

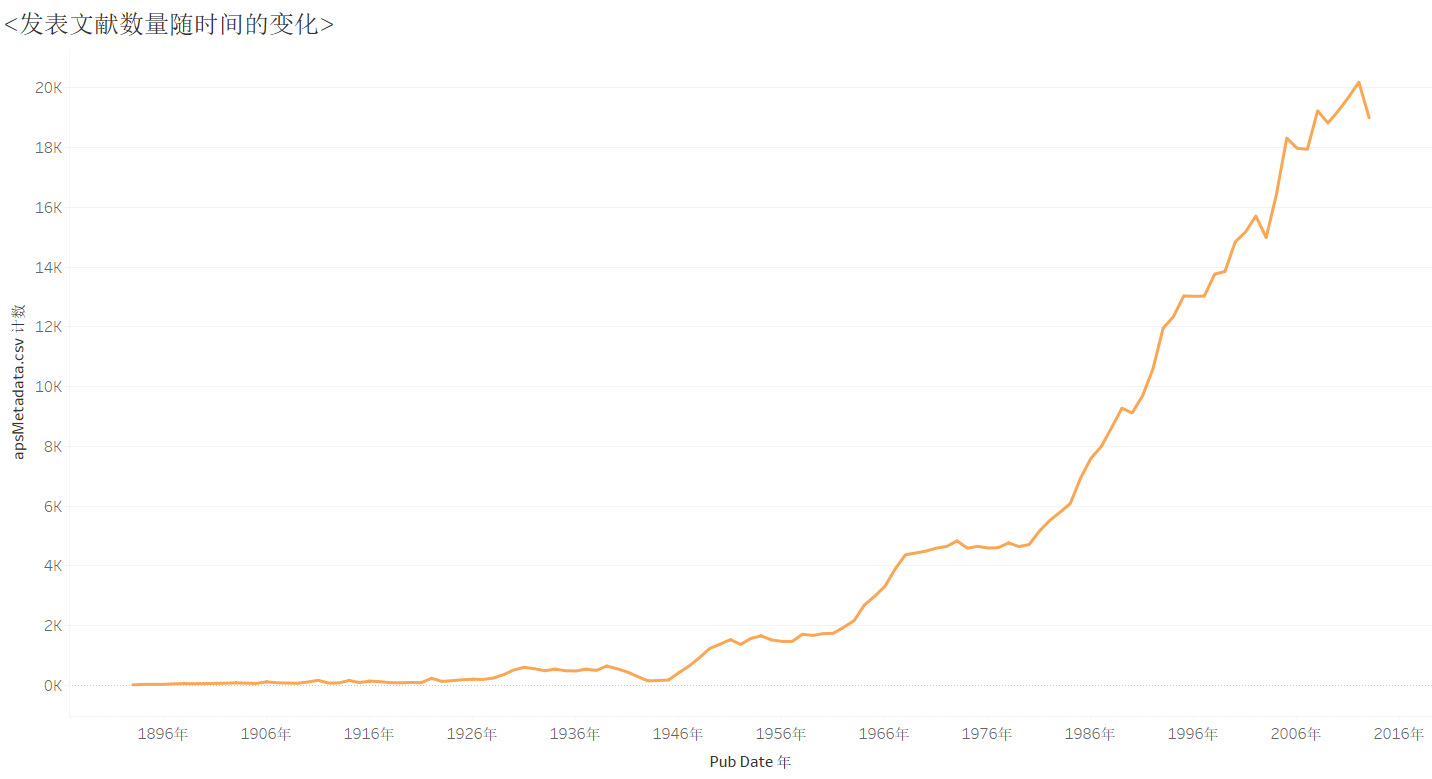
**数据可视化 个人作业二**

**时间和地理的可视化**

|  |  |
| --- | --- |
| **姓名：** | 吴凡 |
| **授课教师：** | 步一 助理教授 |

二〇二三 年 八 月

1. **时间可视化**

****时间可视化能够最为直观和简单地显示APS数据集中论文数量随着时间的变化趋势，笔者利用apsMetadata.csv中的Pub Date字段作为横轴，每一年发表数量的字段作为纵轴绘制了如下折线图。

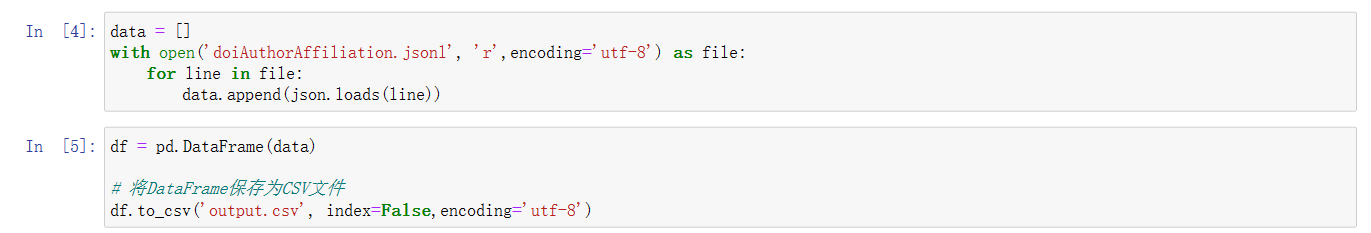
从图上可以看出：

1. 自1896到1946年，相关论文发表数量的起伏较小，总体上处于低谷期；
