**实验报告-互评2-频繁模式与关联规则**

学号：1120203554 姓名：李元盛 班级：07032001

# 实验任务

在题目给出的数据集中自行选择2个并进行分析与挖掘。本实验中从SNAP(Stanford Large Network Dataset Collection) 中选取了

* Twitter Interaction Network for the US Congress
* LastFM Asia Social Network

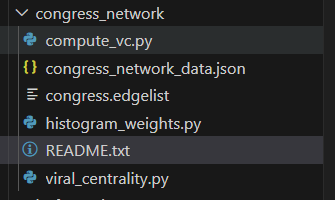
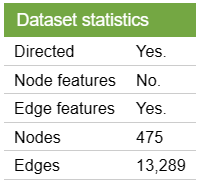
两个数据集展开实验。

# 数据获取与预处理

## 2.1 数据集Twitter Interaction Network for the US Congress

数据集Twitter Interaction Network for the US Congress代表了第117届美国国会参议院和众议院的Twitter互动网络。基本数据通过Twitter的API收集，然后根据一名成员转发、引用、回复或提及另一名成员的twitter的次数比例来量化一个叫做“传输概率”的经验值。

数据的基本信息和文件结构如下：



如上所示数据本身是一个**图结构**，节点是Twitter的用户，边是他们发送消息的**单向“传输概率”**（边是有向的），共475节点+13289条边。数据集由两个主要文件构成。

congress\_network\_data.json包含了以下几类键值对。

inList：包含给某个用户发送连接的其他用户（可认为是入边）。

inWeight：与inList一一对应包含了上述每个用户之间的边的传输概率，可视为边的权重。

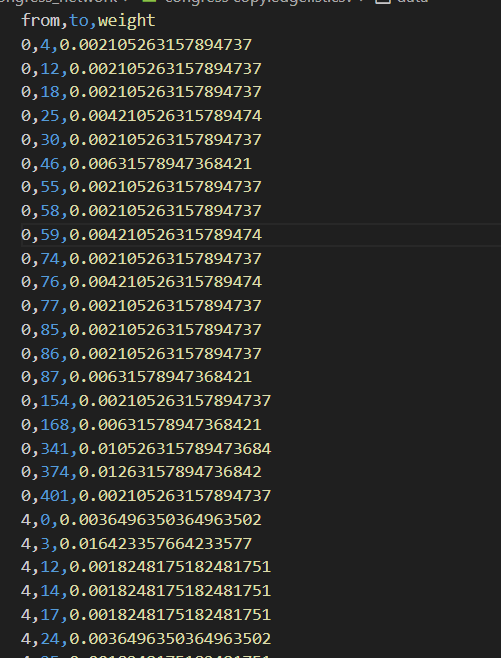
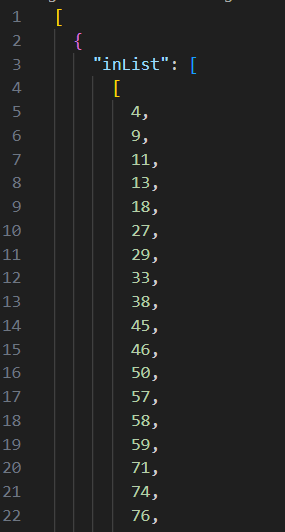
outList：包含从某个用户发出连接的其他用户（可认为是出边）。

outWeight：同上，包含了用户之间边的传输概率（边的权重）。

UsernameList：所有用户节点ID对应的用户名。

congress.edgelist文件包含了所有边的数据。

接下来对数据进行清洗和预处理。经过检查，数据中没有错误和缺失的项目。随后，我们将Json文件的格式进行调整，使其更易于观察和处理，如下左图。并且，我们将edgelist这个文件的格式修改为csv文件，如下右图。

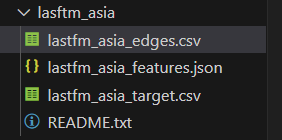


## 2.2 数据集LastFM Asia Social Network

这个数据集是从LastFM用户社交网络通过公开API收集的图结构数据。数据的节点是来自亚洲国家的LastFM用户，边是他们之间的关注关系（有向边）。顶点特征是根据用户喜欢的艺术家提取的。该数据可用于一些位置预测、节点分类的任务。

数据共有7624个节点，27806条边。密度为0.0009，传递率0.1787。

数据集文件共有三个。

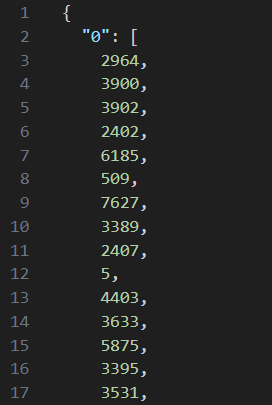
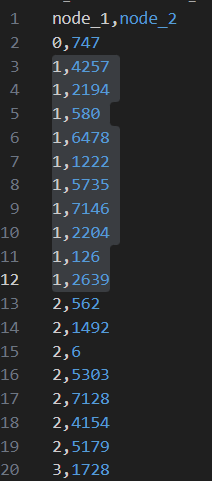


