

25

Social Network Analysis

社交网络分析

度分析、图距离、中心性、社区结构



随大流的人总是亦步亦趋；孤勇者则才可能开天辟地。

The one who follows the crowd will usually go no further than the crowd. The one who walks alone is likely to find themselves in places no one had ever been.

—— 阿尔伯特·爱因斯坦 (Albert Einstein) | 理论物理学家 | 1879 ~ 1955



```

networkx.algorithms.community centrality.girvan_newman() Girvan-Newman 算法划分社区
networkx.betweenness Centrality() 计算介数中心性
networkx.bridges() 生成图中所有桥的迭代器
networkx.center() 找出图的中心节点，即离心率等于图半径的所有节点
networkx.closeness Centrality() 计算紧密中心性
networkx.connected_components() 计算图中连通分量
networkx.degree Centrality() 计算度中心性
networkx.diameter() 计算图的直径，即图中所有节点离心率的最大值
networkx.eccentricity() 计算图中每个节点的离心率，即该节点图距离的最大值
networkx.eigenvector Centrality() 计算特征向量中心性
networkx.has_bridges() 检查图中是否存在桥
networkx.is_connected() 判断一个图是否连通
networkx.local_bridges() 生成图中所有局部桥的迭代器
networkx.periphery() 找出图的边缘节点，即离心率等于图直径的所有节点
networkx.radius() 计算图的半径，即图中所有节点的离心率的最小值
networkx.shortest_path() 寻找两个节点之间的最短路径
networkx.shortest_path_length() 计算在图中两个节点之间的最短路径的长度
numpy.tril() 生成一个数组的下三角矩阵，其余部分填充为零
numpy.tril_indices() 返回一个数组下三角矩阵的索引
numpy.unique() 找出数组中所有唯一值并返回已排序的结果

```

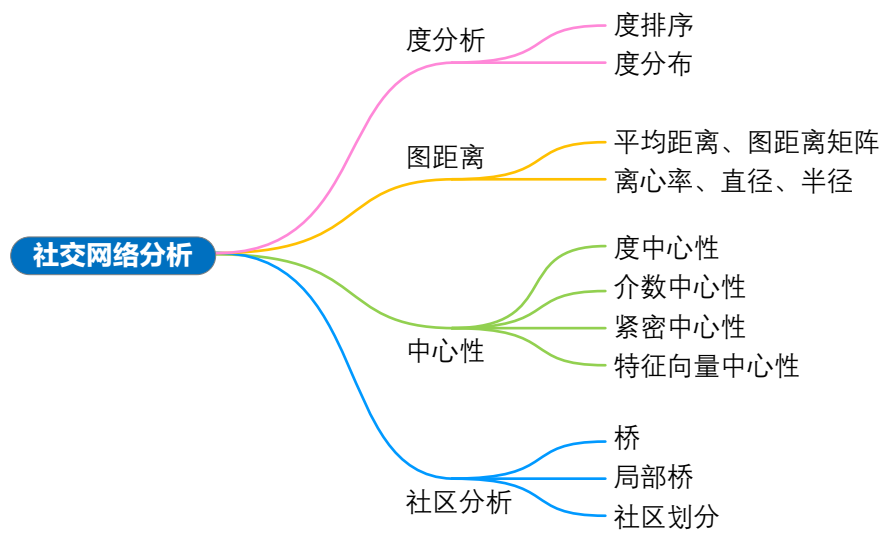
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

25.1 社交网络分析

社交网络分析 (Social Network Analysis, SNA) 是一种研究社交关系和网络结构的方法。它主要关注个体 (如人、组织或概念) 之间的关系, 以及这些关系如何形成和影响整个网络。社交网络分析可以帮助揭示社会结构、信息流动和影响力等方面的模式, 对于理解群体行为、组织结构以及网络中个体之间的互动关系具有重要价值。

本章分析对象是图 1 所示的社交网络。图 1 所示的这幅图有 4039 个节点, 每个节点相当于一个用户; 图中有 88234 条边, 每条边相当于一个好友关系。



图 1. 社交网络图

排版时, 请替换为矢量图, 见附件 SVG 文件

常见的社交网络分析手段包括:

- ▶ **度分析** (degree analysis)。简单来说, 节点的度越高, 表示该节点在网络中有更多的连接。高度中心的节点通常在信息传播和影响力方面更为重要。
- ▶ **图距离** (graph distance)。图距离是图论中衡量两个节点之间最短路径的长度。在社交网络分析中, 图距离用于量化用户之间的联系紧密程度, 识别社区结构, 发现关键用户。

本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com

- **中心性分析** (centrality measure)。中心性有很多度量，如度中心性、介数中心性、紧密中心性、特征向量中心性。这些指标帮助确定节点在网络中的重要性程度，考虑了节点在路径、距离或整体网络结构上的贡献。
- **社区结构分析** (community detection)。通过识别网络中密切连接的子群，揭示了网络中存在的群体结构。社区结构分析有助于理解网络中的功能集群，从而更好地理解组织或社会的内部组织和关系。

大家可能已经发现，本章相当于本书图论主要内容的一个应用案例。

本例参考 NetworkX 官方示例，链接如下：

https://networkx.org/nx-guides/content/exploratory_notebooks/facebook_notebook.html

数据来自 Stanford，链接如下：

<https://snap.stanford.edu/data/ego-Facebook.html>

25.2 度分析

度是社交网络分析中的一项基本指标，用于衡量节点在网络中的连接程度。

简单来说，对于无向图来说，节点的度就是该节点连接的边数，反映了节点在网络中的直接关联程度。节点的度越高，表示其在网络中的联系越多。对于有向网络，节点的度分为入度和出度。入度是指指向该节点的连接数量，出度是指由该节点指向其他节点的连接数量。通过入度和出度的分析，可以揭示节点在信息传播和影响方面的不同角色。

度排序 (degree ranking) 对网络中的节点按照度的大小进行排序。这可以帮助识别网络中的重要节点，即那些连接较多的节点。排序后，可以更清晰地看到网络中的核心成员。图 2 (a) 所示为图 1 社交网络的度排序。