2. 自1946年开始，相关论文数量开始迅猛增长，增速在1980年到1990年间达到了一个峰值。
3. 保持总的增长趋势一直到2015年，2016年相较于2015年数量下降1K，略有下降，是数量波动还是增长趋势变缓还无法看出，还得进一步看之后几年的数据。
4. **空间可视化**

1.数据预处理

由于doiAuthorAffiliation.jsonl中的affliationsName字段的数据太过繁杂，不是规范的地理编码，因此无法直接用于tableau的地理可视化，因此需要对数据进行一定的预处理：

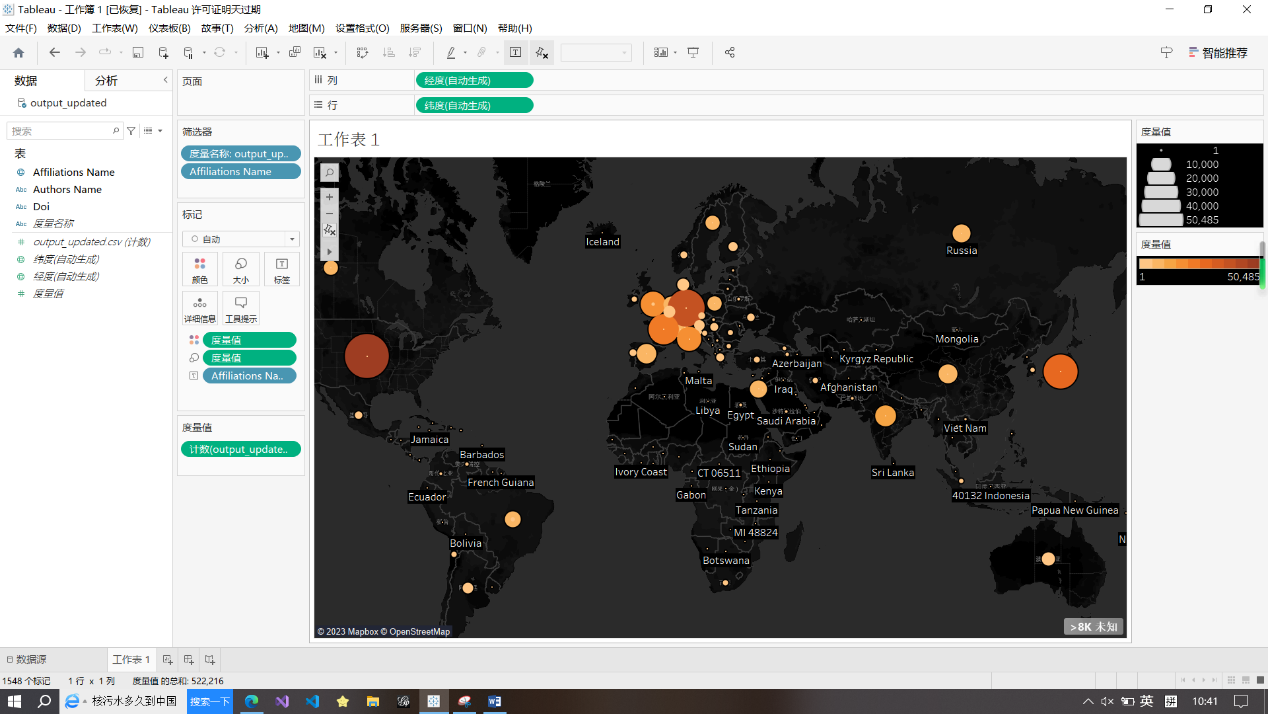
在这里，笔者在Jupyter Notebook中对该列数据进行了 一定的处理，处理的代码文件为Processed.ipynb，在这个文件中做了两件事：