其中，lastfm\_asia\_edges.csv是边，共两列整数，代表边两端的节点。

Lastfm\_asia\_target.csv是目标节点，有7624个数据项，编号为0-7623。

而lastfm\_asia\_features.json是节点的特征值，有7624条、3e6余项数据。

随后进行数据清洗与预处理。数据集没有检查出缺失项和明显的错误项。随后我们根据边的入点顺序对edges文件进行排序如下，如下左图。此外，我们将Json文件的格式进行调整，使其更易于观察和处理，如下右图。



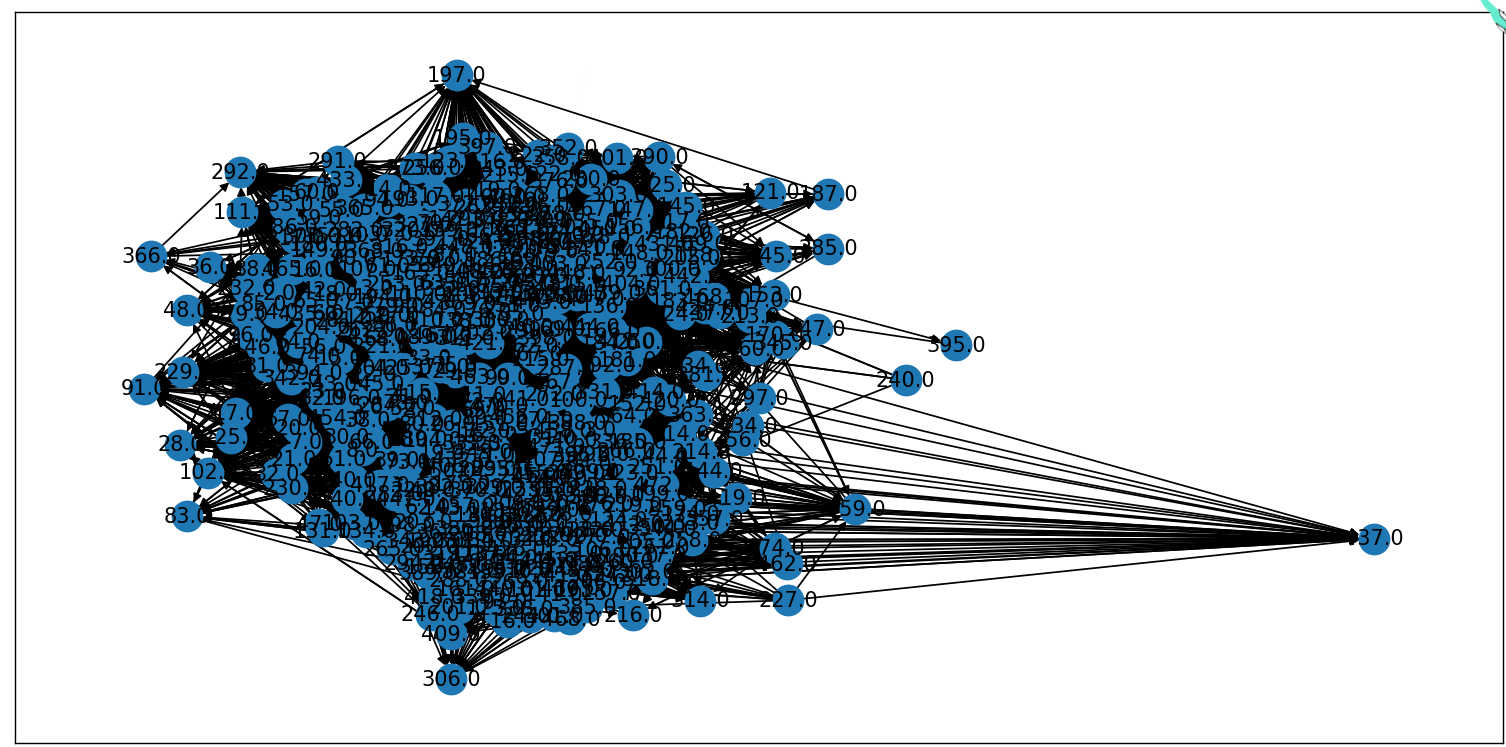