度分布图 (degree bar chart/histogram chart) 可视化网络度分布。横轴表示度的取值，纵轴表示具有相应度的节点数量。通过度直方图，可以观察网络中节点度的分布情况，是一个快速了解网络结构的工具。图 2 (b) 所示为图 1 社交网络的度柱状图。

图 3 根据节点度数渲染节点；暖色节点度数较高，冷色节点度数低。图 4 中红色节点的度数超过 100。度中心较大的节点在信息传播和网络连接方面通常更为重要。通过度分析，可以识别出在网络中具有重要地位的节点，这对于优化信息流、识别关键人物等方面非常有帮助。

度分析可以帮助了解网络的整体结构，尤其是哪些节点在网络中起到连接的纽带作用。这有助于理解网络的稳定性和韧性。异常高或低度的节点可能是网络中的异常点。检测这些异常节点可以帮助发现网络中的潜在问题或重要事件。

在社交网络中，通过度分析可以识别出具有相似连接模式的节点，从而帮助发现网络中的社区结构。总体而言，度分析为研究网络结构、识别关键节点、理解信息传播和预测网络行为提供了基础，并通过可视化工具如度直方图帮助研究者更好地理解网络中节点的分布和连接模式。

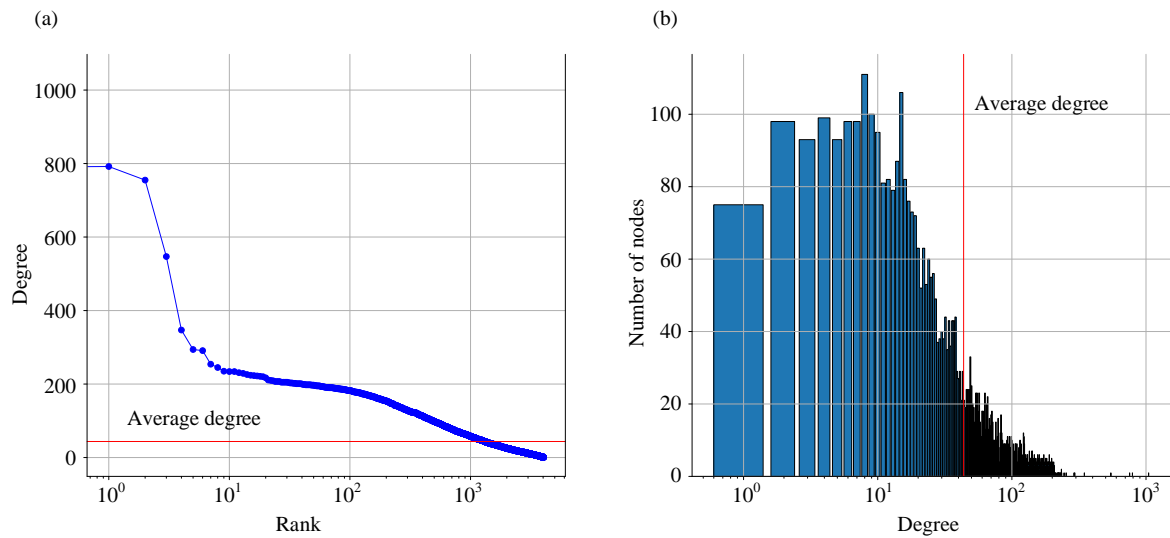


图 2. 度分析，节点度数排序、节点度数柱状图

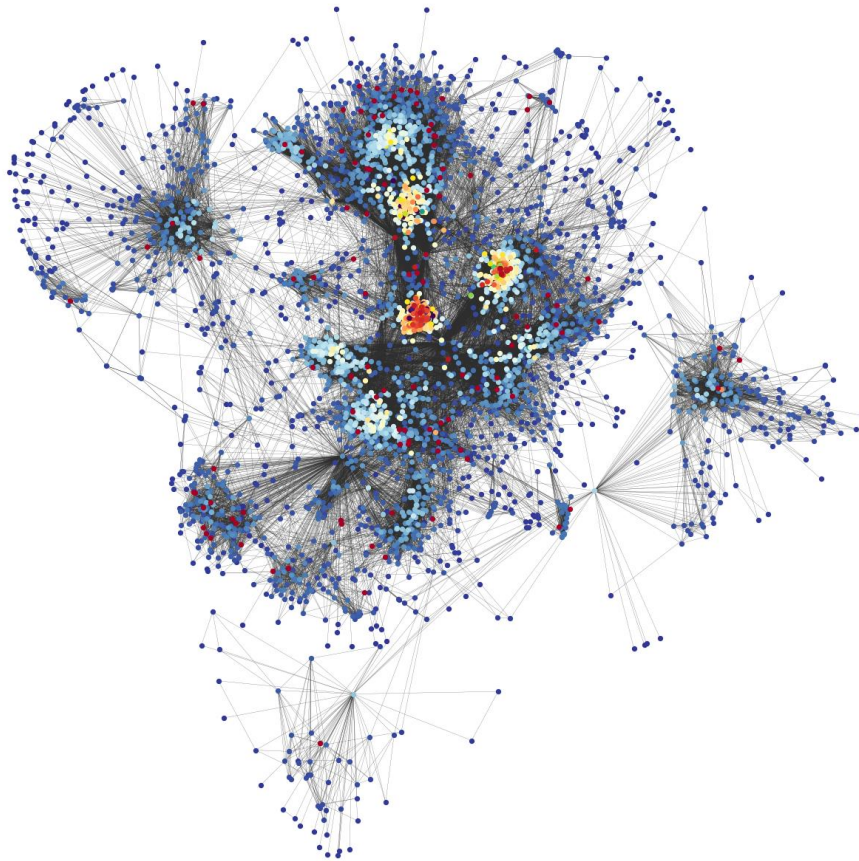


图 3. 社交网络图，节点度数

排版时，请替换为矢量图，见附件 SVG 文件

