比，曼-惠特尼统计量的有效性不需要分布假设。在这个问题上，音乐相似度在流派内部和流派之间的分布与正态性相差甚远。因此，这种统计方法会带来更好的结果。

我们首先计算流派内和流派间的平均音乐相似度，这可以



可表示为

$$Sim_{in,i} = \frac{1}{n} \sum_{j=1}^n Sim_{ij} \quad (8)$$

$$Sim_{bet,i} = \frac{1}{m} \sum_{k=1}^m Sim_{ik} \quad (9)$$

其中  $j$ 、 $k$  表示与艺术家  $i$  同属一个流派和不同流派的个体， $n$ 、 $m$  分别表示上述个体的数量。

然后我们计算曼-惠特尼统计量为：

$$U = \frac{1}{(n+m)^2} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} I(Sim_{in,i} < Sim_{bet,j}) \quad (10)$$

根据中心极限定理， $Z = \frac{U - \frac{n(n+m+1)}{2}}{\sqrt{\frac{n(n+m+1)}{12}}}$  的渐进分布为正常的分布[4]。在此基础上，我们可以构建Mann-Whitney Test，以发现一个流派内的艺术家是否比流派之间的艺术家更相似。

#### 4.2.3 解决方案

所有艺术家之间的相似性显示在图4中。Mann-V 统计量显示，流派内的艺术家比流派之间的艺术家更相似，具有 52.8% 的概率和小于 0.001 的 P 值。

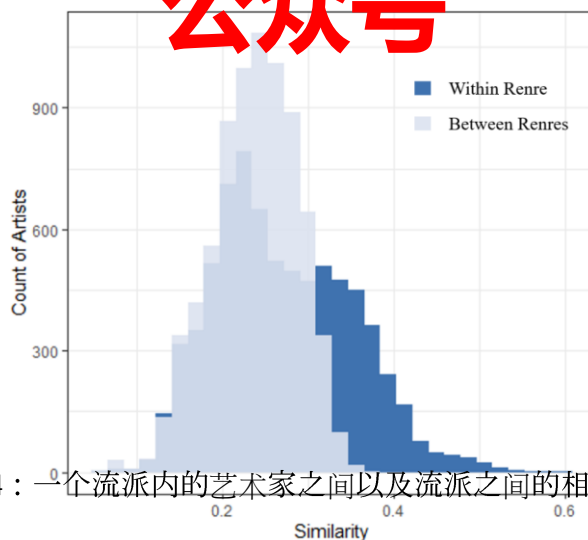


图4：一个流派内的艺术家之间以及流派之间的相似性

### 4.3 任务3

#### 4.3.1 体裁的相似性和影响

为了比较音乐的相似性和流派层面的影响，我们首先合并了数据集中的流派 "influnce\_data.csv" 到数据集 "data\_by\_artist.csv"，然后我们按不同流派将数据分组。

我们将流派之间的相似性定义为

$$模拟_{ij} = \frac{1}{\sum_{k=1}^{n_g} c_{ik} - c_{jk}} \quad i, j = 1, 2, \dots, g \quad (11)$$

其中  $n_g$  是流派的数量， $c_{ik}$  是流派  $i$  中的第  $k$  个平均音乐特征值。

为了评估流派内的相似性，我们计算每个音乐特征在不同流派中的绝对变异系数，它量化了流派内某些特征的变异。让  $Acv_{ij}$  表示音乐特征  $j$  在流派  $i$  中的绝对变异系数，那么。

$$Acv_{ij} = \frac{\sqrt{\frac{1}{G(i)} \sum_{k \in G(i)} x_{jk}^2 - \left( \frac{1}{G(i)} \sum_{k \in G(i)} x_{jk} \right)^2}}{\frac{1}{\sum_{k \in G(i)} x_{jk}}} \quad (12)$$

其中， $x_{jk}$  表示艺术家  $k$  的音乐角色  $j$ 。通过对所有继续音乐角色的  $Acv$  进行平均，我们得到每个流派的平均绝对变异系数。

#### 4.3.2 体裁分类树

根据我们的假设，当我们把一个流派与另一个流派区分开来时，主要的区别在于创作风格--艺术家创作的音乐的各种特征。考虑到每个流派都有其独特的风格，我们分别分析了每个流派的显著特征。

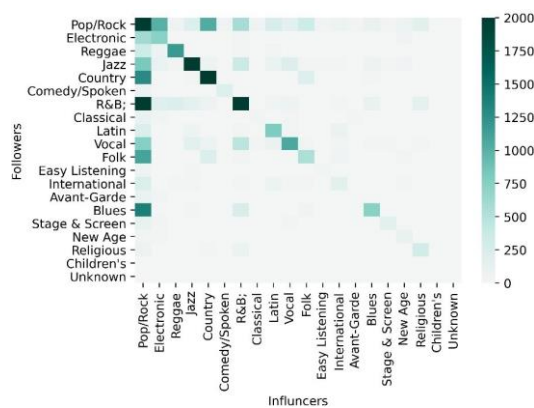
分类树可以区分出特征的相对重要性。靠近根节点的特征比靠近离开节点的特征更重要。构建流派分类树的具体步骤如下。

1. 确定要分析的流派。将该流派中的艺术家标记为正面，其余艺术家标记为负面。
2. 构建分类树并选择适当的树的大小，以保持分类标准的简单性。
3. 根据特征在树中的相对位置对其重要性进行排序，将树可视化，并提供区分策略，以从其他类型中识别出该类型。

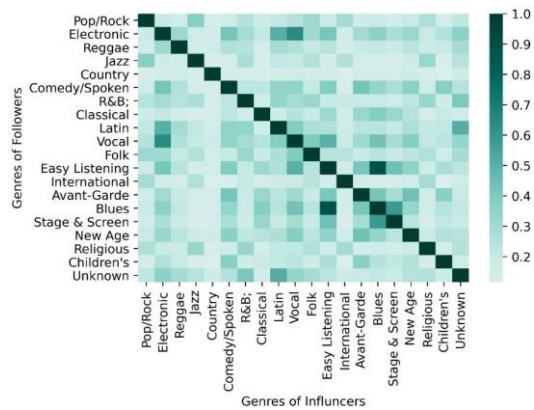
#### 4.3.3 解决方案

图12显示了流派之间的相似度矩阵和影响矩阵。蓝调和轻松音乐在音乐特征上有一些相似的特点，而电子和声乐则更为相似。在影响方面，流行/摇滚乐对其他类型的音乐有很大影响。互相影响。