# 数据集频繁模式挖掘及结果分析

## 3.1 数据集Twitter Interaction Network for the US Congress——子图挖掘

**· 模式挖掘和可视化展示**

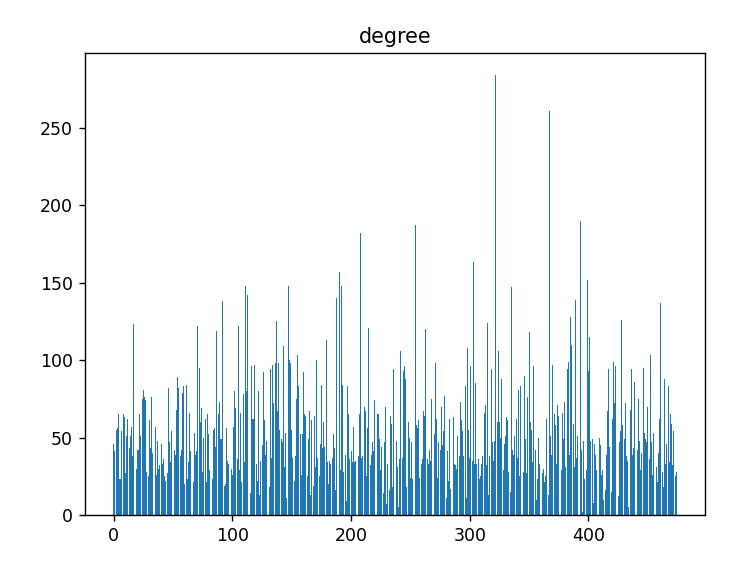
我们使用NetworkX对数据进行挖掘。我们先使用经过处理后的edgelist文件，通过边点信息进行建图操作。

建图的可视化展示结果如下：



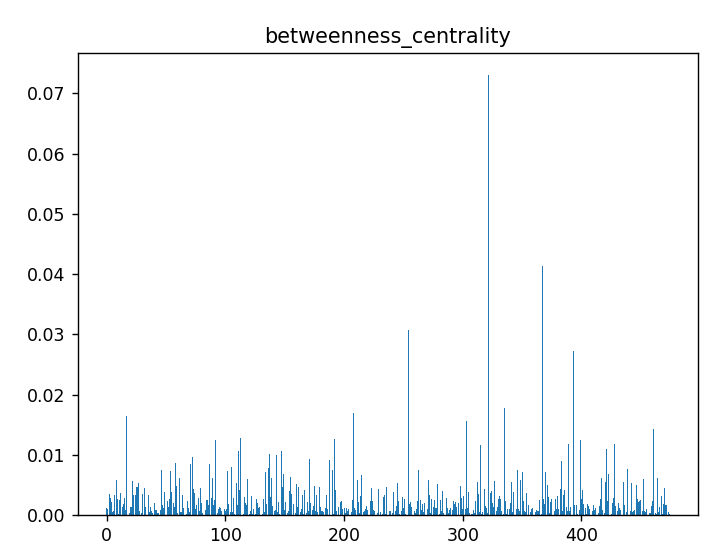
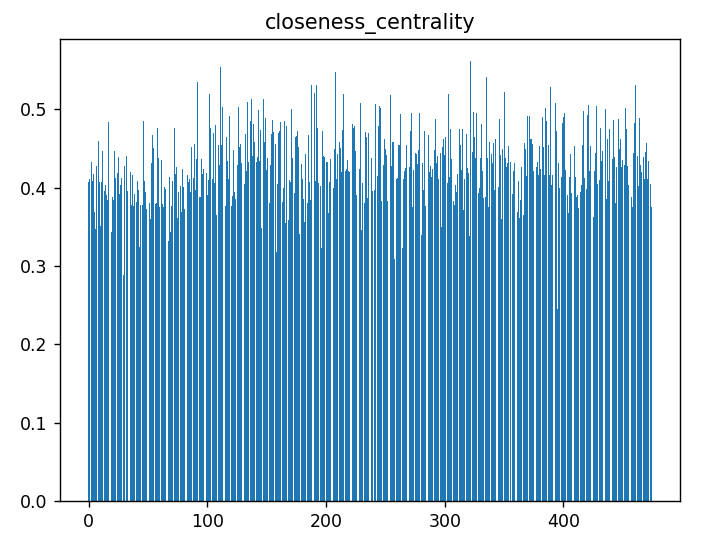
可见，475个节点的连接较为紧密，密度达到了11.75%。

