北京大学

数据可视化 个人作业三

色彩与网络数据可视化

姓名： 吴凡

授课教师： 步一 助理教授

二〇二三 年 八 月

一.数据简介

本次报告选用 Stack Overflow Tag Network 进行分析，下面对该数据集进行简要介绍：

（1）构建方式：Stackoverflow 是全球最大的 IT 领域的问答类网站，其上面的问题可以很好地反映 IT 从业人员的“技术栈”。“Stack Overflow Tag Network”数据集是基于 Stack Overflow 网站开发者对工作、职业描述或个人简历描述的相关信息进行标记，他们会在这些描述中提到自己使用过的技术（包括编程语言、数据库、Web 开发、操作系统等），这些标签最终构成了 Stack Overflow Tag Network 数据集。为了探究对技术如何连接以及它们如何一起使用，整理得到两个数据集，分别刻画节点（"Stack_network_nodes"）和边（"Stack_network_links"）的属性。

（2）指标说明：“Stack Overflow Tag Network”数据的节点代表广泛的技术领域，例如编程语言（如 Java、Python、C# 等）、数据库（如 MySQL、SQL Serve、Oracle 等）、Web 开发（如 HTML、CSS、Java 等）、操作系统（如 Windows、Linux、macOS 等）等方面，共 115 个。边表示两节点（技术领域）之间被同一网站开发者者掌握或使用，共 245 条；边的权重代表同时掌握这两种技术或环境的人数。

数据来源于 Kaggle

（https://www.kaggle.com/datasets/stackoverflow/stack-overflow-tag-network）

二.网络分析

（1）直观信息

该网络为无向图，共有 115 个结点和 245 条边。

（2）模块化分析

首先对数据进行模块化分析，找寻是否存在社区：

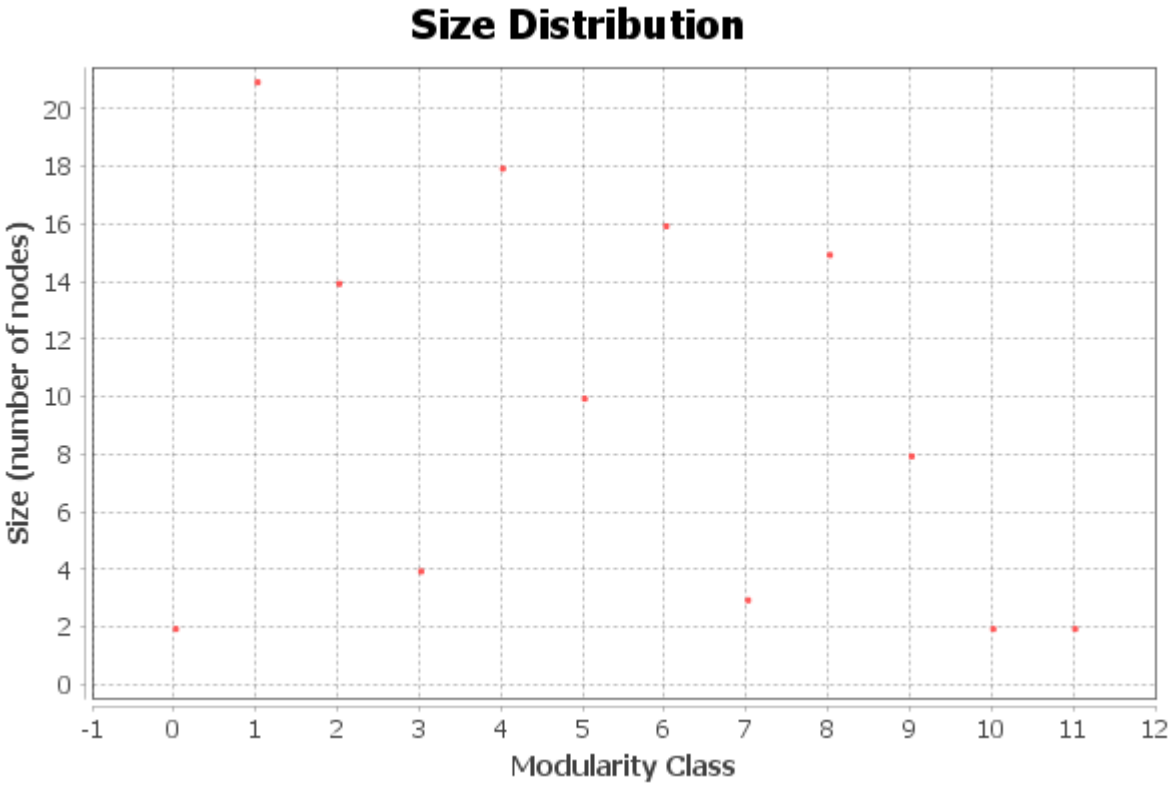
Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0.730
Modularity with resolution: 0.730
Number of Communities: 12