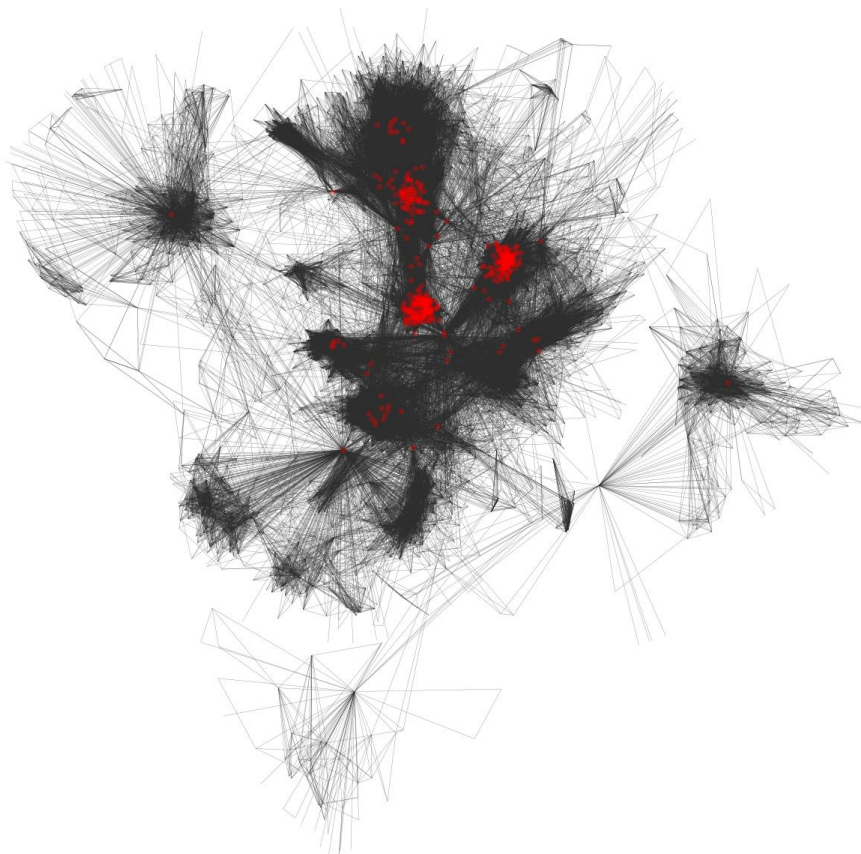


图 4. 社交网络图，节点度数超过 100 的节点

排版时，请替换为矢量图，见附件 SVG 文件

25.3 图距离

图距离指的是图中任意两个节点之间的最短路径长度，图距离直方图展示了图中所有节点对的图距离分布，有助于理解网络的连接紧密程度。图 5 所示为社交网络成对图距离的柱状图。

平均图距离是某个节点到图中其他所有节点图距离的平均值；图 6 所示为社交网络所有节点平均图距离直方图。图 6 中红色划线则是图距离平均值的均值，反映了网络中成员间平均分隔的“远近”，是衡量网络紧密程度的一种指标。图 7 则是根据节点平均图距离大小渲染节点。

如图 8 所示，图距离矩阵是一个矩阵，其中的元素表示图中任意两个节点之间的距离，提供了网络连接结构的全面视图。

离心率是指图中一个节点到所有其他节点的最短路径中的最大值，它衡量了一个节点在网络中的边缘程度。图 9 所示为社交网络离心率柱状图。

直径是图中所有节点离心率最大值，显示了网络中最远两个节点间的距离。半径是所有节点的离心率的最小值，指出了到达网络中任何节点所需的最短距离。观察图 9 这幅离心率柱状图，我们立刻可以知道社交网络的直径为 8，半径为 4。

在社交网络分析中，上述这些图距离相关概念帮助我们理解和量化网络的结构特征，例如，识别关键个体（如中心节点或边缘节点），理解信息或影响力在网络中的传播速度，以及网络的整体连通性和紧密度。