流行/摇滚乐对R&B有很大的影响，但它们并没有明显的相似性在音乐特性方面，我们将在后面讨论其原因。



(a) 体裁之间和内部的影响



(b) 体裁之间的相似性

图5：流派之间和流派内部的影响和相似性

图6和图12a显示了一个流派内的相似性和影响力。一些流派如R&B和乡村音乐在流派内比其他流派有更多的变化，这表明艺术家渴望真正的自由表达。同时，流行/摇滚和爵士乐对自己的影响比其他流派更大，这意味着他们倾向于保持一个一致的风格。

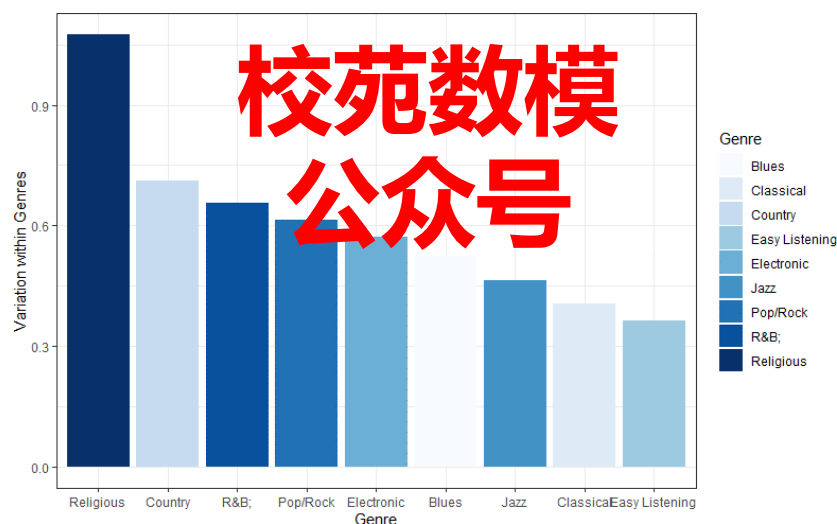


图6：流派内的相似性

以R&B为例，我们构建的分类树如图7所示。舞蹈性、持续时间和工具性是R&B区别于其他流派的重要特征。基于这种方法，我们可以找到每个流派的重要特征。

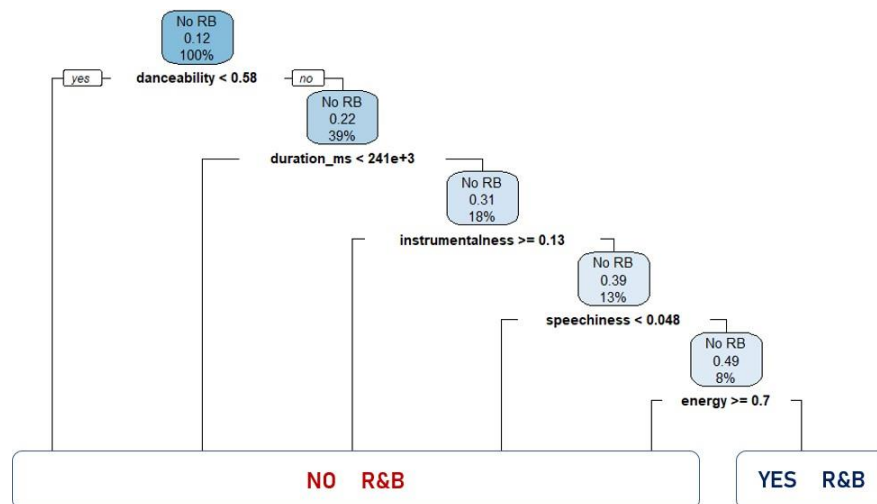


图7：R&amp;B分类树

我们使用数据集 "influence\_data.csv" 来描述流派如何随时间变化。根据不同时间段内不同流派的影响者数量，我们绘制了图8。正如图中所示，蓝调和爵士乐从1930年代到1950年代蓬勃发展，然后遇到平稳的下降。另一方面，乡村音乐从1930年代到1990年代一直流行。流行/摇滚，类似于

**R&B**，在1960年代至1990年代期间落

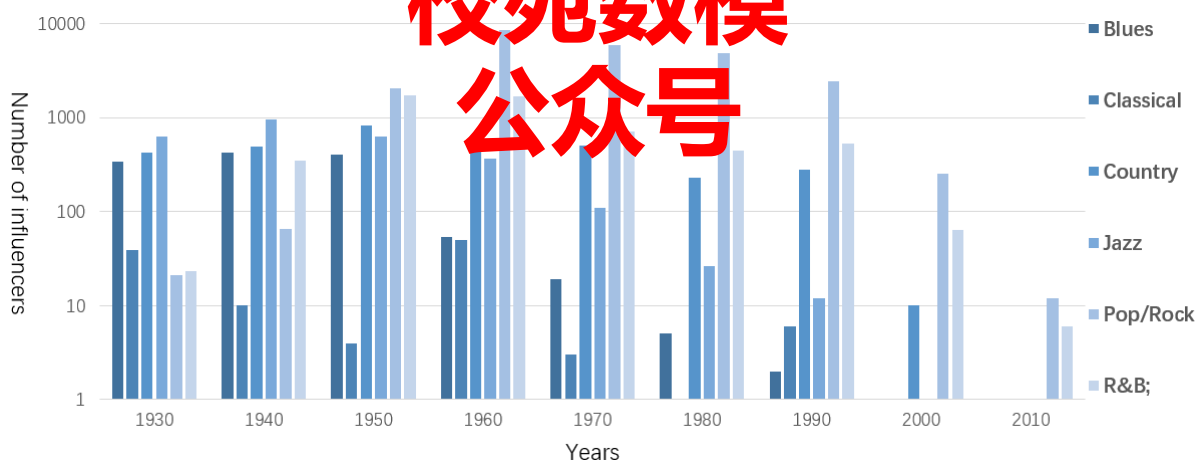


图8：流派随时间变化（部分）。

为了进一步探索不同流派之间的关系，我们在影响者网络的基础上构建了一个有向流派网络，如图9所示。直观地说，如果一个流派与另一个流派密切相关，艺术家往往对对方的曲目有很大的影响。因此，我们认为“音乐影响力”是衡量关系的主要标准。