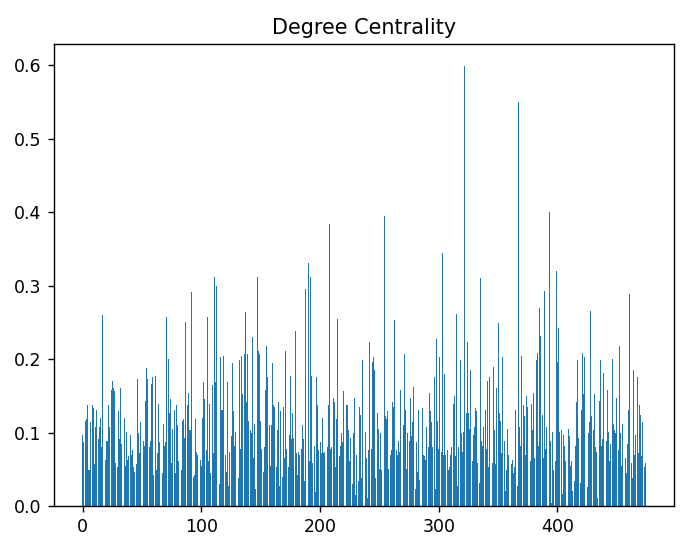
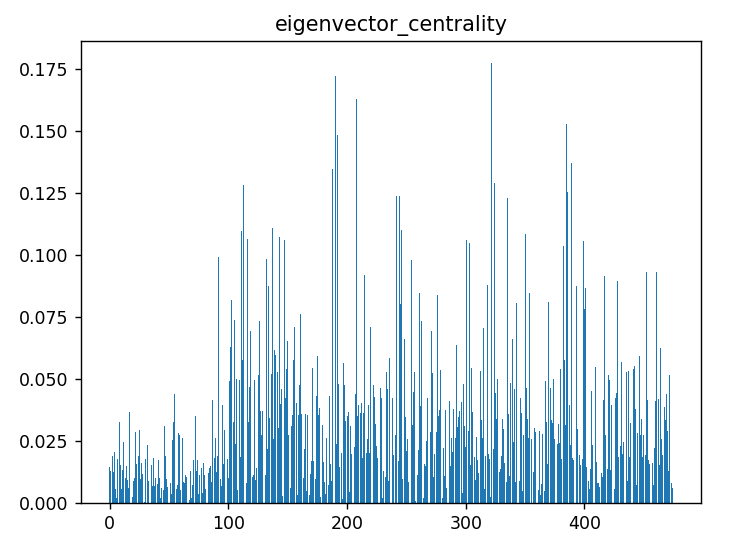
随后，对其求连通度如下：



可见，在大多数数据基本服从均值在45左右的分布，有偏少量样本较为突出，连通度明显高于其它的点。

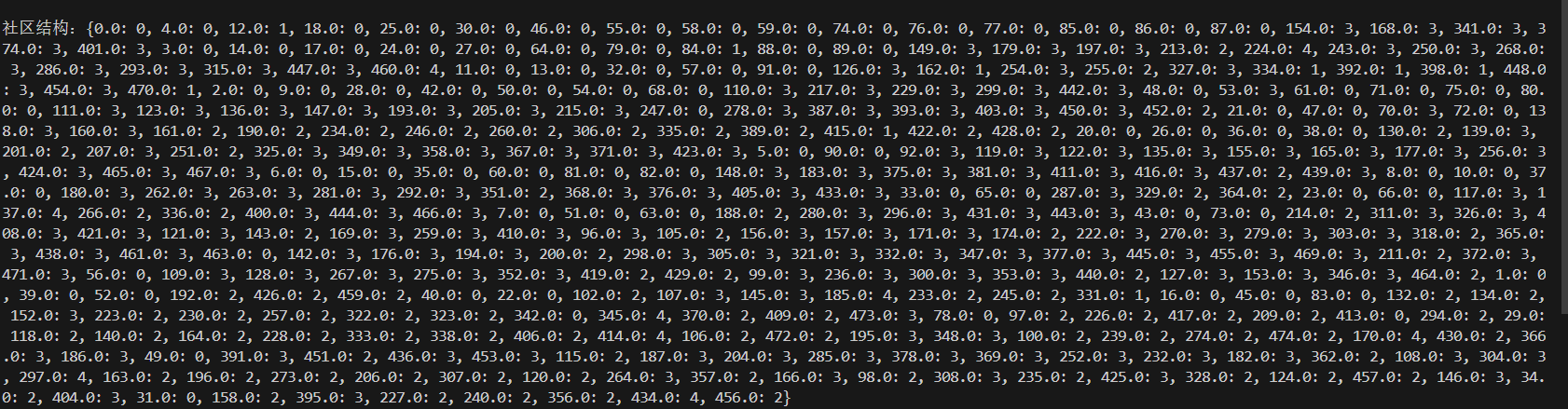
随后，对其进行中心性挖掘，分为四个条目：中心性、邻接中心性、度中心性、优先中心性四种。它们的结果分别如下：（左上、左下、右上、右下）。

