Pylouvain算法 <https://www.cnblogs.com/LittleHann/p/9078909.html>

算法简介：

### **1）初始化：**

将图中的每个节点看成一个独立的社区，社区的数目与节点个数相同；

### **2）开始first phase迭代 - 社区间节点转移：**

对每个节点i，依次尝试把节点 i 分配到其每个邻居节点所在的社区，计算分配前与分配后的模块度变化ΔQ，并记录ΔQ最大的那个邻居节点，如果maxΔQ>0，则把节点 i 分配ΔQ最大的那个邻居节点所在的社区，否则保持不变；

### **3）重复2）- 继续进行社区间节点转移评估：**

直到所有节点的所属社区不再变化，即社区间的节点转移结束，可以理解为本轮迭代的 Local Maximization 已达到；

### **4）second phase - Rebuilding Graph：**

因为在这轮的first phase中，社区 C 中新增了一个新的节点 i，而 i 所在的旧的社区少了一个节点，因此需要对整个图进行一个rebuild。

对图进行重构，将所有在同一个社区的节点重构成一个新社区，社区内节点之间的边的权重更新为新节点的环的权重，社区间的边权重更新为新节点间的边权重；

### **5）重复2）- 继续开始下一轮的first/second phase：**

直到整个图的模块度不再发生变化。

探索内容

通过pylouvain算法对现有社区进一步进行社区划分，然后对每一个社区进行一些指标的统计，来探索一些欺诈情况

统计指标：总体特征：团体节点数量，团体放款金额，团男女比例，最大年龄差

平均特征：平均度，平均年龄，平均的申请，通过，逾期次数，放款金额，最大逾期天数

比率特征：通过率，逾期率。

分析过程：

1. 从网络选择社区
2. 完成社团检测，社团特征转入pandas
3. 对特征数据进行分析
4. 过滤出高风险团体并验证
5. 对数据进行聚类分析，对分类结果进行验证

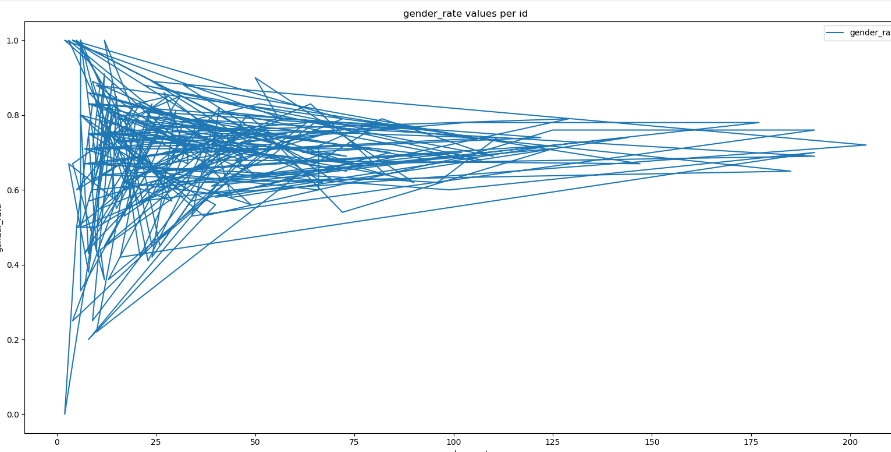
特征： 'node\_count', 'gender\_rate','max\_age\_diff',