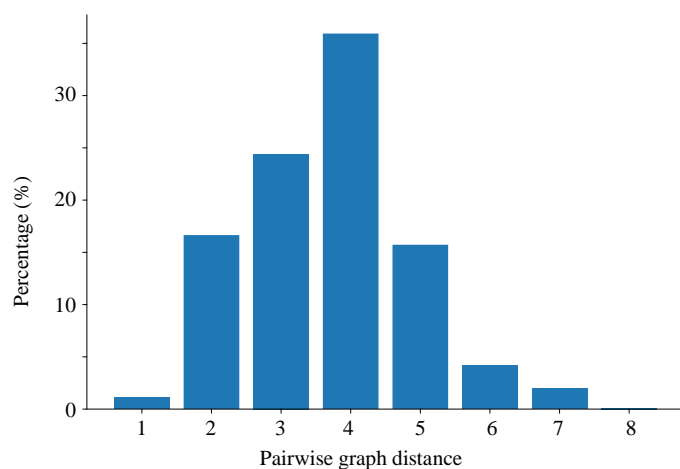


图 5. 图距离柱状图

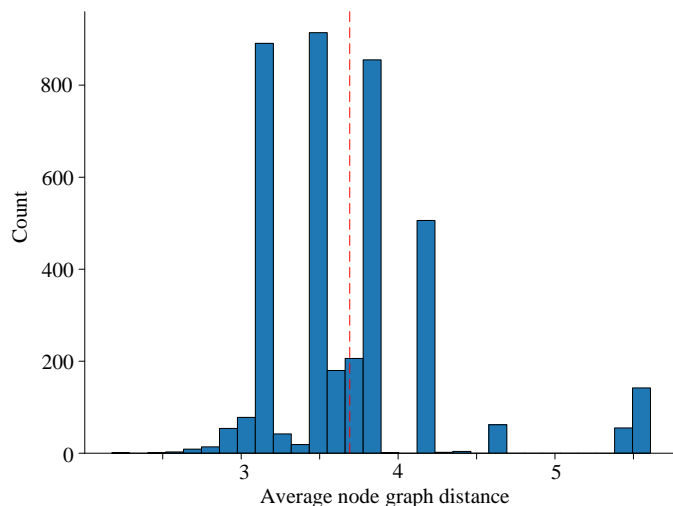


图 6. 平均图距离直方图

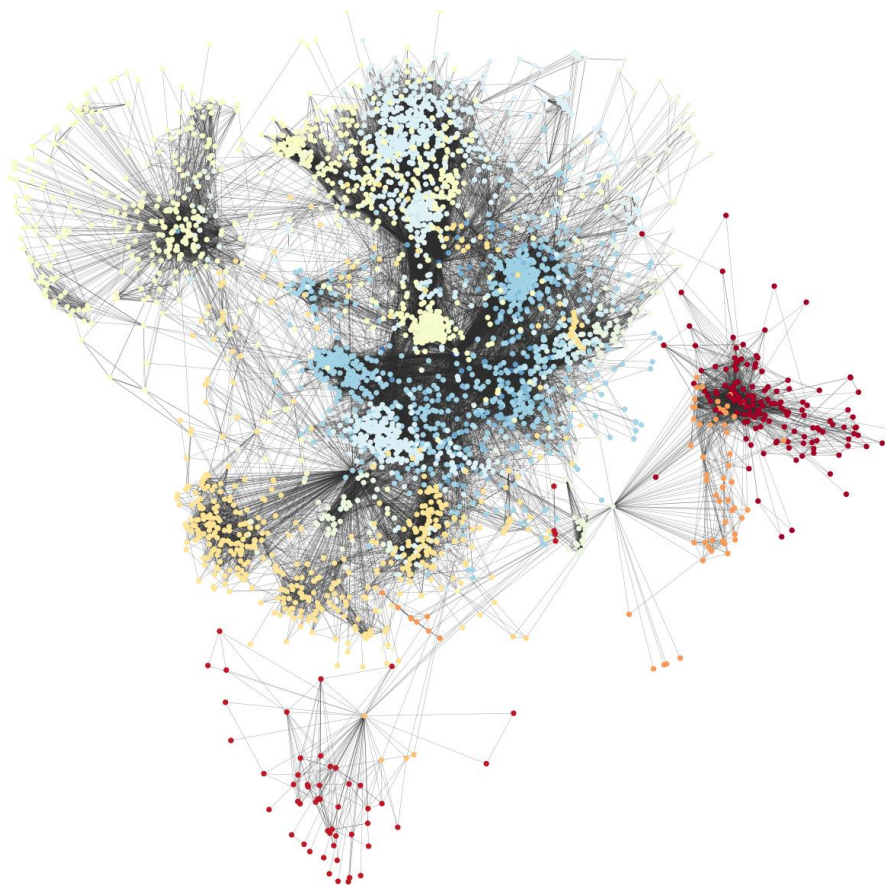


图 7. 社交网络图，平均图距离

排版时，请替换为矢量图，见附件 SVG 文件

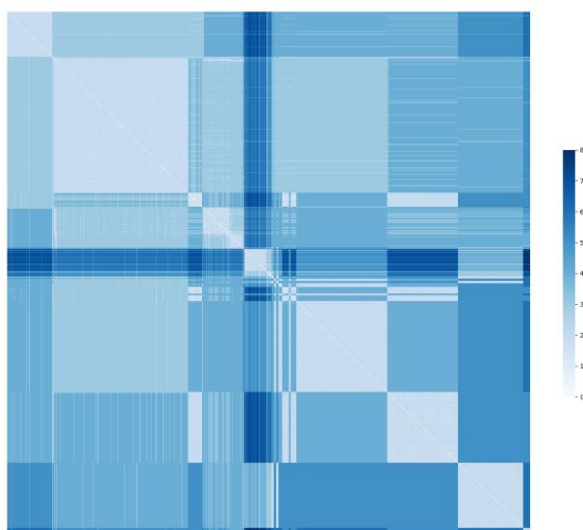


图 8. 图距离矩阵

排版时，请替换为矢量图，见附件 SVG 文件

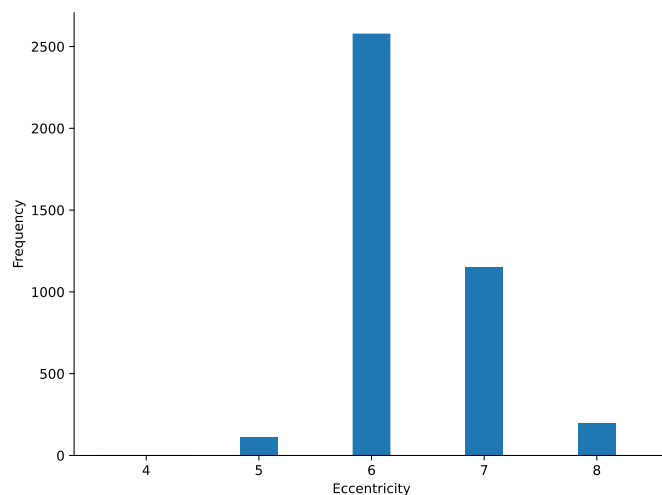


图 9. 离心率柱状图

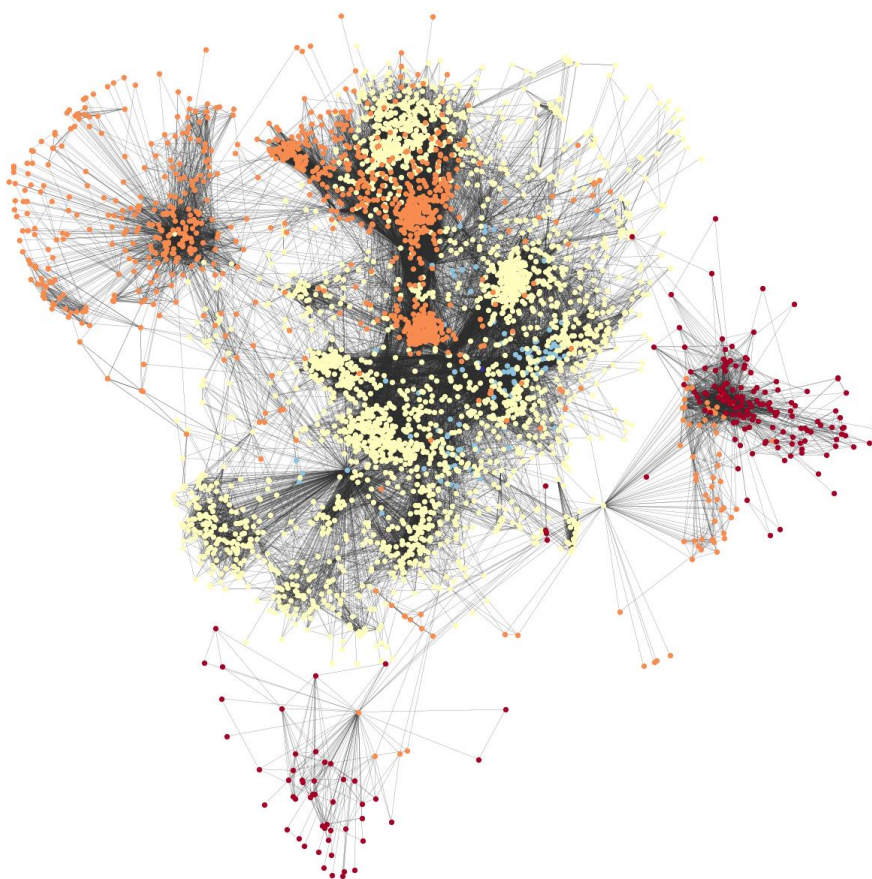


图 10. 社交网络图，离心率

排版时，请替换为矢量图，见附件 SVG 文件

小世界理论 (Small World Theory) 描述了一种网络结构，其中节点之间的平均距离较短，同时节点之间的关系又相对密切。这种网络结构兼具高度集聚的特征和较短的平均图距离，使得网络在信息传播、搜索和传递方面具有高效性。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

小世界理论的关键观点之一是，即使在庞大的网络中，任意两个节点之间的平均最短路径长度也相对较短。这意味着，即使网络规模庞大，节点之间通过较短的路径就能相互连接，使得信息可以快速传播。因此，图距离是小世界网络结构的一个重要特征。

