



北京大学

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：

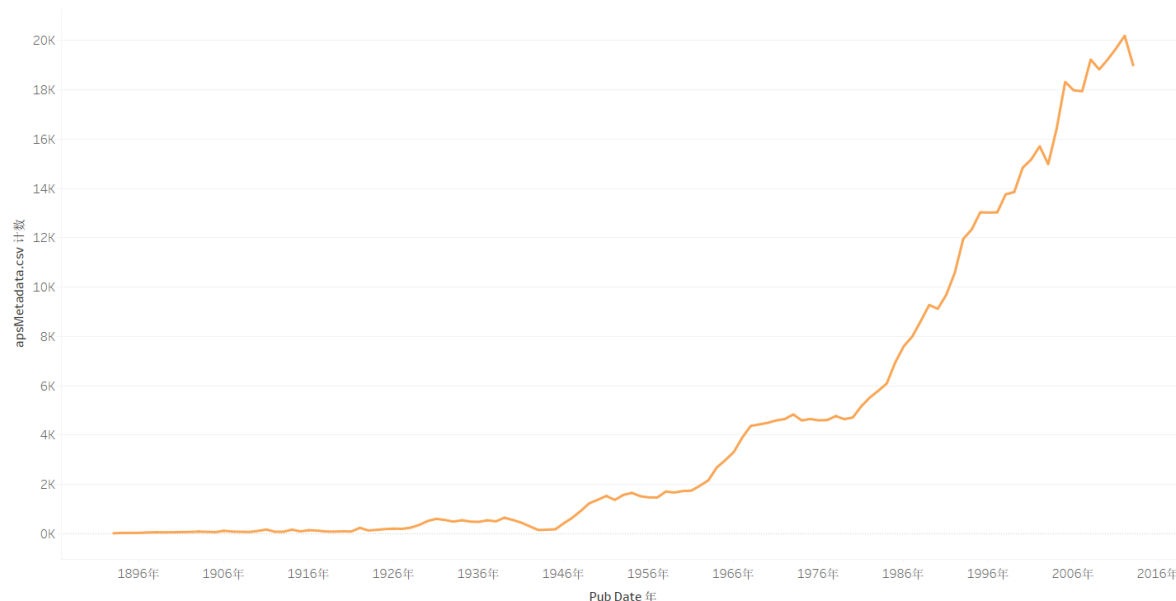
步一 助理教授

二〇二三 年 八 月

一.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。

<发表文献数量随时间的变化>



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。

二.空间可视化

1.数据预处理

由于 doiAuthorAffiliation.jsonl 中的 affiliationsName 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；

```
In [4]: data = []
with open('doiAuthorAffiliation.jsonl', 'r', encoding='utf-8') as file:
    for line in file:
        data.append(json.loads(line))

In [5]: df = pd.DataFrame(data)
# 将DataFrame保存为CSV文件
df.to_csv('output.csv', index=False, encoding='utf-8')
```

(2) 将 affiliationsName 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。

```
In [6]: import csv

input_file = 'output.csv'
output_file = 'output_updated.csv'

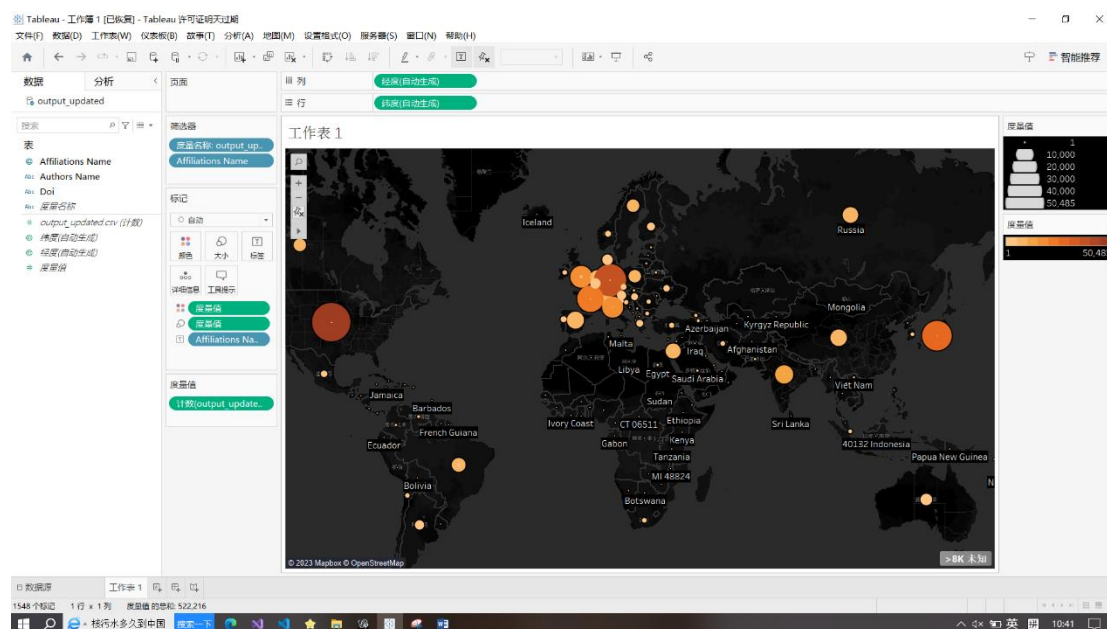
with open(input_file, 'r', newline='', encoding='utf-8') as file:
    reader = csv.reader(file)
    rows = list(reader)

    for row in rows:
        if len(row) > 0:
            data = row[2].strip("[ ]") # 去掉最前面的 '[' 和最后面的 ']'
            data = data.strip(",")
            data = data.strip(".")
            if ',' in data:
                data = data.split(',')[1].strip() # 只保留最后一个逗号之后的内容
            else:
                data = data.strip() # 如果没有逗号，则不做处理
            row[2] = data

with open(output_file, 'w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerows(rows)
```