'avg\_age','avg\_apply','avg\_approve','avg\_overdue',

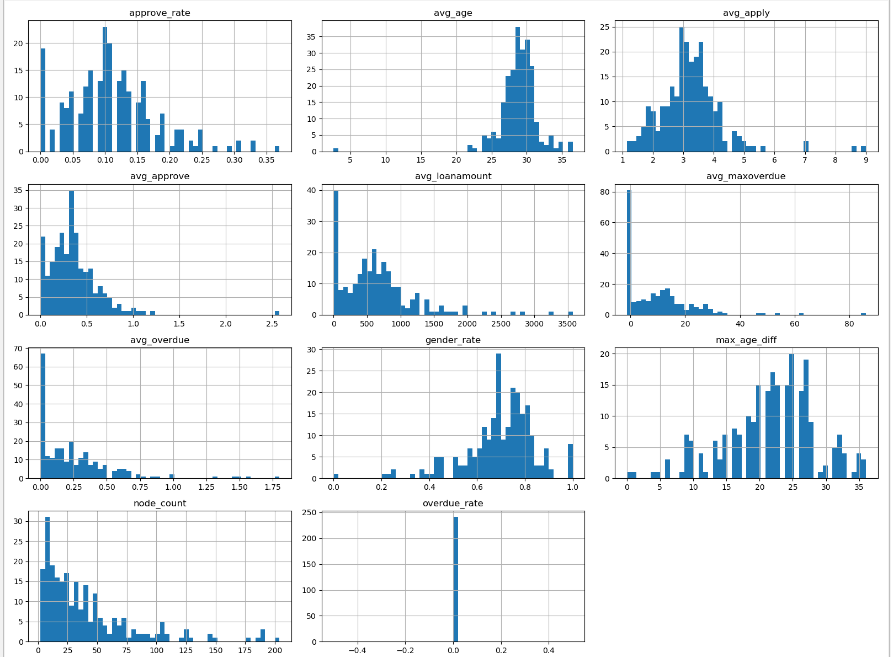
'avg\_loanamount','avg\_maxoverdue','approve\_rate','overdue\_rate'，'avg\_M3','avg\_pd10'

comm = 3229132

Comm = 9375315 9833 nodes, 58415 edges 模块度： 0.9443175736668015 团： 242

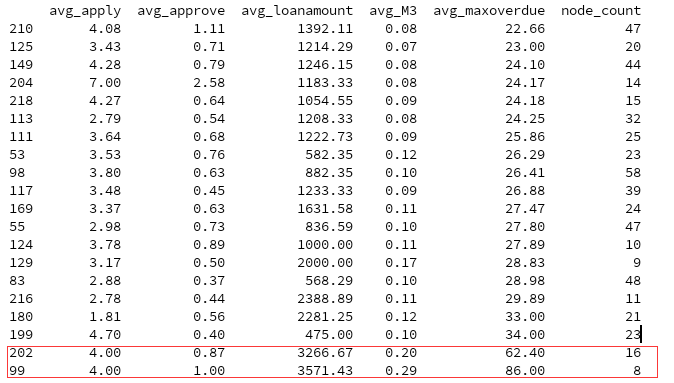


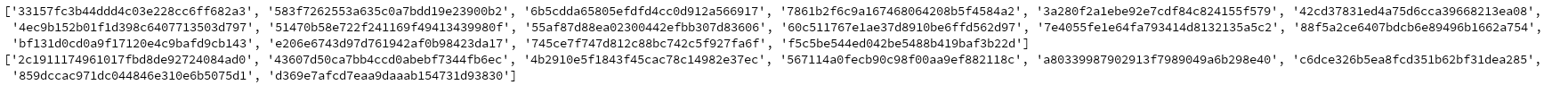
男性比与团数量之间的关系：团男性平均占比大概在0.7左右



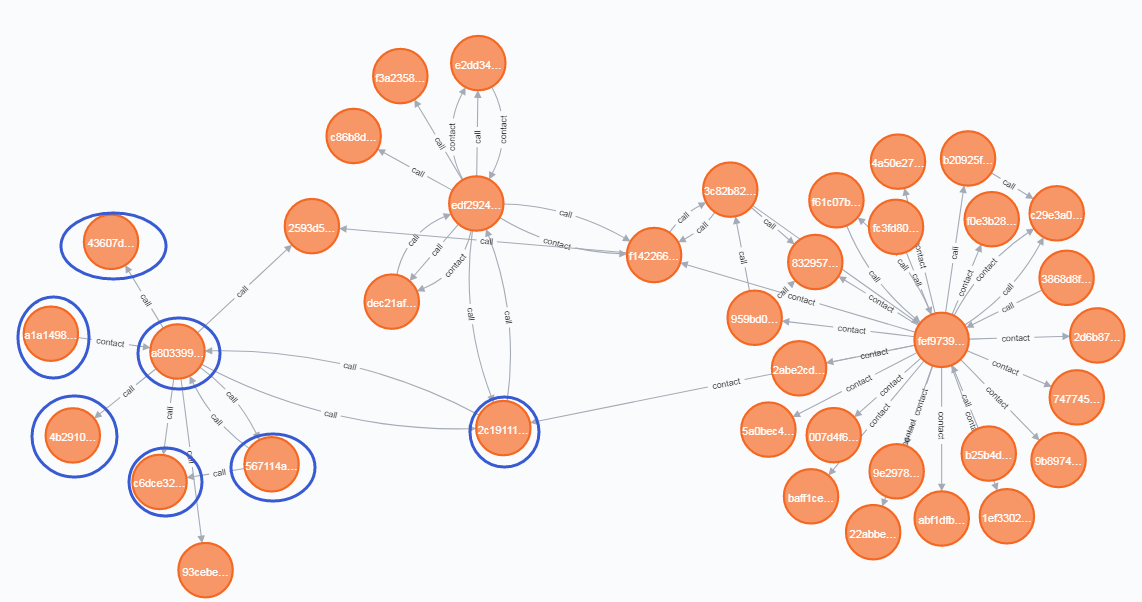
各变量的分布情况，明显的平均年龄，申请次数呈正态分布，其他变量大部分都有点偏离正态分布