社交网络分析通常关注个体之间的关系及其网络结构。许多社交网络，尤其是在线社交媒体网络，展现出小世界网络的特征。在社交网络中，人们通常能够通过朋友之间的短路径迅速建立联系，形成高效的信息传播通道。小世界网络的特性在社交网络中解释了为什么人们可以通过相对较短的路径找到彼此，或者为什么信息在网络中能够迅速传播开来。

25.4 中心性

中心性是社交网络分析中的一组指标，用于度量节点在网络中的重要性程度。中心性度量的核心思想是通过不同的衡量方式来理解节点在网络中的位置和影响力。

度中心性是最简单和最直观的中心性度量。它衡量了一个节点与其他节点直接连接的数量。节点的度越高，说明其在网络中的直接连接越多，通常被认为在信息传播和影响力方面更为重要。

图 11 所示为利用节点节点度中心性渲染节点；图 12 所示为社交网络节点度中心性直方图。

介数中心性衡量了一个节点在网络中的桥接作用，即节点在不同节点之间的最短路径上的频率。节点的介数中心性越高，表示它在网络中连接其他节点之间的路径上更为频繁，可能在信息传播中扮演关键角色。

图 13 所示为根据节点介数中心性渲染节点；图 14 所示为社交网络节点介数中心性直方图。

紧密中心性衡量了一个节点到其他节点的平均距离。节点的紧密中心性越高，表示它距离其他节点更近，可能更容易接触到网络中的信息和资源。紧密中心性可以帮助识别在网络中能够迅速传播信息的节点。

图 15 所示为根据节点紧密中心性渲染节点；图 16 所示为社交网络节点紧密中心性直方图。

特征向量中心性考虑了一个节点及其直接连接的节点的影响力，即节点与其邻居的中心性。一个节点的特征向量中心性越高，表示它与其他中心性较高的节点有更多的连接。这意味着该节点不仅与许多节点相连，而且这些节点本身也在网络中具有较高的中心性。这种中心性度量有助于识别在网络中具有整体影响力的节点。