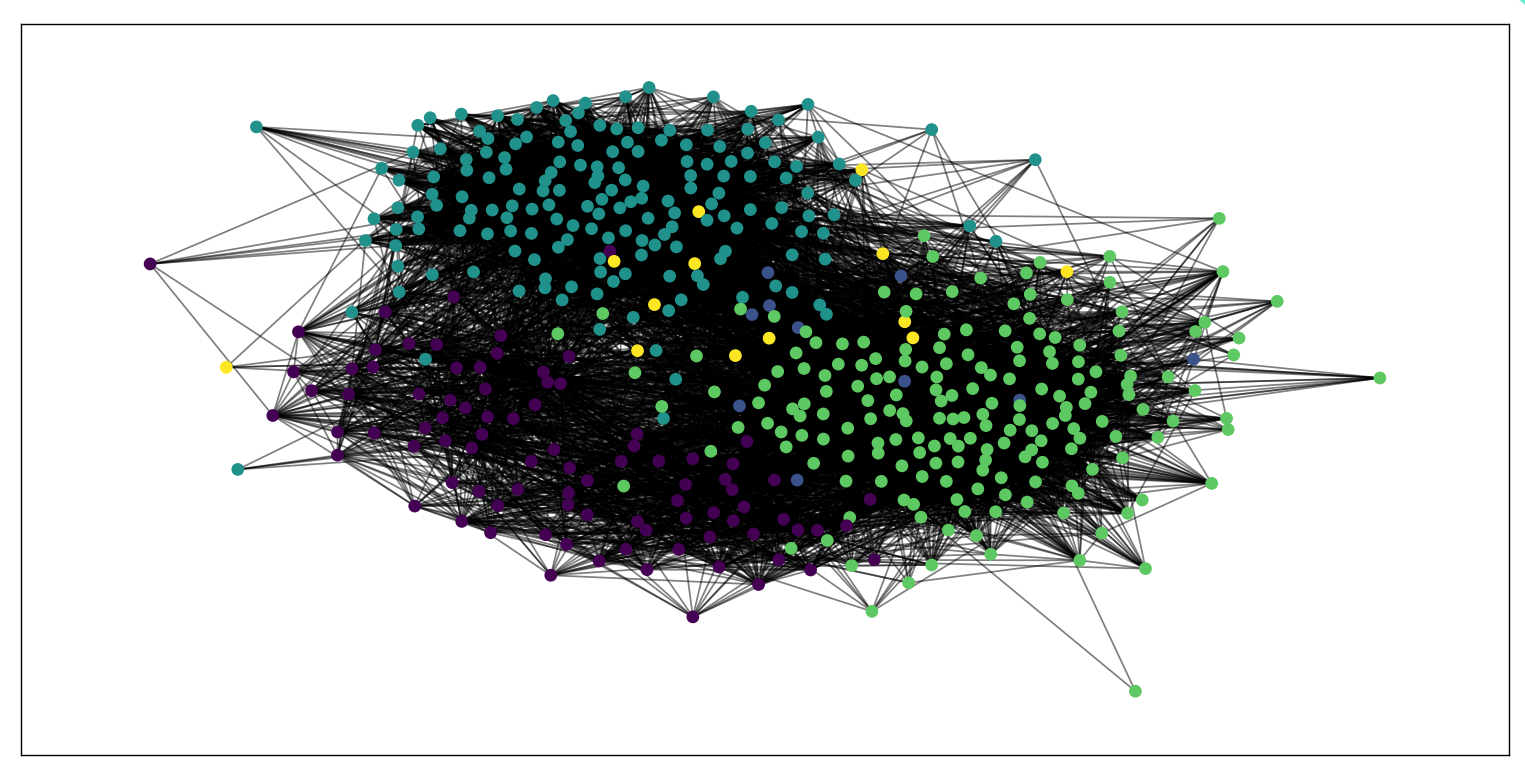
 

可见，中心性将连接饱和度突出的用户进一步地突出，形成了低均值+极高异常值相组合的数据结构，而邻接中心性则反之。度中心性基于度计算，而优先中心性则利好顺序（优先级）较为靠先的点。

随后，我们将原图转换为**无向图，使用Louvain算法**对上述图数据集做社区发现和分析，结果如下：



上述数据的可视化结果如下，图中不同颜色的区域在可视化中代表了不同的聚类类别。



* **挖掘结果分析和模式命名**

通过以上的实验结果可以看出，在度检测和中心性检测中，数据的分布都是由大量低方差平稳数据和少量远高于一般水平的值构成的，且这些值在不同的分析标准下高度相关。**我们命名这种模式为“高流量用户”，**代表其在整个社交平台上较高的受关注度和参与度。

