探究集中采样算法🡪保持原始图特征的的情况下减少图的size

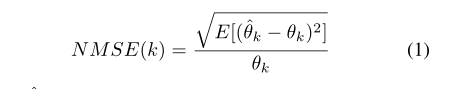
我们之前是扫描了整个图来统计catalog中的Qk，但那样扫描需要太多时间。论文idea就是探讨多种抽样方法，使得仅抽取图的一部分数据统计，但同时也能尽量保持原图特征。并同时规定了聚类系数和概率分布等indicators

CDF—Degree

CDF—CC

NMSE：

K度节点的归一化均方误差：

用以表示节点度分布（入度出度）。

CC：

聚类系数，是图中节点聚在一起的程度的度量

NACC：

图中所有节点的网络平均聚类系数

RE：

相对误差，用以量化算法在不同图属性上的采样性能

即采样所得到的度量，和原始图所得到的度量

抽样方法：BFS, RW, MHRW, USDSG， FS

RW：随机遍历，效果差

BFS： 更偏向度高的节点

MHRW：近似均匀分布

FS：NMSE最小，在程度或聚类系数较小时，FS表现不佳

MHRW：

1. 选取非0度节点起始点v
   1. 记其度kv
2. 从该点的所有相邻中任意选取点w
   1. 生成0~1的随机数p
   2. 若随机数p<kv/kw,则选取该邻点，否则舍弃
3. 重复步骤直至达到预算

估计函数来估计度分度🡪大数定律🡪收敛于真实值

MHRW和FS的性能对数据集的依赖性较大，在紧密连接的图中均表现较好，而FS比MHRW收敛更快、更准确。

FS：

1. 选出一组点集S（随机，无关）
   1. 获取每个点的度
   2. 计算每个点的选取概率，公式：p=
   3. 选取点v
2. 从v出边中均匀随机的选一条（v,w）将点集中v换w
3. 重复以上步骤直至达到预算

聚类系数分布也很好，但在聚类系数较小时不佳

Q1: 预算？算法至何时停止？

Think: 在有实验数据后，投入使用时可以设定lower bound和upper bound。进入lower bound。在进入范围内后，周期性进行比较indicator的值或分布，是否收敛。若收敛，则认定抽样完成

Q2：现在只有抽样方法，如何抽样后（或抽样的过程中）统计|Qk|?

宇凡学长之前有讲过一个idea，对于每个分开后的子图使用

Q3：分布式

Think:

对于FS,本身就是采取一组点集，通过边进行拓展。分布式的情况下可采取多个不相连点集（或在预期跳数内不会相连的多个点集）

对于MHRW， 选择多个在预算内不会互相拓展到的起始点，分别进行MHRW步骤的采样。

1. 可以不考虑将分布式后的多个图合并再讨论是否收敛，只要在该小图内部收敛稳定即可

2. 若探讨多个图合并后的收敛，能有效避免某个子图的构成与其他图完全不一样（比如某个人际关系子图是多个交际花组成的，而另外几个都是由多个不善交际的组成的）（已排除孤立点）

各自/合并后 在采样完成（或采样过程中），将Q2收集到的|Qk|更新至catalog