（1）将jsonl文件转成csv文件作为output.csv输出；

（2）将affliationsName字段进行分割处理，只取最后一个逗号后面的字符串，忽略前面实验室的信息，以便形成较为规范的地理编码，得到output\_updated文件输出，在tableau完成地理可视化。

2.可视化结果分析

下图是完整的结果图展示：节点大小越大，颜色越偏向红色，说明发表数量越多。

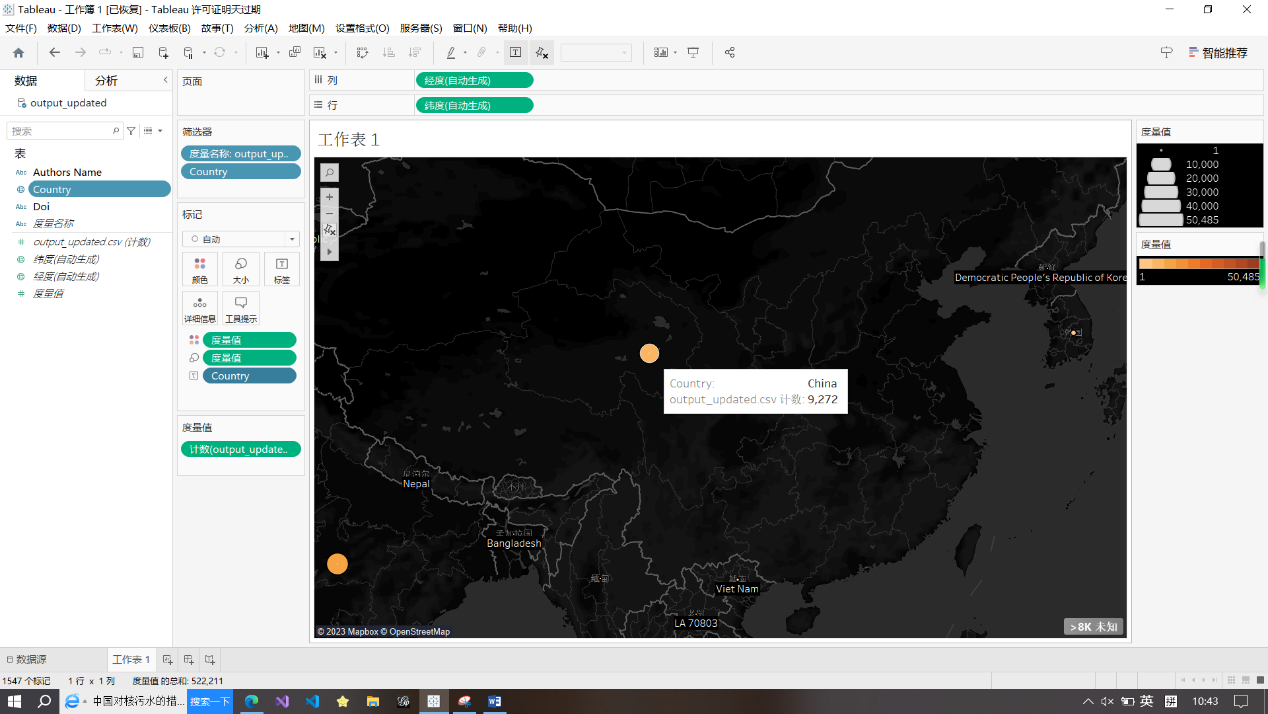


在得到上面这张图之前需要手动进行一些数据后处理工作，因为即使处理过了，还是有一部分数据Tableau无法识别为正确的地理编码，例如下图中的中华人民共和国就有很多条需要手动清洗数据：