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：

将值与位置匹配

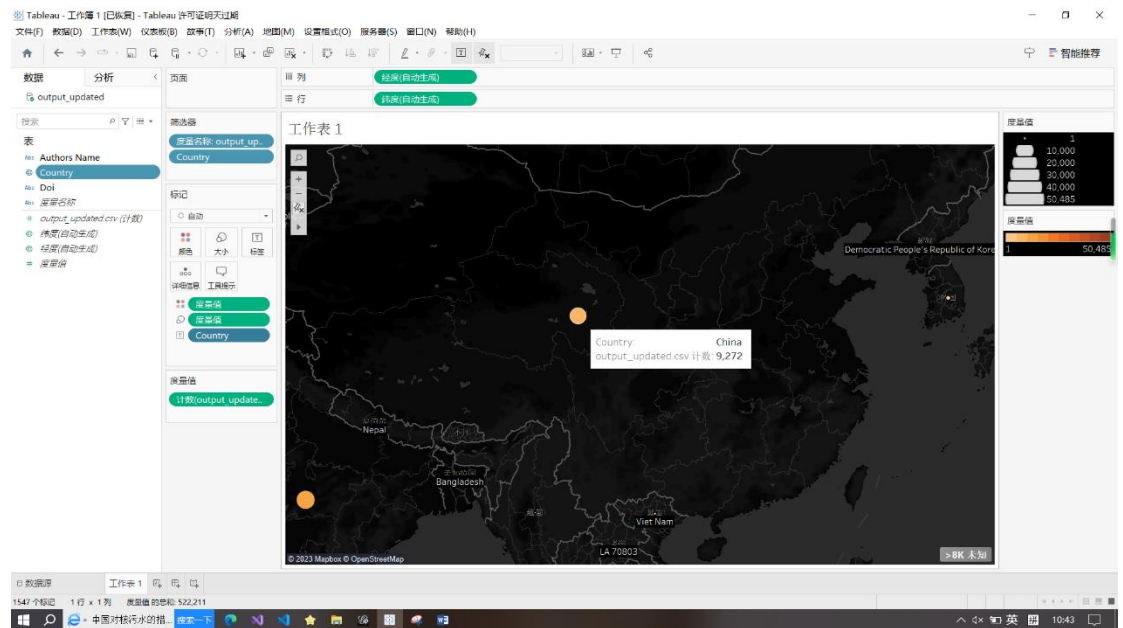
| 您的数据 | 匹配位置 |
|--------------------------------|------|
| People^{'s} Republic of China | 中国 |
| People' ĭ's Republic of China | 中国 |
| People' Republic of China" | 中国 |
| People Republic of China | 中国 |
| People' s of Republic of China | 中国 |
| Peoples Rebbublic of China | 中国 |