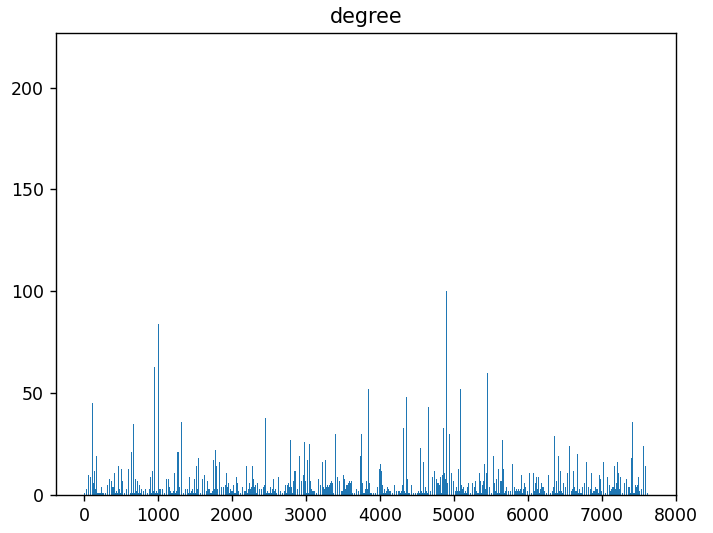
此外，挖掘到的聚类社群中，类别之间的关联度低，而内部关联度高。**我们将这种聚类模式命名为“社交小团体”**，代表社交平台上不同人群关系网之中相互的局部亲近关系。

## 数据集LastFM Asia Social Networks——子图挖掘

**· 模式挖掘和可视化展示**

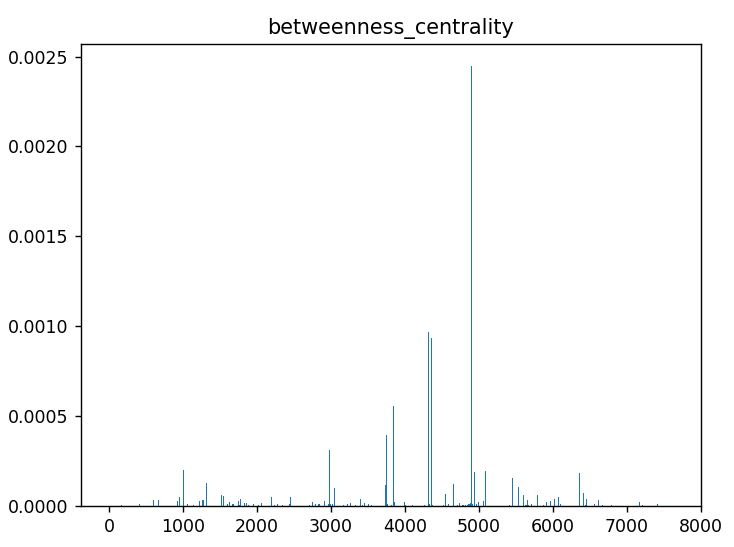
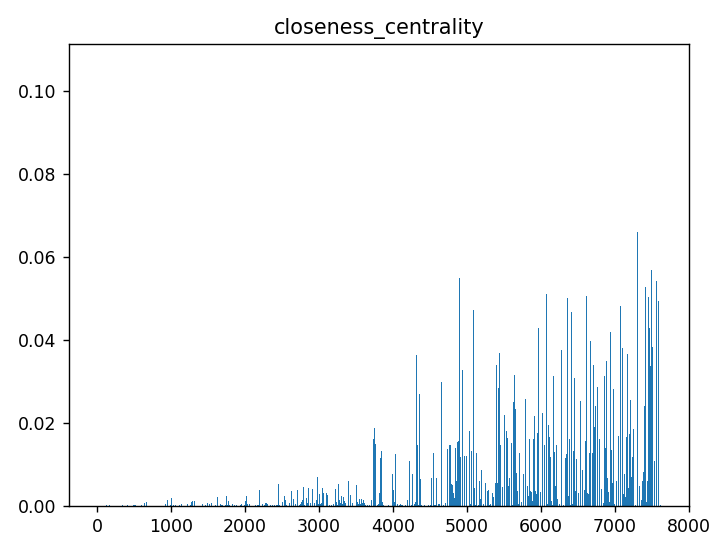
同样，我们先使用经过处理后的edgelist文件，通过边和点进行建图操作。建图结果过大，共7624个结点和27806条边，此处暂不做可视化处理。

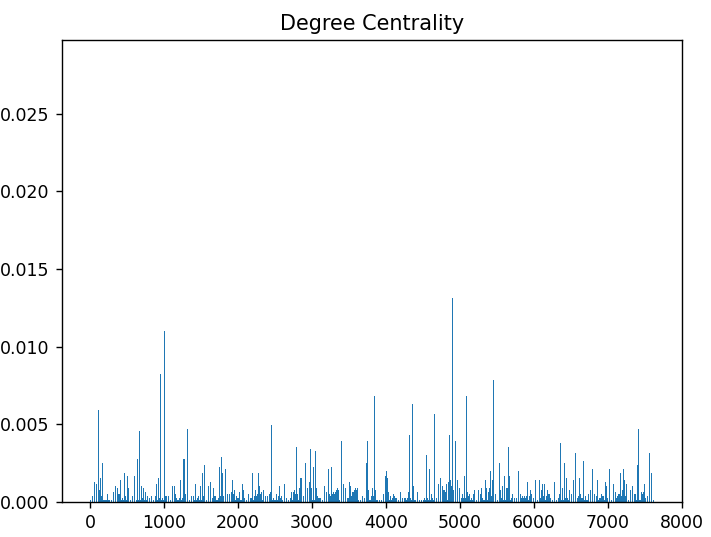
随后，对其求连通度如下：



可见，作为同类的有向图数据集。该数据集与上一个数据集相比**平均连接饱和度低很多**。在大多数数据基本服从均值很低的分布，有少量样本连通度明显高于其它的点，较为突出。