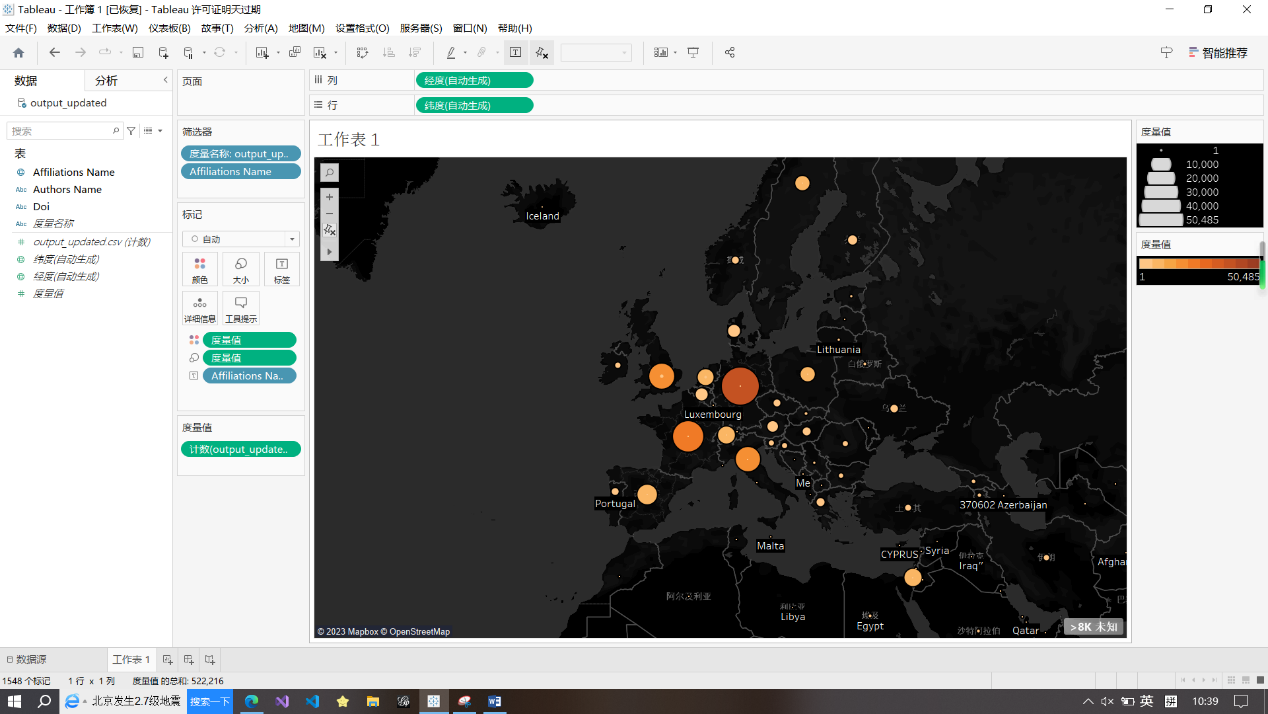


对于该地理可视化结果，我们需要注意到一些局部的信息：

1. 中国在APS数据集中共有9272条，在所有国家中处于中等水平，说明在这一块的研究中国还相对欠缺。



