四、影响力网络模型

4.1 构建音乐家影响网络

？data\_influence 数据集有无数据问题来着？

考虑到影响者对追随者的影响关系是有向的，可以用数据结构中的邻接矩阵[1]来描述。要生成用于网络分析的邻接矩阵，必须定义表征连接的条件。在influence\_data数据集中，一共有5603个音乐家被分配了影响者和追随者的角色。音乐家作为网络节点，我们将定义为影响者对追随者有影响关系，音乐家之间的邻接矩阵可以表示为：



随后，我们将数据导入，对数据进行可视化化操作。由于节点数量和连边数量过大，节点杂糅在一起，可视化效果差。多维缩放（MDS）可以将节点之间的差异转化为可用于在低维空间中可视化对象的坐标（MDS converts a set of pairwise dissimilarities among objects into coordinates that can be used to visualize the objects in a low-dimensional space）[这是翻译]，降维后的样本为中心化样本，因为只保留了距离信息[zh1 13]，方便将其节点嵌入到二维中进行可视化。因此，我们借助Gephi利用MDS布局方式执行图形可视化，由此生成音乐家影响网络。

图 ：MDS layout of the musician influence network, having 5603 nodes and 42770 edges. 754032 has the highest degree, far left, is far away from other points.

音乐家影响网络中，我们用代表各音乐家，我们设计了包含四种颜色和五种形状的20种节点，用以区分音乐家所属的不同派系。若影响者对追随者产生影响，则从指向生成有向边。

4.2 音乐家个人音乐影响力

为了确定音乐家影响力的影响指标，我们将数据导入UCINET针对音乐家影响网络进行个体网分析，其中具有最显著差异的指标为网络中节点的度。

用音乐家的出度，表征音乐家对其他音乐家的影响大小，



用代表音乐家的入度，代表了音乐家创作时吸收其他音乐家长处的多少。



若音乐家参考太多而影响相对较小，可以认为该音乐家的音乐原创性较弱，从实际出发，此类艺术家一般不会产生深远的影响；而上世纪的音乐大师，存在只有出度没有入度的情况，是因为他们通过自己的不断研究留下了影响深远的音乐，此类音乐家是具有高影响力的。基于对网络性质和节点度的研究，我们定义了音乐家个人音乐影响力。



4.3 音乐影响力研究的子网络

基于对音乐家个体的研究，我们对音乐家网络中从属于同一派系的音乐家进行聚类，由此形成了二十个不同派系的子网络，子网络内部相互连通，不同子网络之间也是联通的。

图 音乐家影响网络的子网络

在派系子网络的基础上，我们将音乐家影响网络演化为派系影响网络。音乐影响力的定义以每个音乐派系为主体，即计算各派系音乐影响力的大小。节点为各音乐派系，用表示，若派系中有音乐家对派系存在影响关系，则生成指向的有向边，边的权重为派系中影响派系追随者的全部影响者数量。通过gephi对各派系子网络的所有节点、自环、指向各派系的出边以及各派系指入的入边分别进行融合。其中子网络中音乐家的数量决定了该节点的大小，各节点出边和入边的权重分别决定了其出边的粗细，从而生成了派系音乐影响网络。

4.4 派系音乐影响力

由于派系影响网络是由音乐家影响网络演化而来，节点为相同派系音乐家节点融合而成的，所以派系影响网络中度数仍是影响力的关键指标。我们仍然延用出度与入度之比表征音乐家的原创传播力，那么派系的音乐原创传播力可以用派系中所有音乐家的该性质加总来计算，由此我们定义了个人原创传播负指标：



在度数分析的基础上，根据**距离**或**年份**（←一定要加粗）对派系影响网络进行聚类分析，我们可以发现各派系之间存在相互影响较为紧密的团体，而联系团体内部或联系其他团体的派系，起到了中介作用，对网络的稳定性有着决定性作用，根据中间人分析的相关知识：

扮演中间人角色的派系，在网络中有着重要的位置，往往能决定网络的稳定。因此，利用中介性作为派系影响力大小指标的一部分是合理的。由于派系音乐影响网络属于边加权有向图，对其中介中心度的求解一般有两种方法：一是利用度数中心度代表，再就是利用幂律分布的性质对其近似求解。但这两种方法可操作性较差，于是我们在中间人分析的启发下，利用权重矩阵和邻接矩阵求相对出权重和相对入权重，从而定义了中介度。