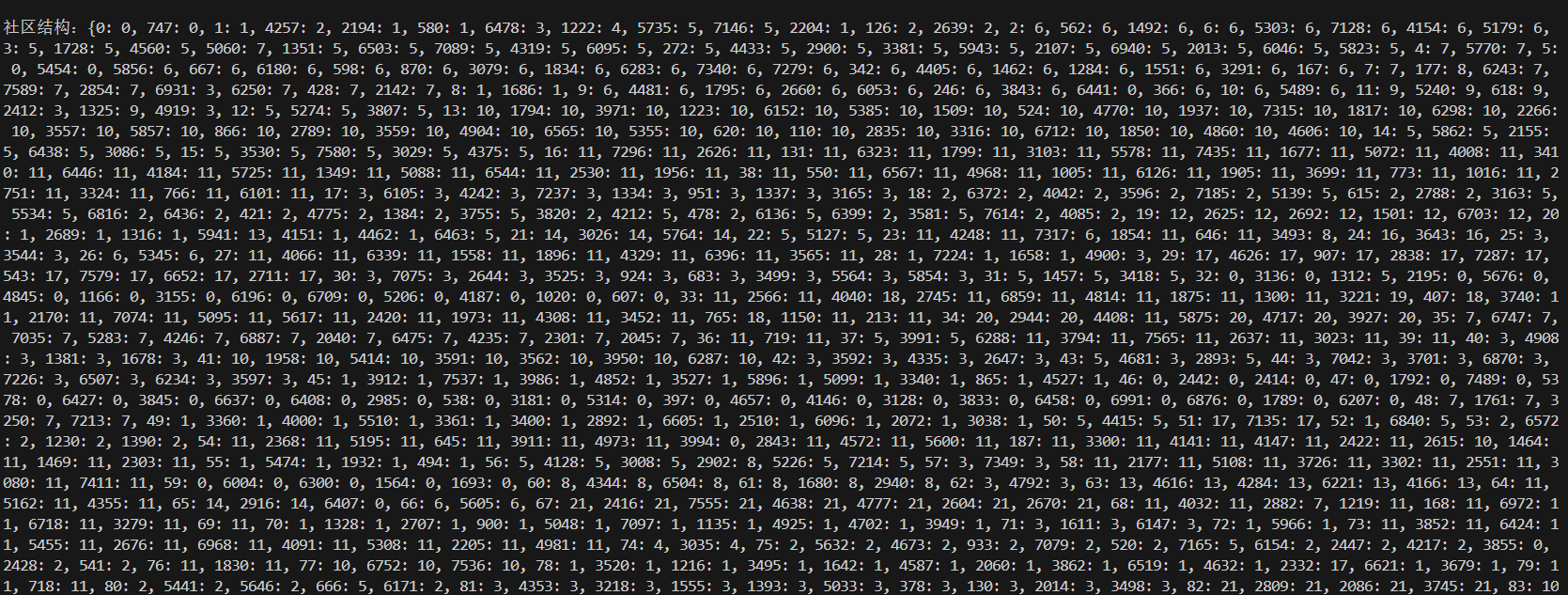
随后，对其进行中心性挖掘，分为三个条目：中心性、邻接中心性、度中心性。它们的结果分别如下：（左上、右上、右下）。由于优先级中心度的计算代价过高、难以收敛，此处删去。