☒ 仅在下拉列表中显示没有匹配项的位置
☐ 显示所有值

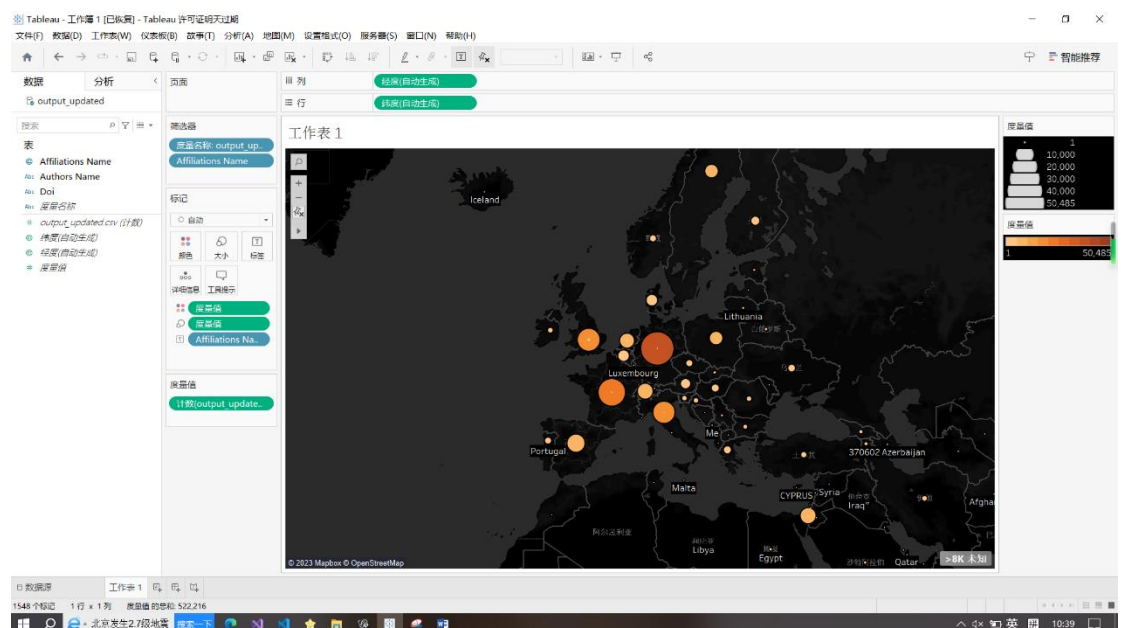
重置匹配项 确定 取消

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。



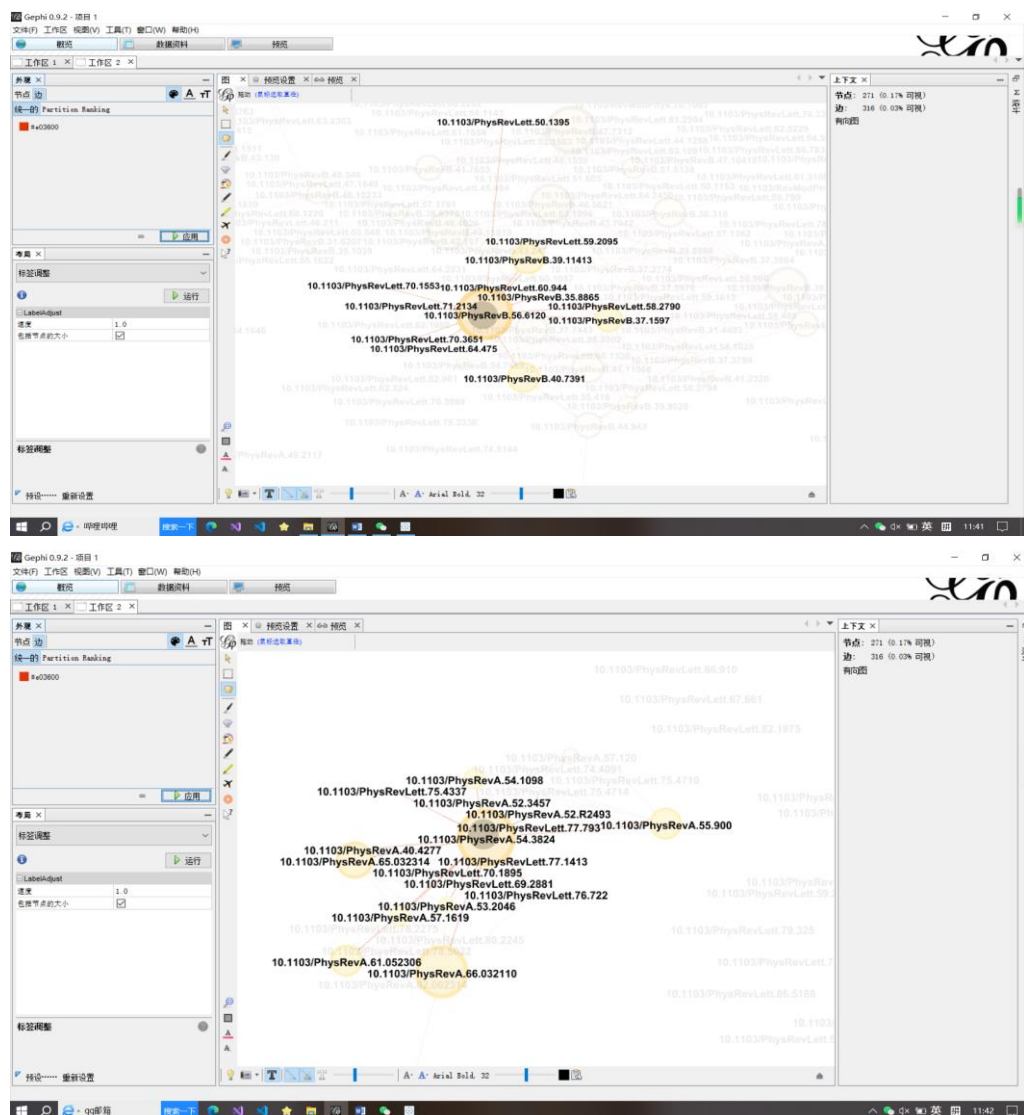
2. 欧美国家肉眼可见地分布数量较多且分布位置较为集中。



三.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:

| Label | Eigenvector Centrality | PageRank |
|----------------------------|------------------------|----------|
| 10.1103/PhysRevA.66.032110 | 1.0 | 0.019475 |
| 10.1103/PhysRevE.54.11169 | 0.634963 | 0.023552 |
| 10.1103/PhysRevE.47.7312 | 0.57177 | 0.026706 |
| 10.1103/PhysRevE.51.5138 | 0.463946 | 0.028833 |
| 10.1103/PhysRevA.65.032314 | 0.456286 | 0.010617 |
| 10.1103/PhysRevA.61.052306 | 0.4498 | 0.009311 |
| 10.1103/PhysRevE.41.9049 | 0.440431 | 0.009825 |

(2) 以 PageRank 的降序排序前 5 的节点信息:

| Label | PageRank | Eigenvector Centrality |
|----------------------------|----------|------------------------|
| 10.1103/PhysRevE.51.5138 | 0.028833 | 0.463946 |
| 10.1103/PhysRevE.47.7312 | 0.026706 | 0.57177 |
| 10.1103/PhysRevE.54.11169 | 0.023552 | 0.634963 |
| 10.1103/PhysRevA.66.032110 | 0.019475 | 1.0 |
| 10.1103/PhysRevE.56.6120 | 0.016321 | 0.382909 |
| 10.1103/PhysRevA.59.620 | 0.015433 | 0.071215 |

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)

Entanglement in a simple quantum phase transition

| D | E | F | G | H | I | J | K | L | M | N | O |
|-----------|----------|---|----|-------|---|----|---|---|---|---|---|
| tanglen | ##### | 3 | 66 | 32110 | | 14 | | | | | |
| arch for | ##### | 3 | 66 | 32111 | | 10 | | | | | |
| present | ##### | 3 | 66 | 32112 | | 4 | | | | | |
| erimeter | ##### | 3 | 66 | 32113 | | 11 | | | | | |
| iantum | ##### | 3 | 66 | | | | | | | | |
| diative | ##### | 3 | 66 | | | | | | | | |
| larizatic | ##### | 3 | 66 | | | | | | | | |
| ne-dep | ##### | 3 | 66 | | | | | | | | |
| iantum | ##### | 3 | 66 | | | | | | | | |
| lid-stat | 2002/9/4 | 3 | 66 | | | | | | | | |
| curity p | 2002/9/4 | 3 | 66 | | | | | | | | |
| obe opt | ##### | 3 | 66 | | | | | | | | |
| act perf | ##### | 3 | 66 | | | | | | | | |
| scribing | ##### | 3 | 66 | | | | | | | | |
| uivalent | ##### | 3 | 66 | | | | | | | | |
| ndeteri | ##### | 3 | 66 | | | | | | | | |
| constru | ##### | 3 | 66 | | | | | | | | |
| lonomi | ##### | 3 | 66 | | | | | | | | |
| /mptoti | ##### | 3 | 66 | | | | | | | | |
| iabatic | ##### | 3 | 66 | | | | | | | | |
| coherei | ##### | 3 | 66 | | | | | | | | |
| ltipulse | ##### | 3 | 66 | | | | | | | | |

查找和替换

查找(D)

替换(P)

查找内容(N): 10.1103/PhysRevA.66.032110

选项(I) >>

查找全部(I)

查找下一个(I)

关闭

| 工作簿 | 工作表 | 名称 | 单元格 | 值 |
|-----------------|-------------|------------|-------------------------|---|
| apsMetadata.csv | apsMetadata | \$A\$84606 | 10.1103/PhysRevA.66.032 | |

1 个单元格被找到

(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)

| | | | | | |
|--------|--------------------------|---------|-------------------|---|---|
| 171771 | 10.1103/PhysRevB.51.5130 | article | Physical Review B | Coherent transport through a coupled-quantum-dot system with strong intradot interaction | 7 |
| 171772 | 10.1103/PhysRevB.51.5138 | article | Physical Review B | Spontaneous interlayer coherence in double-layer quantum Hall systems: Charged vortices and Kosterlitz-Thouless phase transitions | # |