

## 实验二

请于 2023 年 5 月 9 日 17:59 之前提交至课程邮箱 [ustcweb2022@163.com](mailto:ustcweb2022@163.com)  
并于 2023 年 5 月 10 日（周三）课上进行现场汇报

### 总体实验要求：

请组成 4-6 人小组，围绕指定数据集进行自定方案分析实验，记录实验过程并撰写实验报告。

### 数据背景：

数据集来自某组织成员对组内提案的投票情况统计，其机制为：

- 组织中的每名成员归属于两个子社团之一，两个子社团的成员一般情况下不会发生重叠。
- 在该组织中，每次提案会议有固定人数的成员参与，部分成员在会上提出提案，而其他参会成员对该提案进行表决。一个提案可能会被多次提出。
- 提案可以由某位成员单独发起，也可以由若干名成员联合发起。
- 参与会议的组织成员可以自由选择“赞成”、“反对”、“弃权”。并非所有“cosponsors.json”里面的提案都会被投票表决。

### 实验数据：

本数据集一共包含该组织在若干次提案会议中表决过的 8849 项提案。

数据使用方式：将下载的压缩包解压缩即可。

### 文件说明：

文件以.json 的格式进行存储，主要包含三类文件，分别对应提案投票情况（Votes）、历次参会成员（Members）以及提案信息（Bills）。另外还提供了成员所在地的映射信息（State）。以提案信息 Bills 为例，其 json 文件格式如下：

```
{'bill_id': 'b144-104', 'bill_type': 'b', 'actions_dates': ['1996-01-03'], 'sponsor': 'O000007', 'cosponsors': []}
```

其中，每个 dict 中的"bill\_id"代表某个提案的 id，actions\_dates 表示提案提出时间，"bill\_type"代表该提案的类型，"sponsor"代表该提案的提出者，"cosponsor"代表该提案的联合提出者（若为空表示该提案无联合提出者）。**需要注意的是，并非所有提案都有相应的投票记录，请小心甄别。**

更详细的数据说明请见压缩包下的 readme 文件。请根据实验需要，自行提取并处理数据。

### 实验内容：

要求对于指定数据，自行设计实验方案及实验目标，并根据数据给出量化的结果分析。

具体实验内容包括：

#### （1）实验目标选定

本次实验的最终目的是解决一个基于社会网络的预测性问题，问题由小组自行商议决定。预测性问题要求对数据根据时间戳（提案会议发生的时间）进行拆分，并利用历史数据对未来情况进行预测。问题本身需要有**明确的可验证性及对应的量化指标**。

一些可供参考的选题包括（仅供参考，自行确定选题）：

- 预测成员对提案的投票结果
- 预测成员的活跃性（如参与提案发起的频率）
- 预测成员是否会参加某个提案的联名发起
- 预测成员未来的网络关系会发生何种变化
- 预测成员在提案的主题/标签偏好上的演变（可以从提案信息中获得）

**注意：**我们将参考工作量进行评分，过于简单的选题将影响到最后的得分。

## （2） 动态社交网络的构建

由于数据集本身并未包含显式的网络结构信息，为了更为完整地实现基于社会网络的分析 and 预测，需要基于掌握的信息自行构造一个社会网络。**常见的构造方式**如根据成员之间的标签相似性、成员共同发起过的提案，或者成员对提案的态度等进行构造，并可采用相似性或共现次数等因素进行边权重的量化估计。

**本环节的要求：**

- 根据需要自行定义社会网络中节点和边的定义，并设计社会网络构建方法，同时说明设计方案的合理性依据。
- 网络中的每一条边应具有权重，权重的计算方式自行定义，并说明其合理性和意义。
- 所设计的网络要随着时间推移（即提案会议的发生）而发生动态变化，包括并不限于新增/删除节点、新增/删除边，边上的权重变化等。时间信息可以在 `members.json` 中获得。

## （3） 围绕社交活动的量化分析部分

在完成社会网络的构建后，请围绕拟开展的研究课题，首先进行统计分析，确认构造的社会网络对于研究课题是否具有显著作用（可通过显著性检验、因果推断等手段加以体现），并确定作用方式以辅助下一阶段的建模预测。

一些可供选择的分析内容包括（仅供参考，请结合实际自行设计方案）：

- 不同网络构造方式对于结果的影响
- 网络结构对于成员态度/倾向的影响，以及成员态度/倾向对于网络结构的反向影响
- 网络结构演化对于结果的影响
- 网络是否加权对于结果的影响
- 网络稀疏性/新节点（冷启动）等问题对于结果的影响

## （4） 自定义任务的预测实验部分

最后，根据选定的社会网络构建方案及相应的统计分析，设计模型解决预设的目标问题，并给出相应的测试方案和测试结果。

**本环节的要求如下：**

- 请自行设计模型完成预测。不要求采用深度学习方法，我们仅根据模型的合理性进行评价，模型的复杂度并不是评分的主要依据。
- 训练集/验证集（如需）/测试集的比例自行确定，但测试集比例不低于 20%。建议分析比较不同划分方式和比例对于结果的影响。
- 请完成必要的消融实验，分析比较考虑/不考虑社会网络信息对于结果的影响。
- 必要的参数敏感性讨论和必要的案例分析。

### 提交说明：

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为“社会计算第二次实验报告\_学号\_姓名”。

- 例如：“社会计算第二次实验报告\_SA20011999\_法外狂徒张三”
- 标题仅写明小组组长（即汇报人）的学号及姓名即可，其他成员请在邮件及实验报告正文中注明学号及姓名。
- 因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 请注明所采用的算法，并列举必要的参考文献。
- 请采用必要的图表以更清晰地展示实验结果。
- 提交报告的同时请提交源代码以供检查。
- 除非特殊情况并事先征得许可，否则迟交报告将不再被接收，并取消答辩资格。

### 报告要求：

由组长进行汇报，汇报总时长为 15 分钟，包括 12 分钟 PPT 讲解与 3 分钟提问+点评。

- 报告内容应包括选题设计、数据处理方式、社会网络构建方式、统计分析情况、采用的预测模型、相关参数的设置、实验结果及其分析、组内成员分工等。
- 报告顺序按照实验报告接收的顺序为准，名单将在报告前一天晚上于课程群内公布。
- 助教将根据汇报内容和实验报告内容进行打分，并计入总评成绩。

### 额外说明：

每组提交一份实验报告，所有组员得分相同。但考虑到组长的额外工作量（协调组员工作并进行汇报），组长将获得额外的 1 分加分。

如有未尽事宜，将对本说明进行进一步更新。