类型网络表明，流行/摇滚艺术家对许多其他类型有很大影响，特别是R&B、蓝调和民谣等，将这些类型联系在一起。此外，流行/摇滚艺术家还深受电子音乐的影响。从流行/摇滚音乐家那里学习，R&B的艺术家与前卫音乐的艺术家有许多相互作用。

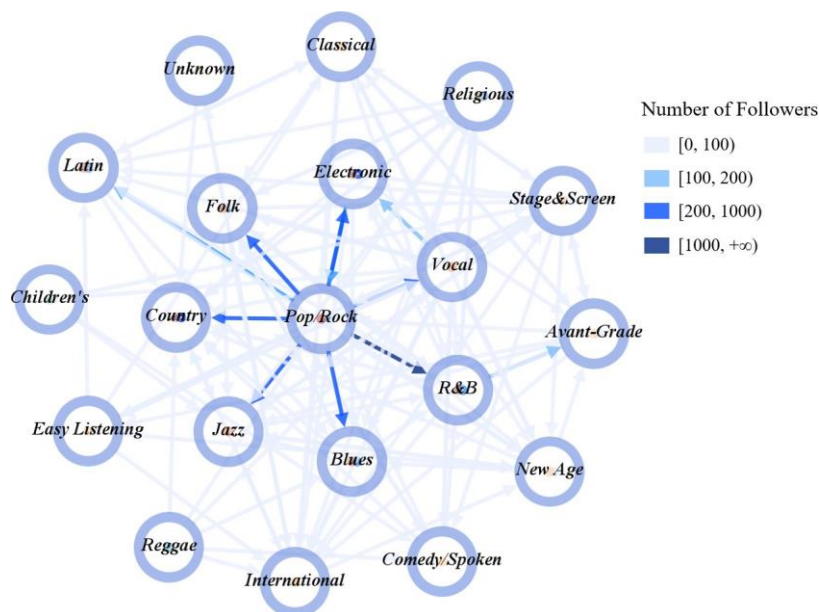


图9：流派网络

## 4.4 任务4

# 校苑数模

### 4.4.1 相似性贝叶斯网络

本小节旨在探讨所谓的“影响者”是否真的对“追随者”创作的音乐有影响，换句话说，影响者的音乐和他们的追随者创作的作品之间是否有明显的相似度。

为了解决这个问题，我们构建了一个基于相似性贝叶斯网络。在这里，我们把艺术家A和艺术家B之间的相似度分数定义为

$$\rho_{AB} = \frac{\sum_i^{n_c} \text{Cha}_{Ai} - \text{Ch}^{-a_A} \text{Cha}_{Bi} - \text{Ch}^{-a_B}}{\sqrt{\sum_i^{n_c} \text{Cha}_{Ai} - \text{Ch}^{-a_A}}^2 \sum_i^{n_c} \text{Cha}_{Bi} - \text{Ch}^{-a_B}}^2} \quad (13)$$

其形式与皮尔逊相关相似， $n_c$  表示音乐特性的数量， $\text{Cha}_{Ai}$  表示艺术家A的第1个标准化音乐特性得分， $\text{Ch}^{-a_{Ai}}$  表示平均值。

在我们的贝叶斯网络中，节点表示艺术家，边表示相似度分数。根据Fisher Z-Test，如果A和B在C的条件下没有任何相似性，那么

$$\frac{\frac{n - |\text{ChaC}| - 3}{2} \log \frac{1 + \rho_{AB} |\text{C}|}{1 - \rho_{AB} |\text{C}|}}{\sqrt{\frac{n - |\text{ChaC}| - 3}{2}}} \xrightarrow{d} N(0, 1) \quad (14)$$

基于渐进分布，我们可以计算出任何两位艺术家之间的相似度分数，然后构建一个相似度贝叶斯网络。然而，这种遍历算法有NP

问题。因此，我们使用贝叶斯爬山算法用随机启动代替。定义息作为

$$BIC = \sum_{i=1}^{n_c} \log f_A(i | \Pi_A(i)) - \frac{d}{2} \log(n) \quad (15)$$



---

**算法1** 相似性贝叶斯网络
 

---

```

虽然  $i \leq \text{Max\_RandomNumber}$  do
  随机生成一个相似度贝叶斯网络 计算  $BIC\_init$ 
   $BIC\_best \leftarrow BIC\_init$ 
  虽然  $j \leq \text{Max\_iterations}$  do
    随机选择添加或删除一条边 计算  $BIC\_new$ 
    如果  $BIC\_new \leq BIC\_best$ , 那么
       $Net\_best \leftarrow \text{CurrentBayesianNetwork}$ 
       $BIC\_best \leftarrow BIC\_new$ 
    结束 如果
  End
while end
while
  对每个  $i$  的  $BIC\_best$  进行排序, 找到最佳的相似性贝叶斯网络 后期处理结果和
  可视化

```

---

其中  $\log f_{A_i}(A_i | \Pi_{A_i})$  是条件密度函数, 算法的伪代码显示如下。

为了提高解决方案的稳定性, 我们使用自举法来计算500个不同的  
解决方案, 并取一个平均值

。

#### 4.4.2 传染性特征测试

在本小节中, 我们将分析一些音乐特征是否比其他的更具有 "传染性", 或者它们在影响一个特定艺术家的音乐方面都有类似的作用。基于上一小节的结果, 我们通过分析披头士乐队和他们的主要追随者来解决这一问题。海滩男孩、齐柏林飞船、格拉姆-帕森斯和大卫-鲍伊。