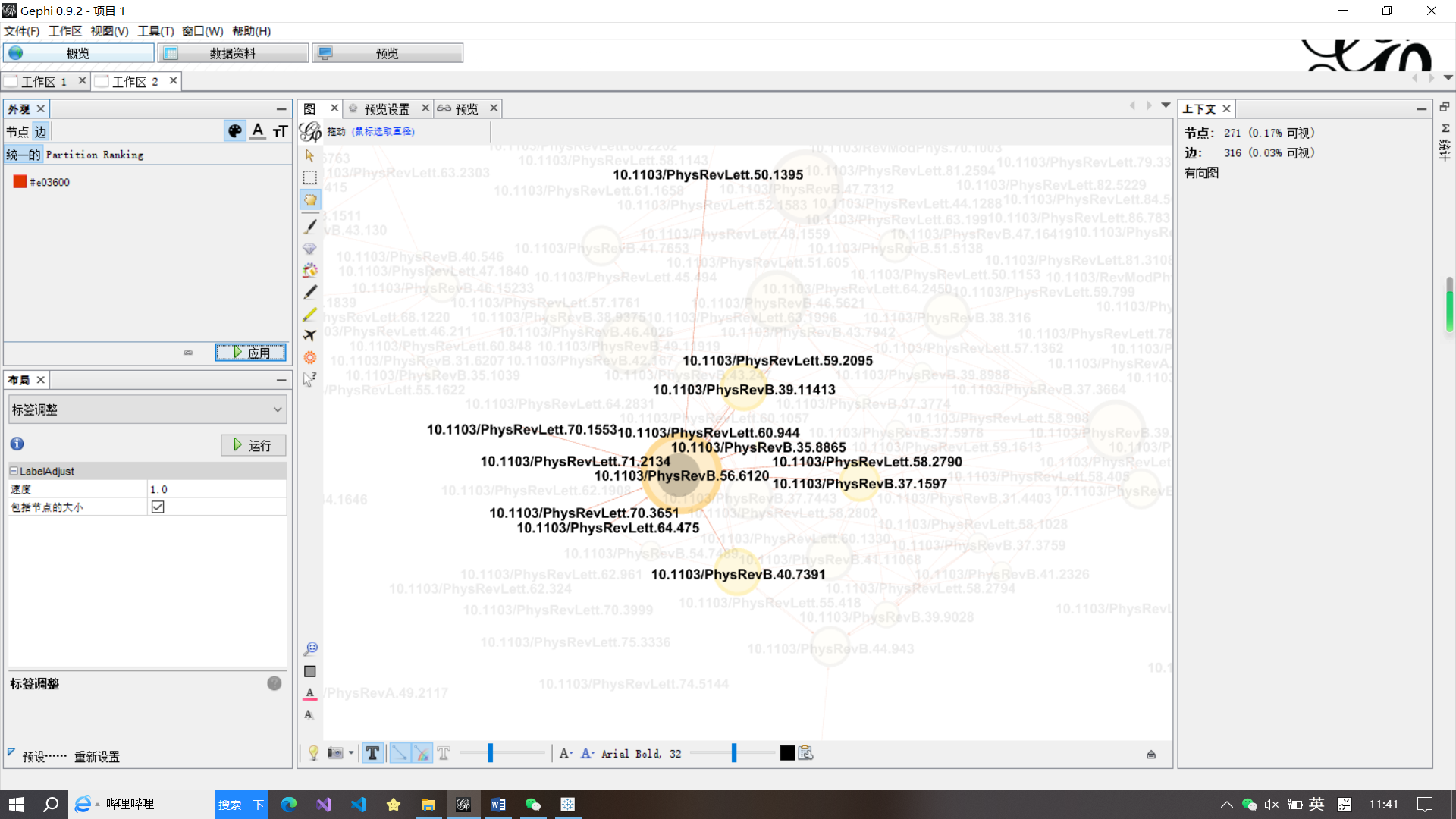
2.欧美国家肉眼可见地分布数量较多且分布位置较为集中。

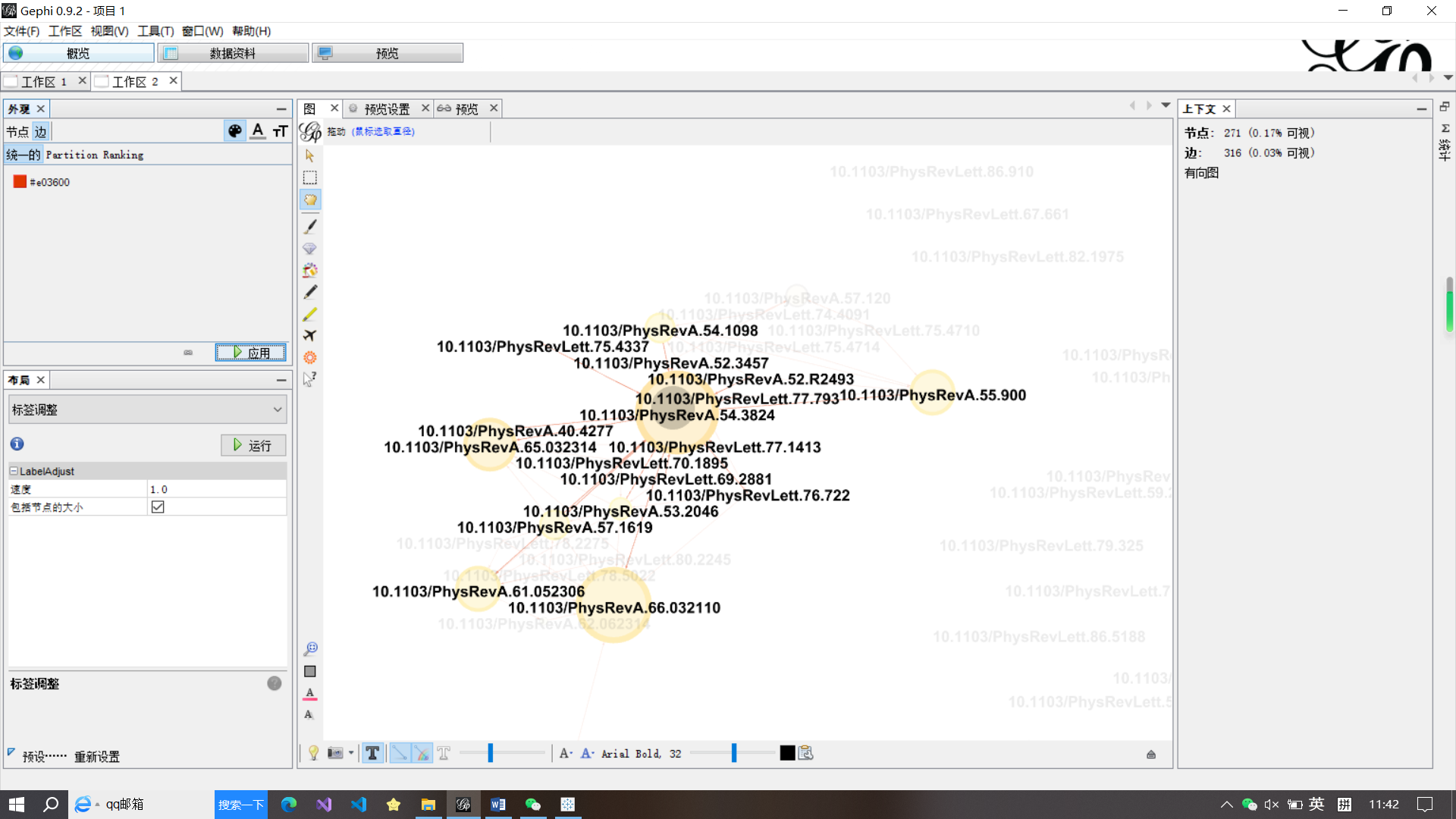


1. **网络可视化**

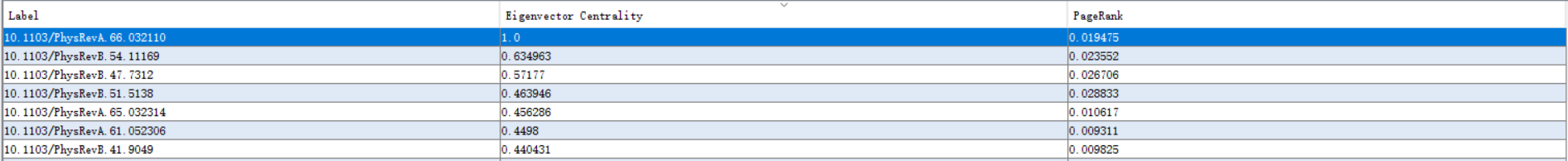
笔者根据APS数据集中的citationPairs.csv在Gephi工具下做了网络可视化，并用Gephi内置的统计工具结合apsMetadata.csv对结果进行了相应的分析。