利用节点中最大中介度对所有节点的中介度进行标准化，就可以得到中介传播力



综合个人原创传播负指标和中介传播力两方面，加入传播与影响的相关因子，最终定义派系的音乐影响力



五、相似度度量模型

5.1 音乐网络的构建

通过建立音乐家影响网络并将其演化为派系影响网络，对音乐家之间的影响做了研究分析。为了深入探索音乐的特点，可以利用full\_music数据集建立歌曲散点图（Song Graph），将音乐特征相似的歌曲连线，即可生成歌曲相似网络。考虑到歌曲和歌手是一一对应关系，我们换种思路，利用歌曲与音乐家的对应关系将歌曲散点图和音乐家影响网络叠加，从而生成一种综合型网络——音乐影响网络。在音乐影响网络中，分析歌曲之间的相似性时，除了歌曲本身的音乐特点，还有该歌曲的歌手之间的关系影响，这样一来，我们建立的相似度度量方法更贴近现实，也使得对音乐特点的分析更全面、透彻。

5.2 音乐相似度度量方法

音乐的演绎形式有很多，比如：唱歌、乐器以及电脑合成等，但是几乎全部的音乐都能找到与其相似的音乐，并归为已知的几个类别。将相似的音乐按照年代对比分析，就可以发现音乐特征随年代的演变过程，因此音乐相似度度量对我们研究音乐的发展历史非常重要。

为建立一套全面的、可靠的音乐相似度度量方法，我们结合full\_music\_data数据集给出的信息，从音乐特点、人声特点和歌曲描述三个方面量化音乐的相似度。考虑到参照的指标过多会影响相似度度量的精度，因此我们首先对数据中给出的17种指标进行机理性分析，进行第一波指标剔除。

经过查阅文献和机理性分析，我们分析歌曲描述中year、release\_date和song\_title 三个指标以及人声特点中explicit对于音乐相似的解释力度不足，所以剔除以上四个指标。随后，为检测指标之间的相关程度，对剩余的13个指标求解相关系数并展开分析。

（上表用latex打出来，看看效果，不行的话直接截图）

表中可以看出，loudness和energy之间的相关系数最大，且只有它超过了0.5，所以我们认为loudness和energy对于相似度的贡献是重复的。从现实意义来讲，力量和响度从音乐本质上就是相通的，因此我们将loudness剔除，留下具有精神和响度两方面属性的energy。因此，我们选取了12个指标，下面将所有指标按照所属大类列出：

1.音乐特点相似度

根据音乐常识可知，一首音乐的乐调是音乐辨识度最大的指标，mode和key值就是乐调的数据，对于mode值，大调和小调可以直接决定两首歌曲的相似与否，而对于key值，两首歌曲的变化范围超过1，则就超出了转调的范围，因此定义来表示key值对歌曲相似度的决定性的影响。

，

利用、、、分别代表剩余的音乐指标danceability、energy、valence、tempo，他们对于音乐的辨识度不像mode和key值极端，所以分别给出各自的权重来表示各指标对音乐相似度的贡献。因此可以得出音乐特点相似度：



分别为0.39193、0.08823、0.37313、0.14671（熵权法确定）

2.人声特点相似度

在人声特点中，像key或mode这种决定性的参考指标是speechiness，它代表了歌曲音乐是否是演讲形式演绎的，The more exclusively speechlike the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. speechiness对相似度的贡献值可以表示为：



利用、、分别代表剩余的音乐指标acousticness、instrumentalness、liveness，这三项指标都是音乐音轨方面的分析，各指标之差对音乐相似度的贡献我们认为是相同的。为了得到标准化的数据以及更直观的表达，我们采用余弦距离来表征指标之间的差值，因此可以得出人声特点相似度：



3.歌曲描述相似度

歌曲描述方面的指标对于相似度的贡献并不是很大，因此该方面也是剔除指标最多的。对歌曲秒是方面，我们使用数据之差，再进行标准化，得到了歌曲描述相似度：



4.相似度度量公式

综上，我们得到了音乐特点相似度、人声特点相似度和歌曲描述相似度，为三个指标分别制定贡献系数，即可得到音乐相似度度量方法：



5.利用熵权法确定未知权重

在相似度度量公式中，我们在公式中假设了系数，求解相似度前必须给出可靠的系数。一般给定系数的方法是层次分析法等采用专家打分来定夺权重大小，但是此类方法主观性太大。考虑到相似度度量方法要服务于之后的音乐发展历史探索，我们选择了通过数据特征客观给出权重的熵权法。

根据信息熵的定义，对于某项指标，可以用熵值来判断某个指标的离散程度，其信息熵值越小，指标的离散程度越大， 该指标对综合评价的影响（即权重）就越大。因此，我们利用熵权法对未知系数进行求解。

以音乐指标danceability、energy、valence、tempo在中的贡献系数的求解为例，我们对熵权法的过程进行展示。对各个指标的数据进行标准化处理。4个指标，其中，假设对各个指标数据标准化后的值为。