Algorithm:

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, *Fast unfolding of communities in large networks*, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

Resolution:

R. Lambiotte, J.-C. Delvenne, M. Barahona *Laplacian Dynamics and Multiscale Modular Structure in Networks 2009*

可见数据模块化值为 0.73，模块化很强，共可以分为 12 社区，其中 Class1 的 size 最大，

（3）PageRank

对数据进行 PageRank 算法，寻找一下哪个结点最重要吧，在 gephi 的数据资料中依据 PageRank 对数据进行降序排序：

Label	PageRank
linux	0.025135
jquery	0.022375
c#	0.02236
css	0.020023
angularjs	0.019842
asp.net	0.019005
python	0.017821
hibernate	0.016779
spring	0.016779
mysql	0.016622
javascript	0.016165
java	0.016076
php	0.014147
html5	0.014114

可见 linux、jquery、c#分别位于前三的位置，是当前 IT 行业最重要的三种技术。

（4）平均聚类系数

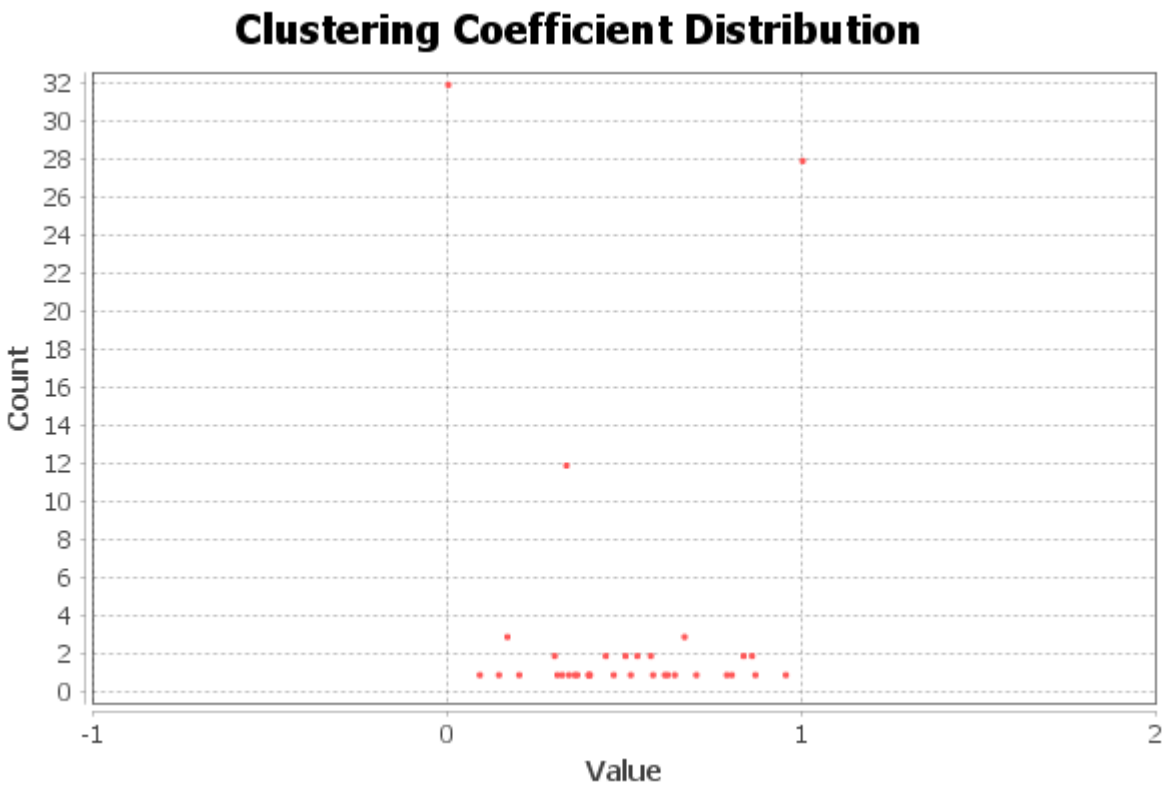
Clustering Coefficient Metric Report

Parameters:

Network Interpretation: undirected

Results:

Average Clustering Coefficient: 0.629
Total triangles: 238
The Average Clustering Coefficient is the mean value of individual coefficients.



Algorithm:

Matthieu Latapy, *Main-memory Triangle Computations for Very Large (Sparse (Power-Law)) Graphs*, in Theoretical Computer Science (TCS) 407 (1-3), pages 458-473, 2008

以上为平均聚类系数的报告，聚类系数为 0.629，说明各类技术存在一定的聚类特征。

（5）特征向量中心度

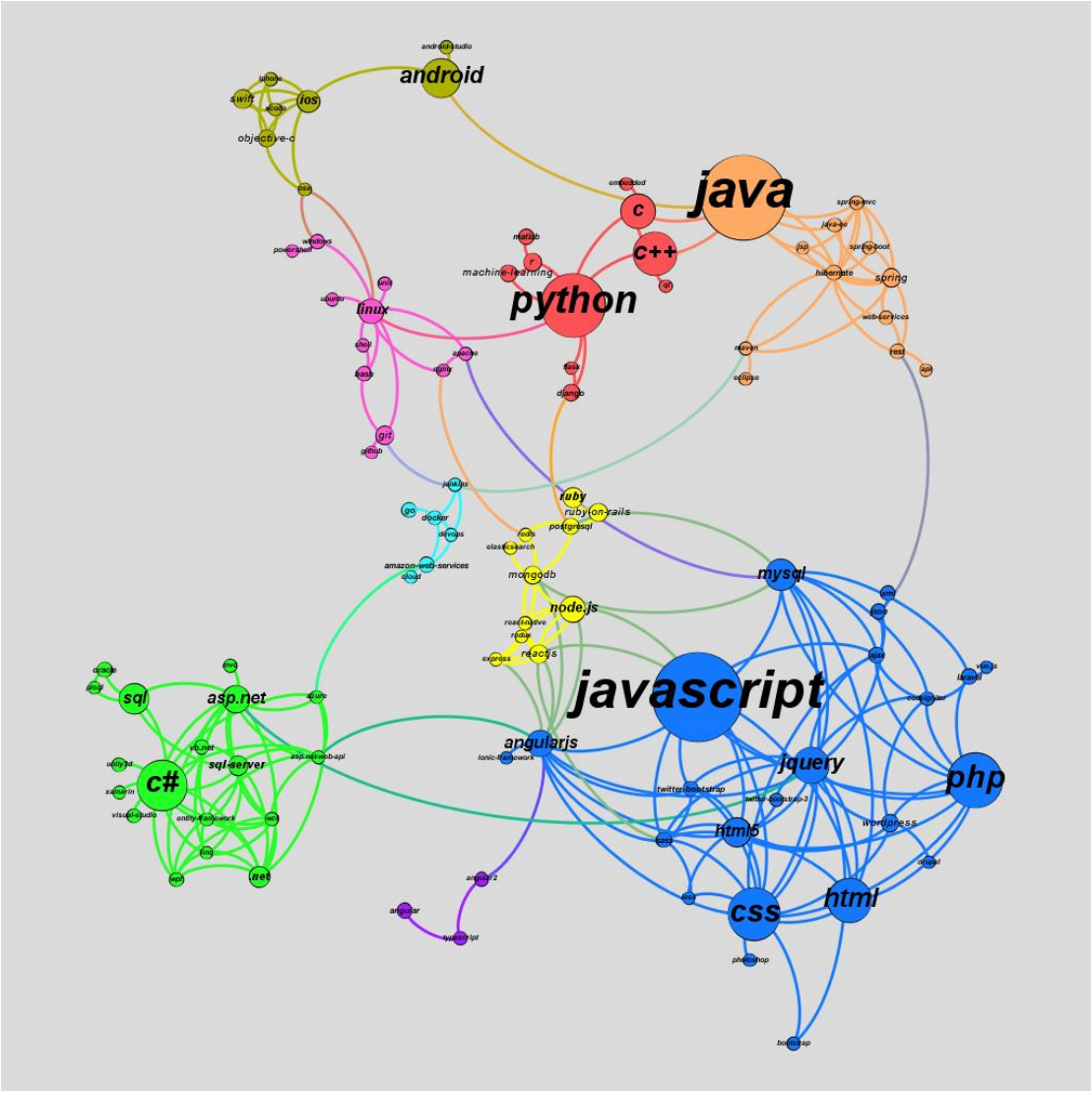
Label	Eigenvector Centrality
jquery	1.0
css	0.890086
javascript	0.867403
angularjs	0.747208
html5	0.704197
php	0.695593
sass	0.667624
mysql	0.646949