1. 过滤高风险团体，找到通过率和最大逾期天数都高于平均水平的数据





接下来定位这两个团体的具体节点，然后去网络里面看一下



上图是99号团体，共8人，有3人逾期M3，这样一个小的社交团体欺诈风险已经非常之高了，

1. 利用无监督算法对团体进行聚类

1）降维是从高维到低维度，到二维，如果数据的特征没有区分度，没有相关性，那么所有的点应该都聚集在一个小范围内，异常点自然是会脱离于聚集区存在的。

如果所有特征之间存在线性相关性，映射到二维平面整体分布呈单调趋势，如果只在某一些特征上有相关性，那么至少会在二维平面的某一维度呈现单调

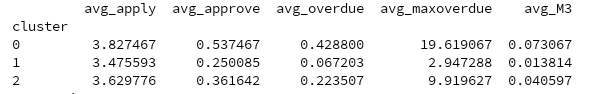
2）聚类的拟合曲线的陡峭程度也有非常好的参考价值，变化率越高说明团体之间的差异越大。

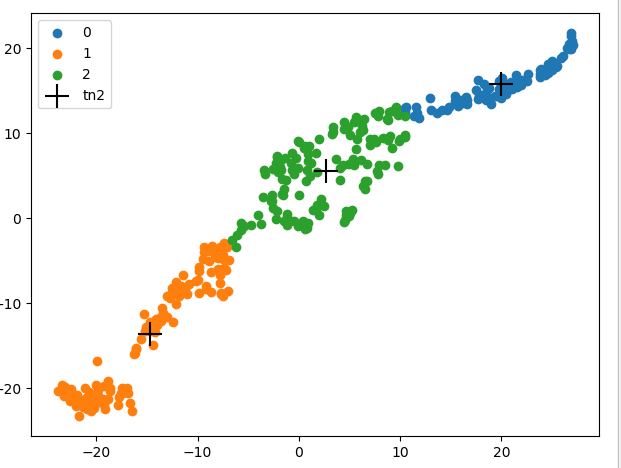
降维之后的数据在二维空间分布比较明显，接下来我们验证一下聚类的效果

通过统计每一个分类的通过逾期指标，目前整体的指标均值是

avg\_apply:3,6 avg\_approve:0.36 approve\_rate:0.10

avg\_overdue:0.21 avg\_maxoverdue:9.6 avg\_M3: 0.04





这些社群中我们看到cluster2 分布在中间部分，其各个指标都在整体均值附件波动，

Cluster0 整体逾期次数较多，最大逾期天数和M3指标也远高于平均水平

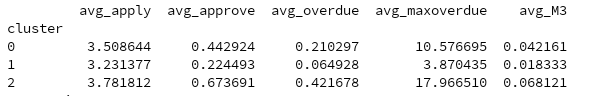
Cluster2 则是相反的，虽然通过率较低，但是逾期次数和最大逾期天数远远低于其他类别