为了找出是否存在明显比其他特征更具传染性的特征, 我们应用多变量双样本检验

$$H_0: \mu_1 = \mu_2 \quad (16)$$

其中,  $\mu_1$  表示披头士乐队发行的所有曲目的平均音乐特征向量。

$\mu_2$  表示其一个追随者的平均音乐特征向量。

根据渐进统计学理论, 当  $p$  固定且  $n$  趋于无穷大时, 由中心极限定理和斯卢茨基定理。

$$(X - \bar{Y})^{-T} \left( \frac{S_1}{n_1} + \frac{S_2}{n_2} \right)^{-1} (X - \bar{Y}) \rightarrow \chi^2(p) \quad (17)$$

其中  $\bar{X}$  和  $\bar{Y}$  分别表示样本平均值,  $S_1$  和  $S_2$  分别表示样本协方差矩阵。

由于音乐特征的数量 ( $p=9$ ) 远远小于样本量, 我们不需要对数据做任何分布假设。 $\text{Chi-squared}$  统计量遵循自由度为9的  $\text{Chi-squared}$  分布。通过计算统计量, 我们可以根据渐近分布得到  $p$  值。

如果接受无效假设, 我们得出结论, 追随者的所有音乐特征与他们的影响者同样相似, 表明没有任何特征具有显著的 "传染性"。

另一方面，如果拒绝了无效假设，我们可以进一步应用BH多重假设检验，以进一步找出哪个特征更具 "传染性"。

4.4.3 解决方案

在不失一般性的情况下，我们把重点放在最受欢迎的乐队--披头士乐队和他们的追随者身上。将相似性分数阈值设定为0.21，披头士乐队和他们29个最受欢迎的追随者的相似性贝叶斯网络如图10所示。

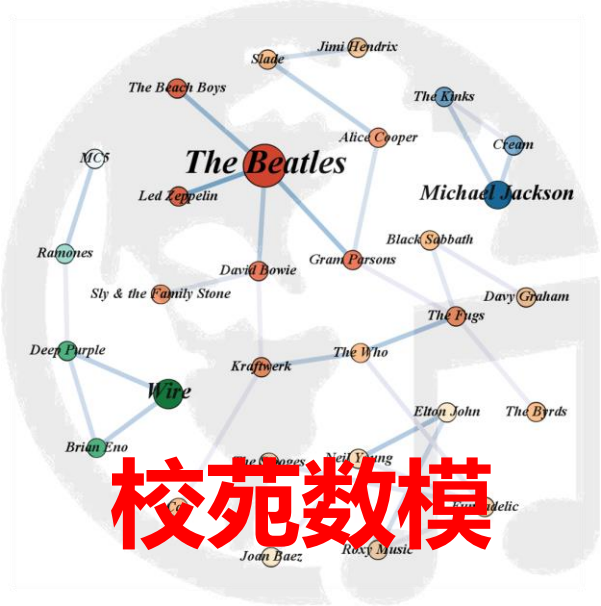


图10：披头士乐队及其追随者的相似性贝叶斯网络。

如图10所示，直接或间接地，72%的所谓 "追随者 "与他们的 "影响者 "披头士乐队有相似的作曲风格。然而，28%声称受到影响的艺术家与他们的 "影响者 "有不明显的相似性（低于阈值）。换句话说，根据相似性数据，近十分之三的艺术家并没有受到他们所谓的 "影响者 "的影响。

表4：传染性特征测试

追随者姓名	卡方值	P值
齐柏林飞船	3.183	0.957
Gram Parsons	5.971	0.743
大卫-鲍伊	7.232	0.613
海滩男孩	14.284	0.113

表4显示了传染性特征检验的结果。在5%的显著水平下，四个无效假设都被接受，这表明披头士乐队的音乐特征对其追随者的影响是相同的，换句话说，没有任何特征比其他特征更具有传染性和突出性。

我们应该强调的是，在上述过程中，由于多重假设检验，第一类错误可能会累积。然而，在这个问题上，所有的假设都被接受。因此，第一类错误仍然可以控制在显著性水平之内。

## 4.5 任务5

### 4.5.1 革命的定义

为了处理任务5，我们首先定义“音乐革命”。主要流派的影响者数量如图11所示。随着时间的推移，一些流派蓬勃发展，另一些则失宠。

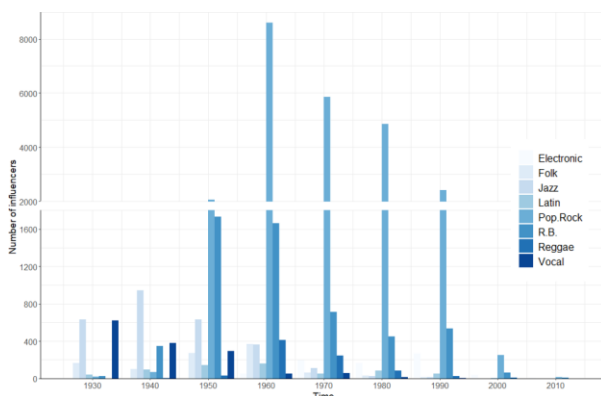


图11：主要流派的影响者数量

我们把流派的革命时间定义为一些流派衰落而另一些流派兴起或占据主导地位的时间。从图11中，我们发现在20世纪50年代，电子乐逐渐繁荣，而声乐则从繁荣走向衰落。流行/摇滚和R&B的流行也遇到了急剧上升。这些现象表明，20世纪50年代是革命时期之一。

在定义了革命时期之后，我们着重于寻找20世纪50年代音乐特征的一个重要变化点，以解释我们上面讨论的内容。这个问题的关键是转向识别音乐特征的时间序列中的变化点。根据我们找到的变化点，我们就可以确定代表革命者的艺术家。

### 4.5.2

我们将变化点的集合定义为  $\Theta = \{\theta_i\}_{i=1}^m$ ，让一些音乐的时间序列字符  $\{Y\}_{t=1}^n$  承认。

$$Y_t = a_i + b_i \frac{*t}{n} + \varepsilon \theta_i \quad i=1 \rightarrow m, 1 \leq t \leq \theta_i \quad (18)$$