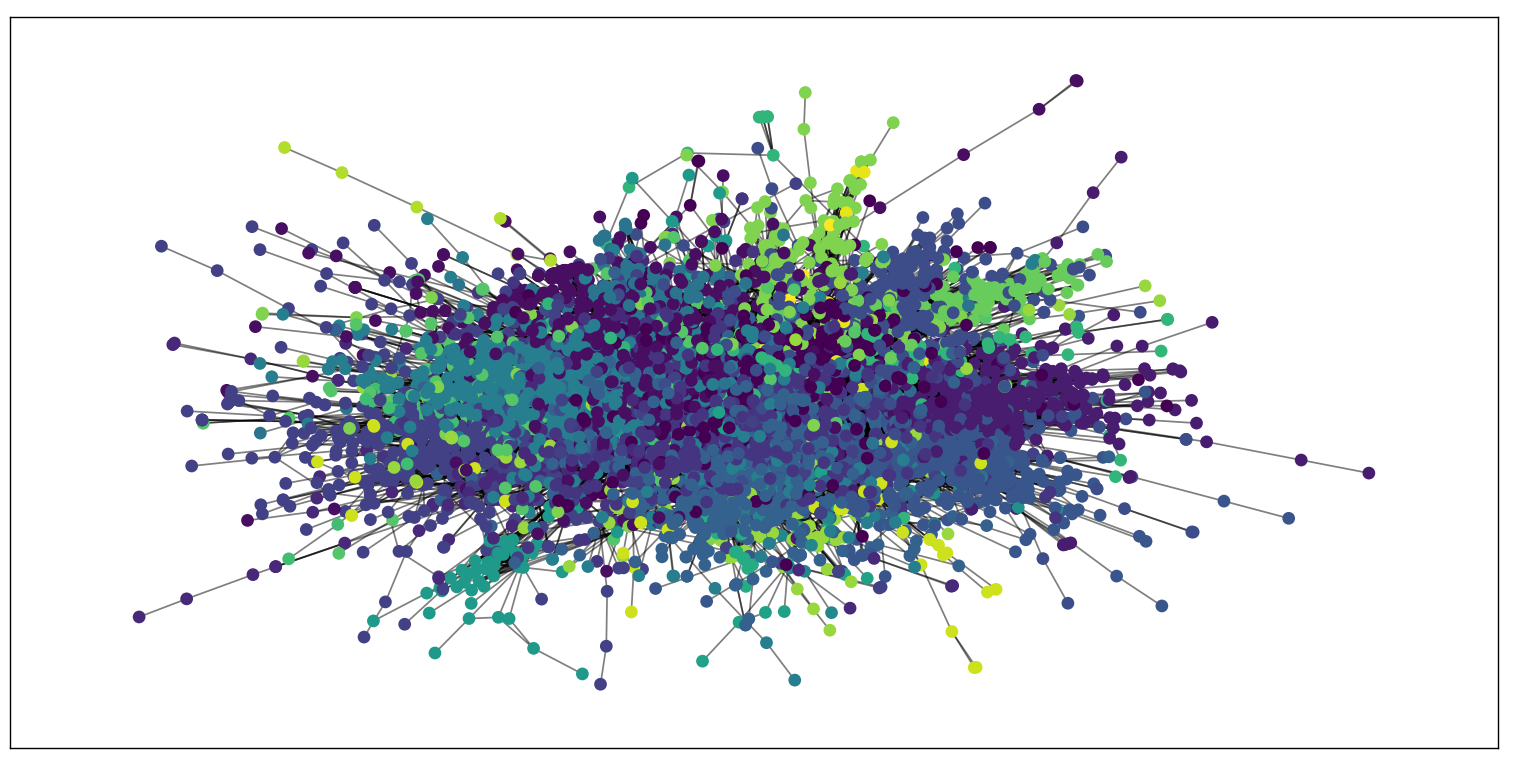


与前一个数据集的分布情况相似，中心性将连接饱和度突出的用户进一步地突出，形成了低均值+极高异常值相组合的数据结构，而邻接中心性则体现除了一定的**右偏高特性**。度中心性基于度计算。

随后，我们同样将原图转换为**无向图，使用Louvain算法**对上述图数据集做社区发现和分析，结果如下：

****

可视化结果如下：



可见。由于点的度整体更低，这一数据库的团体相比于前一个数据库更为松散鲜明。

* **挖掘结果分析和模式命名**

通过以上的实验结果可以看出，在度检测和中心性检测中，数据的分布都是由大量低方差平稳数据和少量远高于一般水平的值构成的，且这些值在不同的分析标准下高度相关。**我们依然命名这种模式为“高流量用户”，**代表其在整个社交平台上较高的受关注度和参与度。

此外，挖掘到的聚类社群中，类别之间的关联度低，而内部关联度高。**我们将这种聚类模式命名为“社交小团体”**，代表社交平台上不同人群关系网之中相互的局部亲近关系。

此外，我们可以发现，在**有类似的点边关系的、抽样得到的图数据中，取样范围越大，则平均度和连接饱和性就会越小**（比如这一数据集抽样范围的亚洲相比于上一数据集的国会大，则关系更为松散，平均度也就越低），**我们将这种模式称为“稀释效应”**。

# 实验总结

本实验的数据集大多偏向于图结构数据，因此挖掘方法与图的操作相关性较强。得益于数据集较为良好的完整性，本实验的子图挖掘得以通过较为简捷的方式执行。

我们的数据挖掘主要有度分析、中心性分析和社区发现-聚类分析三大模块，将挖掘的目标主要集中在图本身的特性上。