图 17 所示为根据节点紧密中心性渲染节点；图 18 所示为社交网络节点紧密中心性直方图。

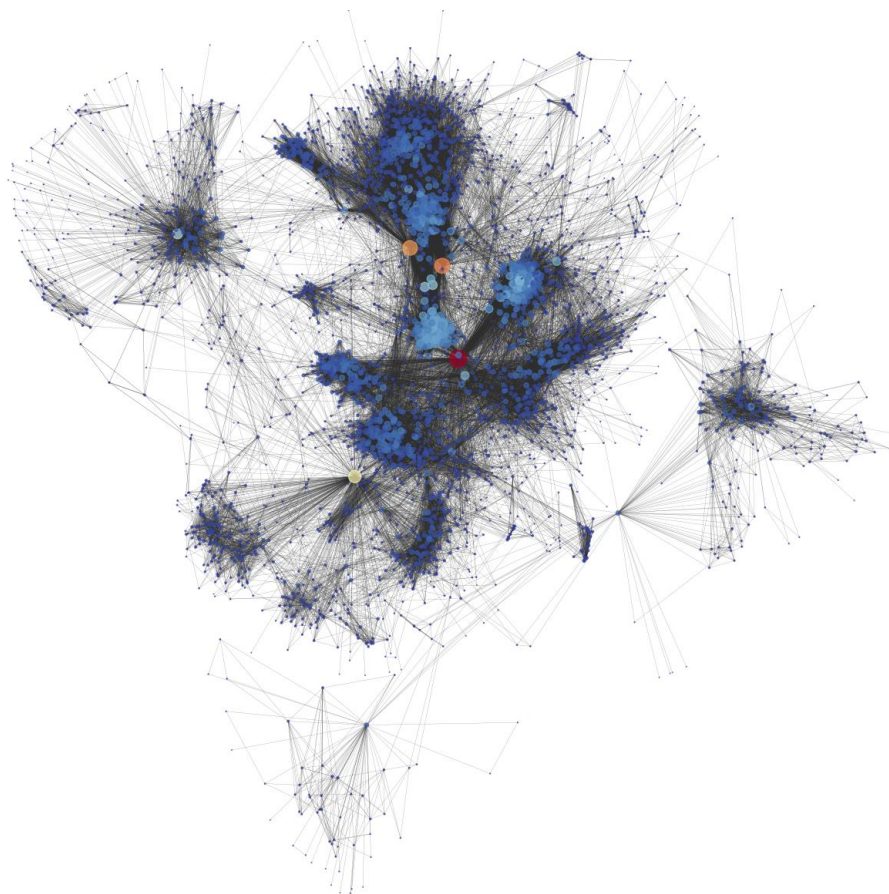


图 11. 社交网络图，度中心性

排版时，请替换为矢量图，见附件 SVG 文件

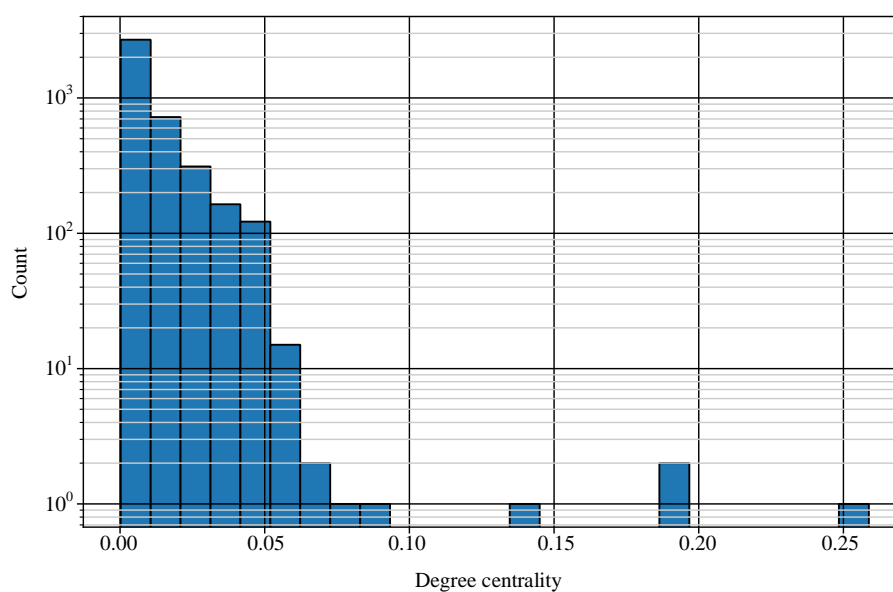


图 12. 度中心性直方图

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

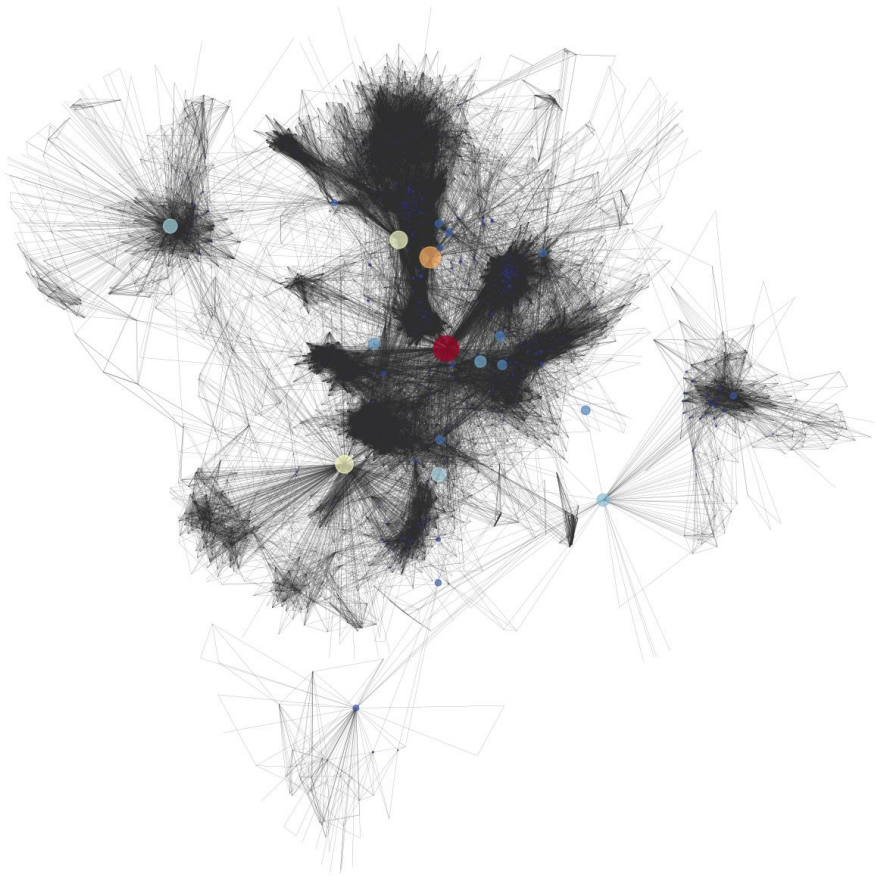


图 13. 社交网络图，介数中心性

排版时，请替换为矢量图，见附件 SVG 文件

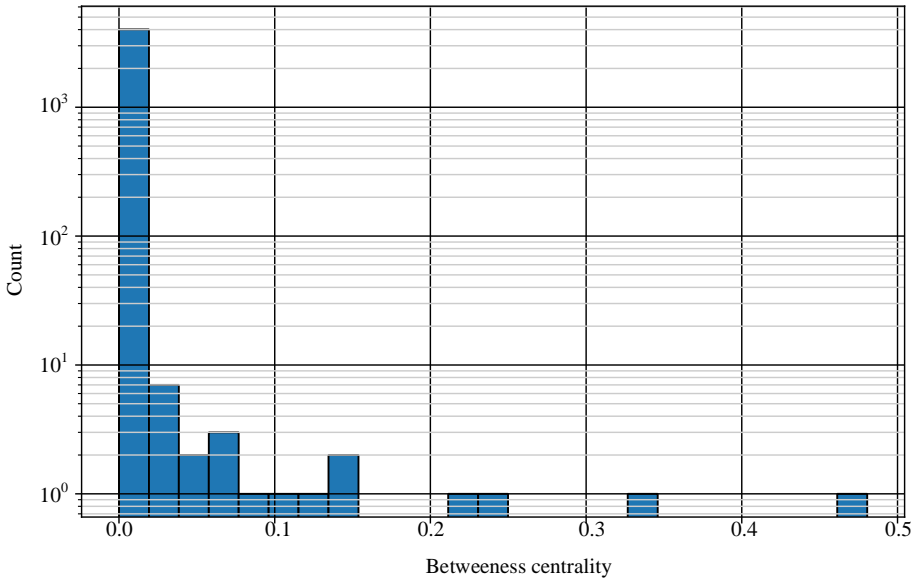