以上为对数据进行特征中心度分析后，将数据按照特征中心度进行降序排序的前八行数据，以此来寻找处于中心的最重要的结点。

可见特征向量中心度与 PageRank 存在一定关联，可以看出 jquery

和 css 两门技术在 IT 行业现在应用非常广泛。

三.可视化结果



该图经过依据 nodesize 太小的值过滤后形成了九个 group，颜色选取的 RGB 值如下:

1 (255，82，87)

2 (36，255，38)

3 (250，255，5)

4 (175，179，0)

5 (255，87，213)

6 (20，121，255)

7 (163，31，255)

8 (255，171，102)

9 (51，249，255)

分别用以上的九种颜色对分属于不同 group 的结点进行着色。

结点的大小与文件中的 nodesize 成比例，以此来表示每项技术的 Tag 在 Stackoverflow 网站中出现的频率。

边的颜色采用对所连接的两个节点的颜色进行混合所得到的颜色。

边的粗细用与数据集中的 value 成比例，以此来表现两种 Tag 同时出现的频率的多少。

图的布局采用 ForceAtlas 2 布局。

四.结果的解读与讨论

通过网络图我们可以看到几个明显的特征：

1. 不同颜色的结点呈现了明显的聚类特征，说明 IT 行业中的技术大都是集中出现，也可以看出哪一些技术集成开发的频率较高：例如前端技术的代表技术 Javascript，与其联系紧密的有我们熟知的 CSS、html、html5、jquery 等，也有与后端技术的联系，例如数据库语言 mysql 和实现动态网页的语言 ajax，还有存储数据的格式 json 和 xml。
2. 我们可以看到图中几个中心性较强的部分：Javascript、Java、Python、C#、android、linux，这些都是在 IT 行业应用较广的技

术；也可以看到有一些熟知的结点较小，如和 C++连接的 qt，现在已经基本上被市场淘汰了。