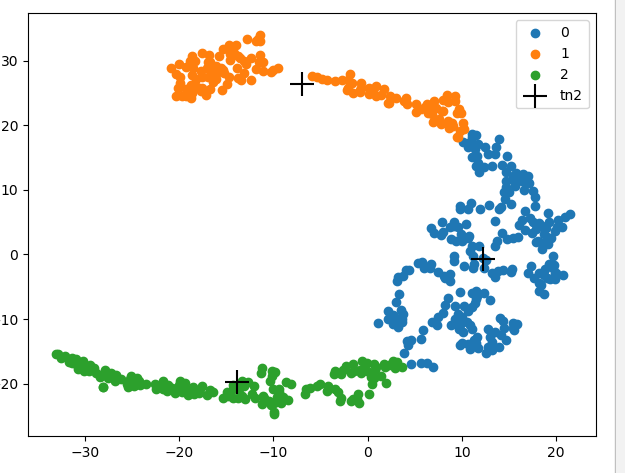
整体上三个类别申请频率相近，通过率有差别，但是逾期指标差别巨大，可以看出聚类很好的区分了不同的团体，从聚类的结果看，越靠近右上角，这样的团体欺诈风险应该是更高的，

同样左下角的团体应该是整体通过率首先不高，但是这不能代表这些团体不是欺诈，一个团体的大部分申请被风控拒绝，这样的团体风险可能更大。

Comm=4818846 33905 nodes, 189900 edges

模块度： 0.9507187583623611 长度： 523





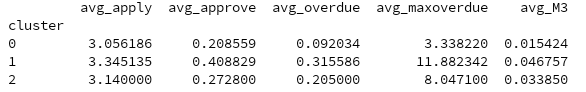
最终评价：我们最终的目的是为了评估团体分割的好坏，评估标准是什么呢？

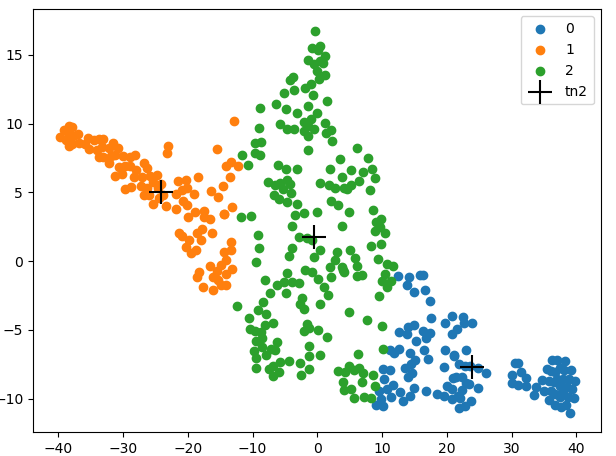
比如说我们从网络拿到了一个社区进行团体检测，baseline是什么呢。就是整个社区的各个平均水平，如果社团划分之后分为限定的几类，每一类的各个平均指标和整体指标没有差异性的话，那么社团划分的结果就是不好的，目前我们看到通过louvain算法划分之后的社团聚类结果可以将指标差异体现出来，说明社团划分效果还是不错的