图 14. 介数中心性直方图



图 15. 社交网络图，紧密中心性

排版时，请替换为矢量图，见附件 SVG 文件

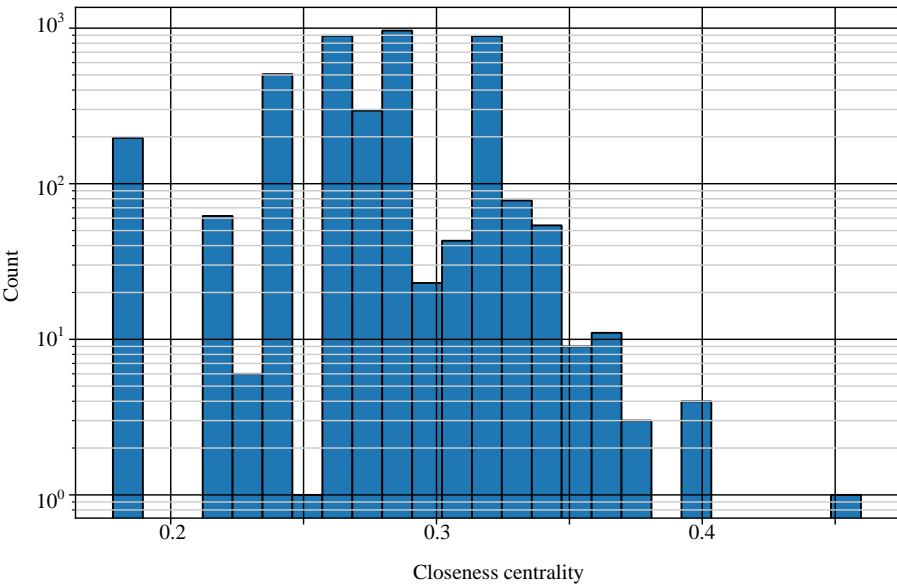


图 16. 紧密中心性直方图

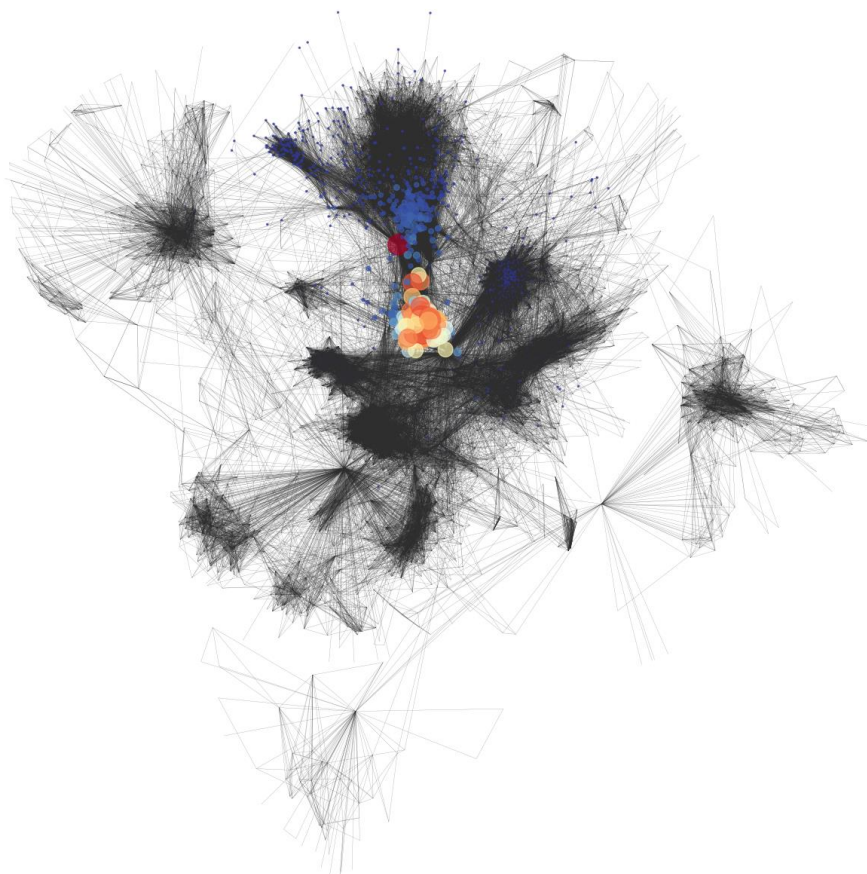


图 17. 社交网络图，特征向量中心性

排版时，请替换为矢量图，见附件 SVG 文件

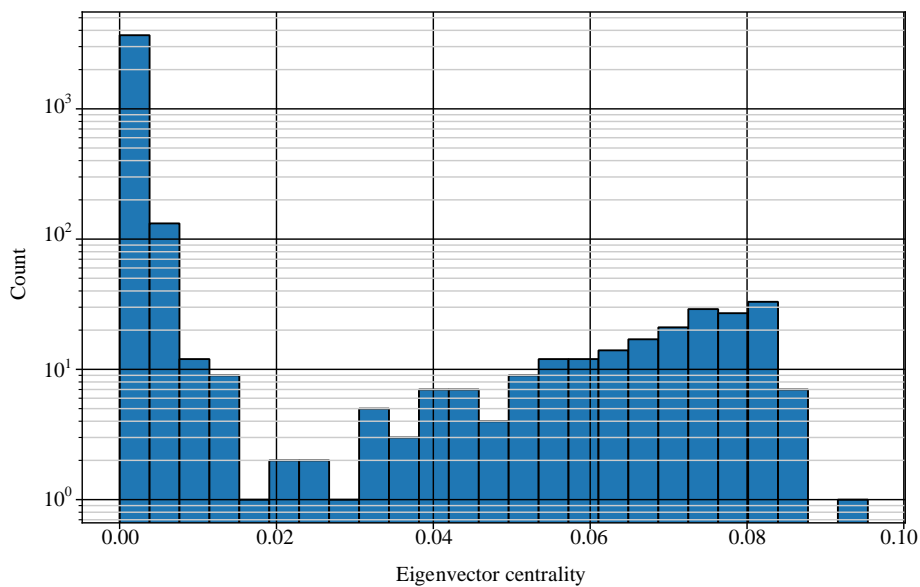


图 18. 特征向量中心性直方图

25.5 社区结构

在一个连通图中，桥是连接两个不同连通分量的边。如果移除一个图中的桥，就会使得图变得不再连通。桥的存在性和识别对于理解图的连通性和社交网络中的重要连接至关重要。在社交网络中，桥可能代表着两个不同的社交群体之间的连接，移除桥可能导致社交网络的分裂。

