Title (e.g., Report for Project xxx)

Name

ID

1. **Preliminaries**

Describe the goal of your project and any algorithms, software or codes that is used in your project.

近年来，微博、QQ、抖音等等社交软件的兴起使得社交网络快速发展。在社交网络中，信息流传的速度和质量决定了这个社交网络的发展。一些商户在发放商业信息的时候往往会选择一些影响力较大的用户，以用更低的成本获得更高的收益。影响力最大化问题就是研究这样的情况的。

在影响力最大化问题中，我们需要在给定的网络中，选取一定数量的节点作为初始被激活的种子集，使得这个种子集在这个网络中的影响力最大。

The applications of the problem.

影响力最大化问题是个很难解决的问题，它的算法只能随时间推移不断优化。当不断被优化时，这样的算法可以应用于商户们在社交网络上投放商业信息。当然，当遇到比较重大的问题时，政府也可以在最低的时间成本上将紧急消息通知给最多的民众。

1. **Methodology**

Describe the details of your representation/algorithm/architecture, etc.

Notation:

Data structure:

Pool: 进程池，用于多进程计算结果

Nodes: 一个用于表示图的字典，键为节点id，值为一个二维列表，第一维表示从这个节点作为起点的边的终点结点，第二维表示对应的权重

Async\_result: 多进程运行列表，存有进程池里所有正在运行以及等待运行的函数

Results：多进程运算结果的列表，存有所有已完成进程的运算结果。

Seeds：在ISE中表示初始被激活的种子集，在IMP中表示结果种子集

Activity\_set: 每一次迭代被激活的节点列表

Total\_set：每次完整的采样流程被激活的节点列表

R: rr集的列表

Sk: 在R中挑选出的k个最优结果节点

Model design:

1. ISE

在IMP问题中，我们需要对种子的影响力做计算。一般来说，社交网络有两种基本的流动模型，一个是独立瀑布模型，另一个是线性阈值模型。

1. 独立瀑布模型

在社交网络中，两个被连接的节点之间有一个传播概率，被激活的节点都有概率去激活它的出边的终点节点，而这个概率与历史无关。无论激活成功与否，尝试激活出边终点节点的节点将不会做再一次的激活尝试。

1. 线性阈值模型

一个节点被所有邻居节点所影响，如果影响的值加在一起大于这个节点的初始阈值，则这个节点被激活。这样的模型有历史效应。

为了得到更为准确的结果，需要对种子集进行许多次的采样取平均值，因此这样的问题可以使用多进程进行计算，在规定时间内计算尽可能多的结果。

1. IMP

在IMP问题中，我们使用IMM算法，先对图进行采样，在规定大小范围内计算rr集并加入R中，再经过一轮选择，最终获得需要的种子集结果。类似地，我们也可以在IMP中使用多进程，并发计算多次imm的结果放到results中，再迭代results获取出现次数最多的k个节点作为最终的结果集。

Detail of algorithms:

IMM

IC\_sample: 将初始的种子集作为激活的种子集，每一轮用上一轮被激活的种子集去激活从它们出发的边的终点结点，对于每一个出边，随机产生一个0~1的浮点数，如果这个浮点数小于这条边的权重（即影响概率），那么这个终点节点就被激活，并被加入到新被激活的节点列表中。最终返回的是总共被激活的节点数，或者是在激活状态的节点列表（伪代码以前者示例）。

LT\_Sample：将初始的种子集作为激活的种子集。在算法开始时为每一个节点随机产生一个0~1的初始阈值，每次迭代至当前节点时，统计入边中起始节点处于激活状态的边的权重和，如果大于初始阈值则被激活，并被加入到新被激活的节点列表中。最终返回的是总共被激活的节点数，或者是在激活状态的节点列表（伪代码以前者示例）。

IMM：对图采样生成一个包含rr集的列表R，然后使用node\_selection函数对其中的节点进行选择，获得最终的结果集

Sampling：采样函数，对图进行采样，获得rr集的列表R

Node\_selection: 根据R来迭代计算选取到的节点的边际效益，选取出边际效益最高的k个节点

1. **Empirical Verification**

Describe the experiments that you conducted to test/verify the quality of your program. This may include (but not limit to) the following:

* How were the experiments designed?
* What data did you use?
* How did you measure the performance?
* Experimental results
* Did the results meet your expectation about the program? Why or why not?

Dataset

ISE阶段，仅使用了network.txt, NetHEPT.txt, network\_seeds.txt作测试。IMP阶段，除了提供的三个文件之外，自己写了一个可以随机生成n个节点m条边且边权重符合题目要求的图的文件test\_data.py，用来计算在不同数据集下程序的运行效果。

Performance Measure：

ISE阶段，仅依靠对比计算出的结果及排行榜上的数据即可判断程序的效果。

IMP阶段，影响生成种子的速度和质量的有诸多因素。第一是图的复杂程度，图中节点的数量和边的数量都会影响计算速度。第二是k的大小，会影响selection的速度。

Hyperparameters:

Epsilon,l：用来控制每次迭代中R的长度的超参数，为了保证每次的rr集既足够有效又不至于花太多时间，因此使epsilon=0.5, l=1以获得比较优的结果。

Experimental results:

总结：

在此次project中，ISE和IMP都使用了多进程以充分使用资源计算更多的结果，这样使得我的结果更为准确有效。然而，在imp中，用多进程计算结果并选取最多的k个节点的做法并没有理论上的证明，数据有时候也不一定比单次计算更好。

经过这段时间的学习，我初步了解了影响力算法和python的多进程编程，为将来的职业发展打下了一些基础。在未来，我可能会阅读更多与IMP相关的文献，主要研究它们的理论基础，以优化我的算法而不是单纯靠大量的计算取胜。

1. **References**

List the references, please follow the IEEE format to prepare your references. The IEEE format can be found at:

http://ieeeauthorcenter.ieee.org/create-your-ieee-article/use-authoring-tools-and-ieee-article-templates/ieee-article-templates/templates-for-transactions/