最后我们再看几个社区节点数比较多的社区的情况

Comm=7875775 52171 nodes, 317673 edges

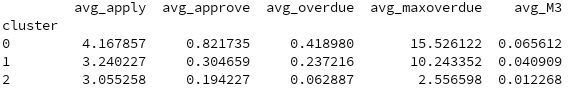
模块度： 0.9420552112223075 长度： 429

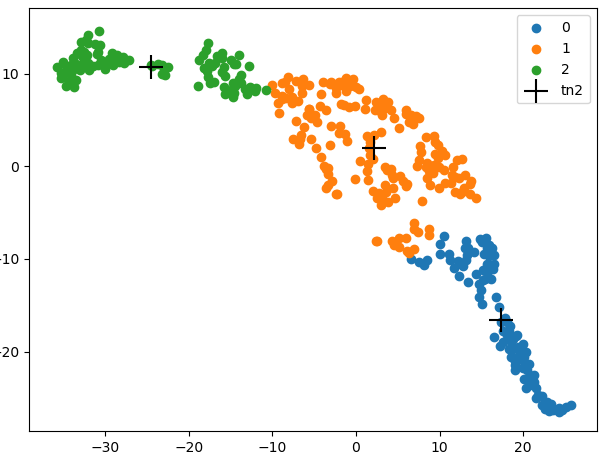




Comm=8998331 28057 nodes, 160177 edges

模块度： 0.9437944908518212 长度： 371





通过以上的可视化结果，我们发现其实很多社区的聚类结果呈现的特征是：

1 两头聚集程度高，中间较分散

2 对于逾期指标最低的团体与其他团体有明显的割裂感，有突然断裂的情况

3 对于逾期指标最高的团体明显的特征是聚集性很高

原因分析：  
聚集程度代表了团体相对于所在分类的特征变动程度，聚集程度越高意味着团体间的各个指标相关性越高。两个极端的聚集性都很好，说明两个极端的那些团体他们的各个指标都有高相关性，尤其是高风险团体，他们申请频率高，通过率高，逾期指标高，

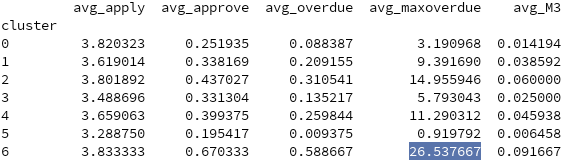
对于中间团体的分散，代表着局部的无关性，但是整体趋势看又有一定的相关性。

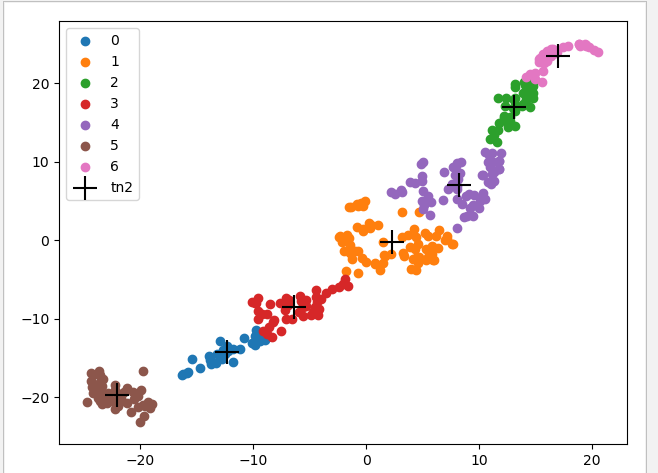
低逾期团体的割裂原因暂时未知

接下来我分了更多的类，如我所料，从整体的特征来看，是线性的且单调的，如果将特征降到一维就是一条单调的曲线

Cluster:5-0-3-1-4-2-6

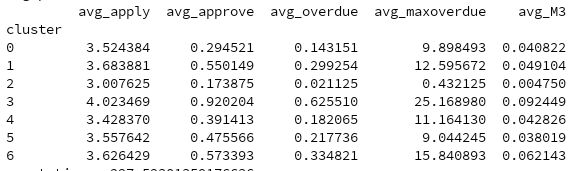
无论从那个特征去分析，我们都可以发现基本是单调递增的

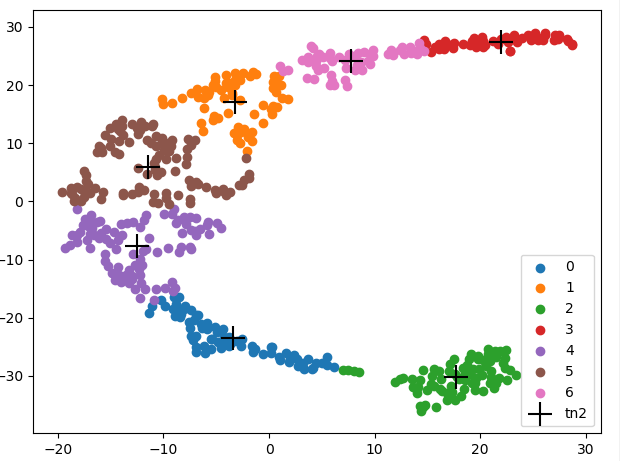




当然也不总是单调的：

这个社区中从4-5这两个分类趋势发生了变化，申请和通过率增长，最大逾期和M3却下降了，但是总体的上升趋势没有变，只是方向变了





最后我又对比了权重对于社团检测的影响，很明显的在默认权重的情况下，分得的社团数较少，而且逾期的区分度不高，可视化的情况下数据较为分散，说明默认权重的情况下效果不好。