其中， $a_i$  是截距项， $b_i$  衡量线性趋势的速度，而误差项  $\varepsilon_t$  是白噪声过程。

为了确保参数和变化点的稳定性，我们还需要限制

$$\theta_i - \theta_{i-1} > \tau \quad (19)$$

我们建议  $\tau = [0.1n]$ ，这将在敏感性分析中详细说明。

由于  $y_t$  在不同阶段的变化具有线性趋势，当  $\varepsilon_i$  遵循独立和相同的分布时，显然可以通过最小化残差平方之和来估计参数。

$$\{\theta_i\}_{i=1}^m = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^n (y_t - y_{\theta_i})^2$$

当变化点的数量给定时, 在这个问题上使用动态编程算法的关键是建立新变化点前后的残差平方之和的递归关系。

让  $\delta(m, n)$  表示与使用前  $n$  个观测值的包含  $m$  个断点的最佳分区相关的平方残差总和,  $RSS(i, j)$  表示对从  $i$  到  $j$  的一段开始应用最小二乘法得到的平方残差总和。

然后, 我们可以通过解决以下递归问题来实现整体残差平方和的全局最小化。

$$\delta(m, n) = \min_{m\tau \leq j \leq n-\tau} \delta(m-1, j) + RSS(j+1, n) \quad (21)$$

其中,  $\tau$  是一个修剪参数, 约束两个变化点之间的距离不能太近。

值得注意的是, DP 算法是  $O(T^2)$ , 不依赖于变化点的数量。因此, 我们可以快速地估计所有的变化点, 这些变化点与真实的变化点是一致的[5]。

很明显, 我们也需要对变化点的数量进行惩罚。考虑到变化点的数量是一个调整参数, 我们建议使用 BIC 准则来确定最佳变化点的数量  $m^*$ 。

$$m^* = \underset{m}{\operatorname{argmin}} RSS_{\text{overall}} + \log(n) * \sigma^2 * m \quad (22)$$

#### 4.5.3 解决方案

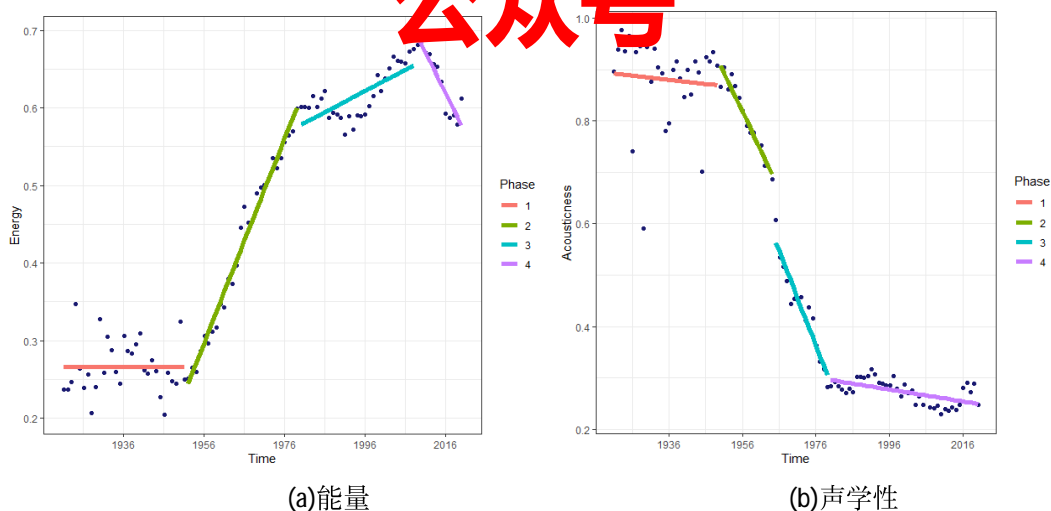


图12：声学性和能量的变化点检测

通过使用上述 DP 算法, 我们检测了数据集 "data\_by\_year.csv" 中 11 个连续音乐特征的变化点, 如表 5 所示。

从表 5 中可以看出, 四个音乐特征 (声学性、能量、舞蹈性和响度) 在 1950 年前后显示出明显的变化。

表5：音乐特征的革命日期

特征	革命日期							
	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
工具性	1933	1946	-	1964	-	-	-	-
持续时间_ms	-	1946	-	1966	-	-	-	2007
声学性	-	-	1950	1964	1979	-	-	-
节奏	-	1947	-	-	1979	-	1996	2008
舞蹈性	-	-	1950	-	-	-	1997	2008
缬氨酸	-	1947	-	1966	-	-	-	2005
能源	-	-	1951	-	1979	-	-	2008
灵活性	-	-	1956	-	1976	-	-	2008
经验之谈	-	-	1956	-	-	-	-	2006
知名度	-	-	1953	-	1970	-	-	2006
响度	1936	-	1950	-	-	-	-	2008

为了进一步了解与不同流派的音乐特征有关的革命，我们在图13中分别画出了能量和声学的图表。很明显，新流派（流行/摇滚、R&B和电子）有明显的高能量，但它们的声学性却比旧流派低得多。这与上面的时间序列分析是一致的。因此，我们相信，四个音乐特征：声学性、能量、舞蹈性和响度可能标志着音乐演变中的革命。