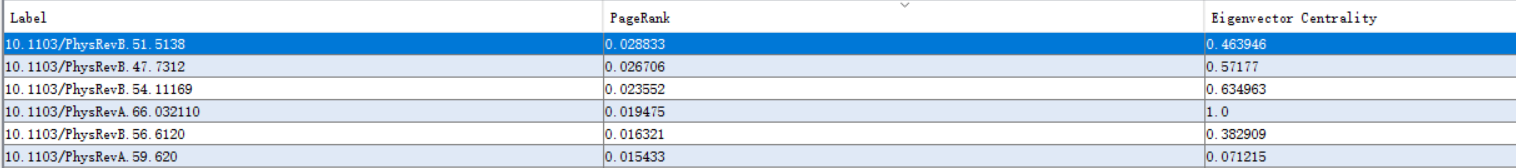
最终结果的gephi文件可以查看“被引用最多的是哪些论文呢.gephi”,由于节点数太多，这里只展现两个比较关键节点周围的部分信息（下面两张图文件夹中也包含gephi\_PIC1.png和gephi\_PIC2.png）：





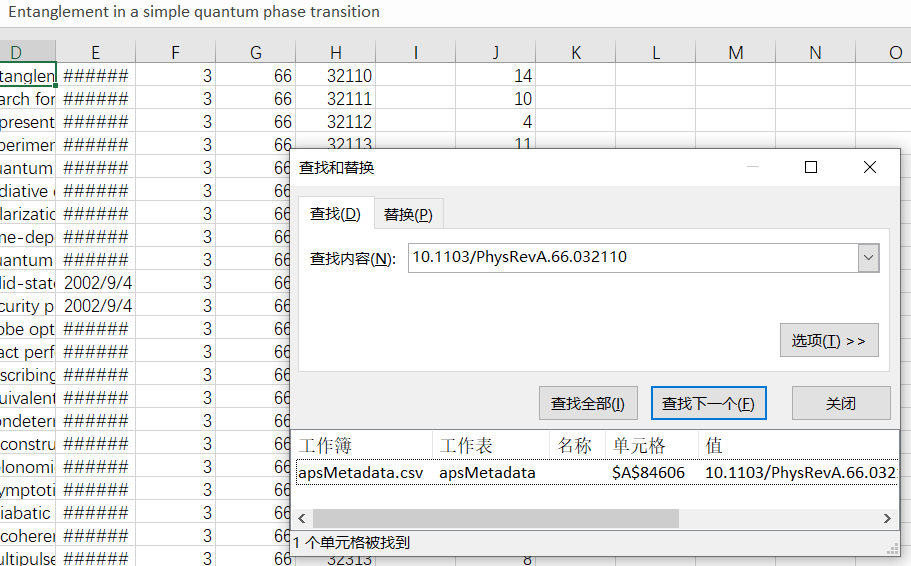
网络数据可视化的结果能够直观地展现出APS数据集中那些被引用较多的节点，也就是那些比较重要的节点，我们进一步利用Gephi内置的统计工具对数据进行分析，利用特征向量中心度和PageRank两个参数看看哪些节点比较重要：

（1）以特征中心度的降序排序前5的节点信息：

（2）以PageRank的降序排序前5的节点信息：

由上面两个统计结果，我比较好奇这两个排序排名第1的两篇论文究竟是什么，于是我在apsMetadata.csv进行检索：

（1）下图是按特征中心度排名第一的检索结果，论文标题为：Entanglement in a simple quantum phase transition(简单量子相变中的纠缠)



（2）下图是按PageRank排名第一的检索结果，论文标题为：Spontaneous interlayer coherence in double-layer quantum Hall systems:Charged vortices and Kosterlitz-Thouless phase transitions(双层量子霍尔系统的自发层间相干:带电涡和Kosterlitz-Thouless相变)