接下来需要求解各指标的信息熵根据信息论中信息熵的定义，一组数据的信息熵：



其中，。如果，则定义

有了各指标的信息熵，就可以确定各指标的权重。根据信息熵的计算公式，计算出各个指标的信息熵为，通过信息熵计算各指标的权重：



通过熵权法计算，我们确定了未知权重：：：的系数分别为0.24846、0.68001、0.07153。音乐指标danceability、energy、valence、tempo在中的贡献系数分别为0.39193、0.08823、0.37313、0.14671。

5.3 Similarities within Genres

利用熵权法求解出未知系数后，我们打开音乐世界的钥匙终于激活了。由于社会的进步，事物变化的速度越来越快，音乐也是如此。在分析派系影响力网络时，我们在POP派系中找到了几首听过的歌，他们有复古爵士味的音乐也有现代感十足的音乐，所以我们对派系的划分产生了怀疑。带着这个疑问，我们利用创造的相似度度量方法对每个派系内的两两音乐家之间均求解相似性，以分析派系内音乐家相似性的大小。

由于要展示的数据量大，且相似度结果为0到1之间的数，直接展示结果不直观，起不到很好的展示效果。因此我们将0到1之间的数表示成蓝色到黄色的颜色变化，将相似度矩阵转换为色阶图的形式展现出来。各派系内音乐家之间相似度色阶图如下：

从色阶图中可以非常直观地观察出各个派系相似度关系。色阶图整体越黄，表示派系内音乐家之间相似度越高，如New Age和Classical两个派系的色谱图，图中大面积呈现黄色，只有少部分的蓝色横线或竖线。这些少数的蓝线可能代表了该派系中另类的音乐家，从相似度度量公式的构成角度来看，这些另类的音乐家可能对一个乐调或演绎方式情有独钟。但是色谱图整体呈蓝色的派系数目较多，如Pop/Rock和Unknown两个派系的色谱图，图中大面积或全呈现蓝色，Pop/Rock只有少数的黄色斑点。Unknown派系收录的是没有明确分类或者分不了类的歌曲，他们非常有特点，所以相互之间的相似性基本为0。而Pop/Pock是个随着信息时代新兴起的年轻派系音乐，音乐家风格各异，派系里的歌也是如此，因此相似度非常低。

根据结果可以看出，派系之间的相似度较低，结果符合现实生活中的道理：同属于一个派系的音乐并非千篇一律。而且各派系的相似度差别较大，既有音乐风格变化较小的派系，也有音乐几乎都不相似的派系。

5.4 Similarities between Genres

研究完派系内音乐家的相似性，我们再来研究一下不同派系之间音乐家之间的相似性，顺便也验证一下我们相似度度量方法的正确性。由于派系种类多，各派系含有的音乐家数目众多，所以一个个地求解不现实，而且数据分析也会非常麻烦。因此，我们设计了一个计算机模拟试验，用来比较派系内和不同派系间音乐家相似性。以下为模拟过程的算法流程：

我们利用计算机随机进行了500次试验，每次试验中若派系内音乐家相似度大于派系间音乐家相似度，记为试验成功，返回值为1；试验失败，返回值为0。我们将500次试验结果以面积图的形式进行展示：

从图中可以看出蓝色区域比红色区域线条更宽、更密集，说明试验成功的次数更多。根据数据显示，500次试验中有318次成功试验。我们随后又分别进行了5000次和10000次的试验，成功比例均在0.6左右。所以，在排除偶然性外，我们得证派系内的音乐家比派系间的音乐家更相似。

5.5 派系随时间变化的研究

为研究派系随时间的变化，从data\_by\_year数据集入手，我们计算了年与年之间的相似度

5.6 音乐家影响网络的相似度研究

通过相似度度量，我们对派系内和派系间的音乐家进行了分析。将视角转变到另一个已知数据集中，我们利用相似度度量的方法对音乐家之间的影响关系进行研究，以期得出音乐家之间影响关系的实际作用。

我们验证影响者和追随者的音乐特征是否相似，若不相似的话，则说明影响者实际上对相应追随者的影响作用不大。由于数据集庞大，逐个计算的过程过于繁琐，我们设计了计算机模拟试验，从data\_influence中随机选取一组相互影响的音乐家，分别求解每组影响者和追随者之间的相似性。试验循环500次，统计每次试验的相似度，以此来揭示影响者实际上对追随者的影响作用大还是不大。计算机模拟程序的算法步骤如下：

通过计算机模拟试验，我们得出了500次试验的相似性数据。

从图中可以看出，500次试验得出的相似性在0.4周围波动，统计相似度超过0.5的试验次数，一共有81次。因此，试验结果可以证明影响者与追随者的相似度不高，可以进一步说明影响者对追随者的实际影响作用并不大。