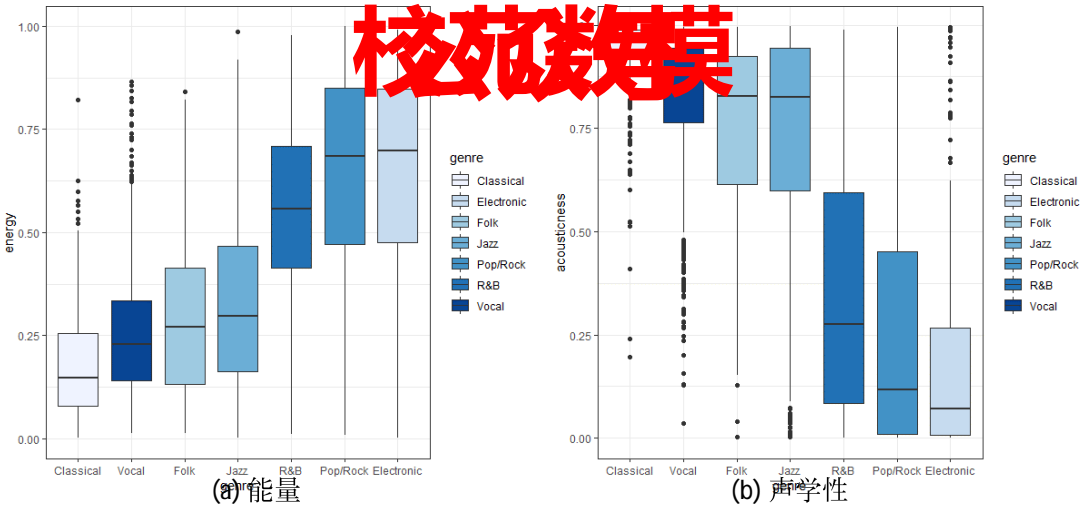
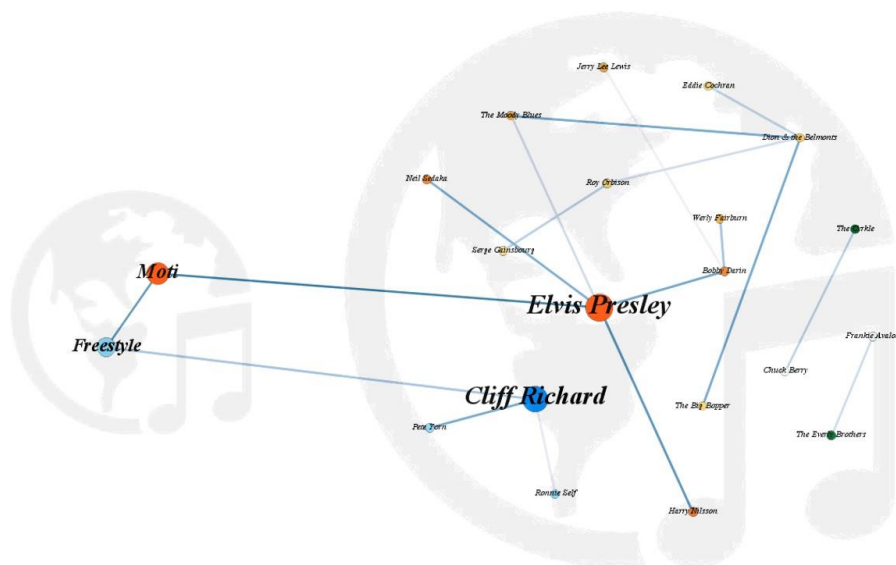


图13：不同流派的能量和声学性

我们定义了两类类型的革命者：对现有流派做出重大改变的艺术家和创造新流派的艺术家。对于第一种类型，我们把重点放在流行/摇滚乐上，它在20世纪50年代出现了巨大的变化。对于第二种类型，我们把重点放在电子乐上，它在20世纪50年代初首次出现。

为了在我们的网络中找出革命者，我们首先过滤了20世纪50年代发行的流行/摇滚和电子曲目，然后根据这些曲目的艺术家建立一个相似性贝叶斯网络。直观地说，处于我们网络核心的艺术家是革命者，因为他们对这一时期的艺术家有重大影响。





Electronic

Pop/Rock

图14：20世纪50年代流行/摇滚和电子的相似度贝叶斯网络

我们的相似性贝叶斯网络如图14所示。猫王，一个在20世纪50年代拥有最多追随者的流行/摇滚乐影响者，在美国领导了一场反文化运动。同时，英国艺术家克里夫-理查德（Cliff Richard）对这一流派的演变做出了巨大的改变。这两位革命者不仅重塑了现有的流派，而且还后发了像莫蒂和自由式这样的艺术家，创造了一个新的流派--电子乐。因此，猫王和克里夫-理查德代表了50年代音乐革命的革命者。

## 4.6 任务6

### 4.6.1

在这项任务中，我们选择流行摇滚作为研究的体裁，因为它存在了很长时间，而且有若干次的再创造。

直观地说，动态影响者在特定时期内极大地改变了该流派的发展趋势。因此，一个时期的动态影响者应该满足

- (a) 他/她在这一时期发布了10多首歌曲。
- (b) 这些曲目的音乐特点与流行/摇滚的滞后趋势一致。

根据上述定义，我们构建一个动态影响者的指标为

$$d_i = \frac{1}{\sum_{j \in F} |A_i|_{t \in A_i}} \sum_{j \in F} \left( x_{i,j,t} - \bar{x}_{j,t-u} \right)^2 \quad (23)$$

其中， $x_{i,j,t}$  是第*i*个艺术家在*t*年发布的曲目的音乐特征*j*， $\bar{x}_{j,t-u}$  是艺术家的平均值，*F*是音乐特征集合，*u*是滞后年份。这里我们设定*u*=1，这意味着动态影响者将在曲目发布一年后引领音乐潮流。

## 4.6.2 解决方案

图15显示了流行/摇滚音乐特征的变化。初步数据分析表明，大多数流行/摇滚歌曲是在1956年以后发行的（虚线）。在1956年之前，音乐特征的波动是由于样本量小造成的。因此，我们只分析了1956年以后的变化。我们设定 $F=\{\text{声学性、舞蹈性、能量、响度、价位}\}$ ，因为只有这些特征在1956年后显示出明显的波动。

