近期有不少同学问两跳后得到的子图大概在什么规模才是合适的，我们测试后的结果为

- 子图文件的大小（txt文件）为 7MB
- 子图包含的实体数量（包含578个电影实体）为 1357 个
- 子图包含的关系数量为 25 个
- 子图包含的三元组数量为 54794 个

子图规模在这个量级，并 **确保子图中包含了578个电影实体** 就行

我们具体的处理流程如下，

1. 得到一跳子图后，我们采样 20 核的设置，即只保留了至少出现在 20 个三元组中的实体，同时只保留出现超过 50 次的关系，由此得到的一跳子图包含了 758 个实体
2. 基于上一步得到的 758 个实体，生成两跳子图，这个过程以压缩形式完成的，得到的 graph_2step.gz 文件大小约为 1.4GB
3. 对两跳子图的处理：先过滤掉出现超过 2w 次的实体和出现少于 50 次的关系；然后再采样 15 核的设置，同时只保留出现大于50次的关系，对两跳子图进行清洗
4. 最终得到的两跳子图的规模如上所示

最重要的是，我们鼓励大家，根据实验一中提供的Tag信息，新增Tag关系，建立其与电影实体的三元组，以充实电影实体的语义信息。