数据分析二

https://github.com/OneCircle1/MengProgram2

# 背景

在对数据进行分析后，数据的一些特征可以被发现，比如关键节点、社群等等。

# 任务

* 了解图的基本概念
* 利用networkx，matplotlib等工具对数据进行可视化分析
* 结合pagerank算法找出关键节点
* 结合LPA等算法对网络做社群发现

# 算法介绍

* Pagerank
  + A picture containing object

    Description automatically generated
  + 每一个网页的影响力，都是指向该网页的单位影响力的和
    - 指向网页的数量越多/指向网页的网页越重要，网页影响力越大
  + 问题
    - 没有指向
    - 指向自己
  + 优化
    - 限定迭代次数
    - 增大epsilon的值，让迭代尽早结束
* LPA
  + 给每个点一个label
  + 每一次更新，一个点的label都会变成它周围label数量最多的那个（当数量相同时随机选择一个）
  + 直到收敛
* FastUnfolding
  + Modularity
    - Q=12m∑i,j[Ai,j−kikj/2m]δ(ci,cj)
    - Q=∑c[∑in/2m−(∑tot/2m)2]
  + FastUnfolding
    - 把每个节点划分在不同的社区中
    - 更新节点所属的社区。如果划分到相邻社区中后Modularity增加，则划分
    - 直到社区不再变化

# 数据情况

csv文件，数据格式为

1,2

1,3

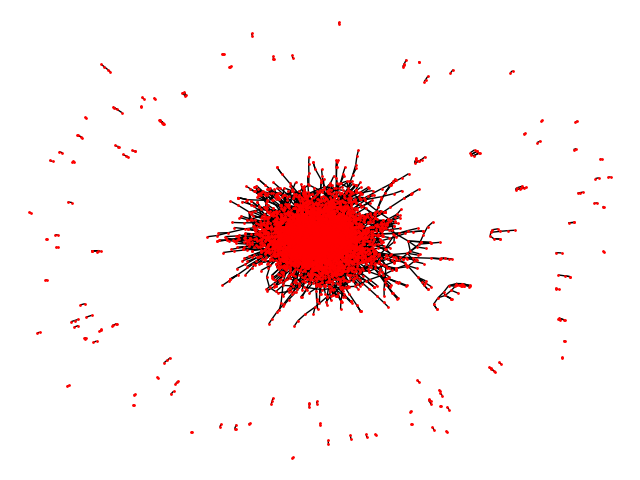
5,9

9,12

…

# 实验结果

* 可视化
  + 所有数据



* + Subgraph

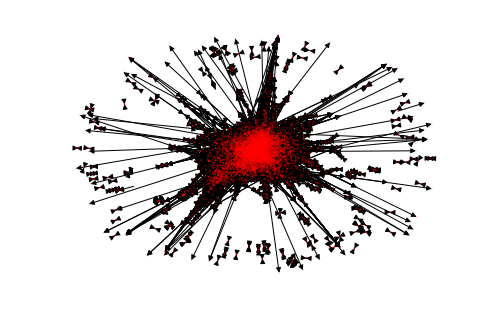
A close up of a fruit tree

Description automatically generated

* Pagerank

[8284 rows x 8284 columns]

[('275995436', 0.019466309905308385), ('476410313', 0.01946381594026822), ('809524687', 0.01600534103045328), ('390148878', 0.016003042960857406), ('249422639', 0.011717182749134037), ('123570525', 0.009844364953666397), ('947272756', 0.008573799682473652), ('676285306', 0.008572709691425888), ('317866087', 0.005860695552687871), ('272834362', 0.005857889789848505)]



* LPA

A picture containing flying, sky

Description automatically generated

# 结论