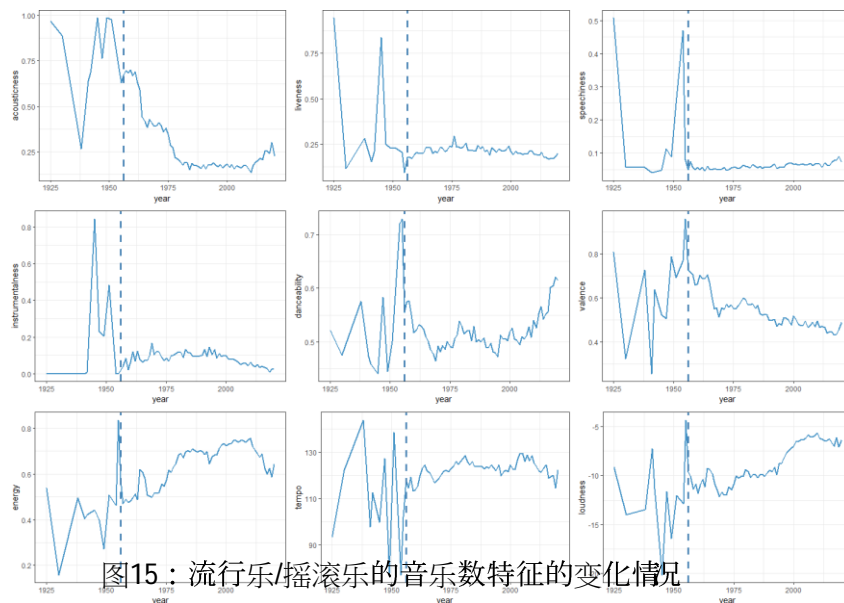


图15：流行乐/摇滚乐的音乐数特征的变化情况

通过计算99位最著名艺术家的动态影响者指标值，我们找出了对流行/摇滚乐有巨大影响的前10位动态影响者。影响者和他/她的活跃期如图16所示。



图16：流行/摇滚音乐特征的变化

根据图15，流行/摇滚的一些音乐特征随着时间的推移发生了很大的变化。1956年后，声学性和价位都经历了明显的下降，而能量和响度则显示出持续的增长。

上述变化主要是由动态影响者的变化引起的。随着技术的提升和电子放大，甲壳虫乐队、齐柏林飞船和老鹰乐队在上个世纪的流行摇滚音乐中倾注了更强烈的情感和激情。这些艺术家倾向于表达悲伤或抑郁的情绪。然而，近20年来，一些伟大的音乐家，如林肯-帕克和泰勒-斯威夫特，将流行摇滚乐的基调从悲伤转向快乐。他们使流行音乐变得流畅、放松，更适合于跳舞。

## 4.7 任务7

### 4.7.1 音乐的文化影响

根据上一节的模型，我们发现了音乐演变的三个重要时期，如图17所示。我们对文化影响的过程讨论如下。

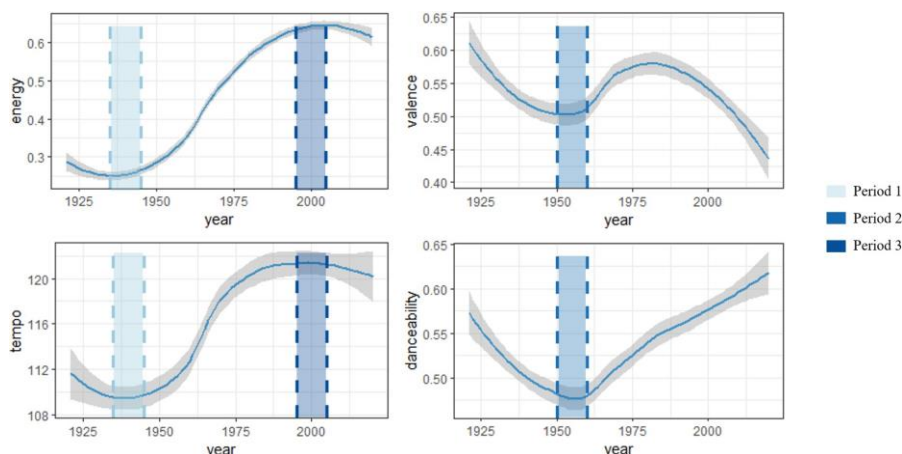


图17：文化影响的过程

**第一时期（1935-1945）。**这一时期是由大萧条时期的情绪塑造的。在这一时期，许多艺术家开始创作高能量和快节奏的音乐，如图17所示。流行/摇滚乐的重塑和快速发展，以及雷鬼音乐的华丽开端，形成了一种向上的文化面貌。

**第二时期（1950-1960）。**在二战的不利影响之后，人们即将踏上音乐之旅，并大幅改变音乐风格、价值和舞蹈-能力，看到这一时期发布的音乐急剧增加--流行/摇滚的黄金时代。整个文化变得充满希望和活力。

**第三时期（1995-2005）。**到1995年底，许多年轻人对充斥在电波中的流行音乐/摇滚乐感到厌倦，音乐的能量和节奏达到了顶峰，开始下降。一个成熟的音乐文化逐渐形成。

### 4.7.2 网络内确定的变化

**社会和政治变化。**根据1950年代流行/摇滚艺术家的相似度贝叶斯网络。如图18所示，20世纪50年代的流行/摇滚乐在几个子类型中是多样化的，这些子类型彼此之间相当不同（因为不同子类型的艺术家之间没有边缘）。网络的这一特征表明了1950年代的反文化运动--人们越来越渴望真正的表达自由，人们的多样性变得越来越突出。

**技术变化。**根据维基百科，我们知道互联网的出现主要导致了电子音乐的兴起，因为互联网使制作电子音乐的软件更容易传播。这种趋势可以通过电子乐的影响者人数随时间的变化来确定，如图19所示。由于互联网的普及，影响达到了顶峰。

上述分析表明，基于所建立的模型，我们可以找到音乐在时间和环境中的文化影响，并确定社会、政治或技术变化的影响。

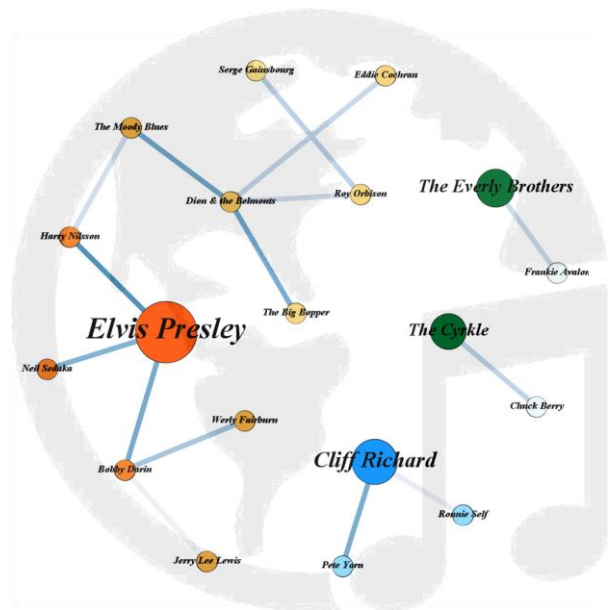


图18：20世纪50年代流行/摇滚乐的熟悉度贝叶斯网络

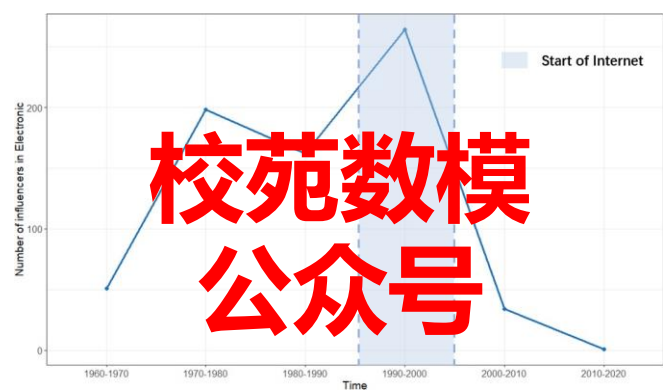


图19：电子行业的影响者数量

5 敏感度分析

在任务5中，我们的变化点检测模型涉及一个修剪参数 $\tau$ ，这可能对最佳变化点检测有影响。为了测试模型的稳健性，我们对修剪参数进行了改变，并检查1950年代的革命是否能被很好地检测出来。

表6： $\tau$ 对变化点检测的影响

特征	$\tau$			
	[0.05n]	[0.10n]	[0.15n]	[0.20n]
声学性	1950 1964 1979	1950 1964 1979	1950 1964 1979	1950 1964 1979
舞蹈性	1928 1951 1997 2008	1950 1997 2008	1950 1997 2008	1950 2008
能源	1951 1983 1994 2008	1951 1979 2008	1951 1979 2008	1951 1979 2008
响度	1936 1951 1980 2008	1936 1950 2008	1935 1950 2008	1950 2008

如表6所示，当 $\tau$ 大于或等于[0.1n]时，提议的模型对变化点非常稳健，这意味着我们选择的值是合理的。

## 6 优势和劣势

### 6.1 优势

- **有效的模型。**对于不同的问题，我们建立了几个模型，包括贝叶斯网络、变化点检测模型，这使得每个问题的分析都去掉了尾巴，令人信服。
- **统计学测试。**我们不是简单地比较数值，而是应用曼-惠特尼检验和多元平均检验，以使我们的结论是可靠的。
- **生动的视觉化。**我们用许多数字来显示我们的结果，使之更容易捕捉到关键信息。

### 6.2 弱点

- 我们的分析并没有涵盖一些模型中的所有体裁。鉴于篇幅有限，我们的部分分析主要集中在两个具有代表性的流派上。流行摇滚和R&B。
- 我们对时间序列做了一个强有力的假设，这在现实中并不总是成立的。我们假设误差项遵循独立的正态分布，然而，在现实中，它往往遵循一个弱依赖的静止过程。

# 校苑数模 公众号

## 参考文献

- [1] 陈冠荣.复杂网络概论：模型、结构与动力学[M].高等教育出版社, 2012.
- [2] Lawley, D. N. "A generalization of Fisher's z test."。Biometrika 30.1/2 (1938):180-187.
- [3] Rencher A C , Christensen W F .多变量分析的方法[J].Technometrics, 2002, 38(443):76-77.
- [4] McKnight, Patrick E., and Julius Najab."Mann-Whitney U测试"。科西尼心理学百科全书》（2010）。1-1.
- [5] Bai J , Perron P .多重结构变化模型的计算和分析[J].Jour- nal of Applied Econometrics, 2003.



## 给ICM协会的一份文件

致综合性的集体音乐协会

来自。2100112团队

在这份文件中，我们将向你展示我们的网络是如何对音乐的影响提供更好的理解，以及在更丰富的数据下可能的改进。此外，我们将给你一些关于进一步研究音乐及其文化影响的建议。

我们的方法对艺术家的影响力有一个全面的衡量。我们提出了基于网络理论的计量方法--度中心、加权重中心和特征中心。音乐影响力的衡量标准同时考虑了追随者的数量和受欢迎程度，使其更接近于现实。

基于贝叶斯网络，我们的方法衡量影响者和追随者之间的真实联系。尽管一些追随者声称他们受到影响者的影响，但他们的曲目并没有显示出与影响者的任何明显的相似性。通过应用我们的贝叶斯网络模型，你可以找出“真正的”追随者，并对艺术家之间的音乐影响网络有一个更客观的了解。

受益于详细的时间序列分析，我们的方法提供了对流派和音乐革命的兴衰的进一步洞察力。我们提出了一种动态编程算法来检测音乐特征和革命的变化点。与动态影响者指标相结合，你将学到很多关于音乐演变过程的知识。

更重要的是，如果引入更多的数据，我们的网络可以得到很大的改善。由于贝叶斯网络强大的学习能力，如果引入更多的数据（如果有更多流派或更多音乐特征的数据），我们可以得到更精确的参数估计，这将带来更好的结果。

他们有很多关于音乐的话题值得进一步调查。例如，如果我们发现某个国家的音乐越来越有活力，那么我们可以去看看是否有社会、政治或技术的变化，比如反文化运动和互联网的普及。当音乐特征发生变化时，社会也会发生变化。例如，积极的音乐风格的传播会形成一种文化的向上看。

希望你会发现我们的论文有帮助，并考虑我们的建议。我们期待着贡献我们的努力，以获得对音乐演变的更深入的了解。