图 19 中红色边代表社交网络中存在的 75 座桥。

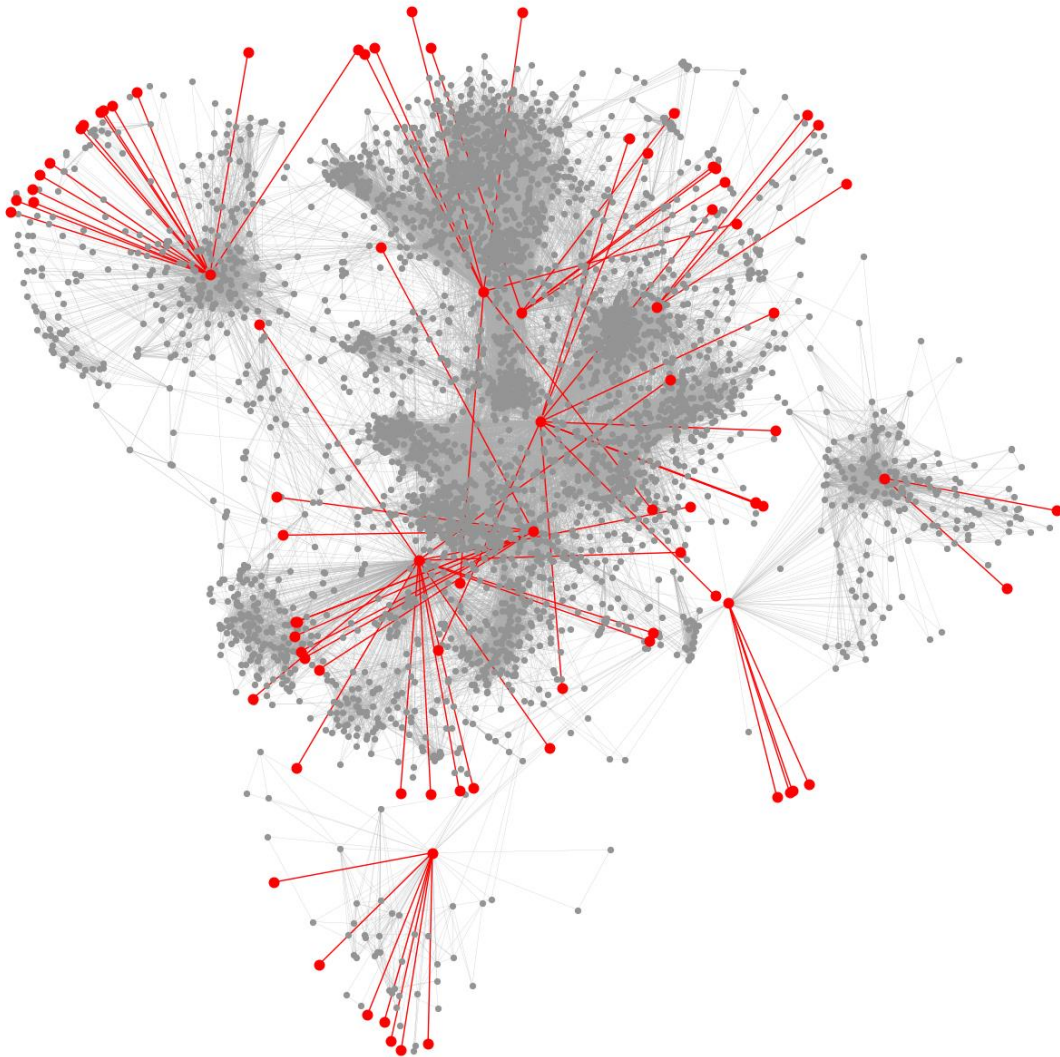


图 19. 社交网络图中的 75 座桥

排版时，请替换为矢量图，见附件 SVG 文件

局部桥是指在社交网络中，连接两个具有很高相似度的节点的边。具体来说，如果边 (u, v) 是一个局部桥，那么节点 u 和节点 v 在社交网络中可能有很多共同的邻居；但是，边 (u, v) 是它们之间唯一的连接。局部桥在社交网络中起到重要的桥接作用，使得相似但非直接相连的节点之间建立联系。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 20 中蓝色边为社交网络中存在的 78 座局部桥。

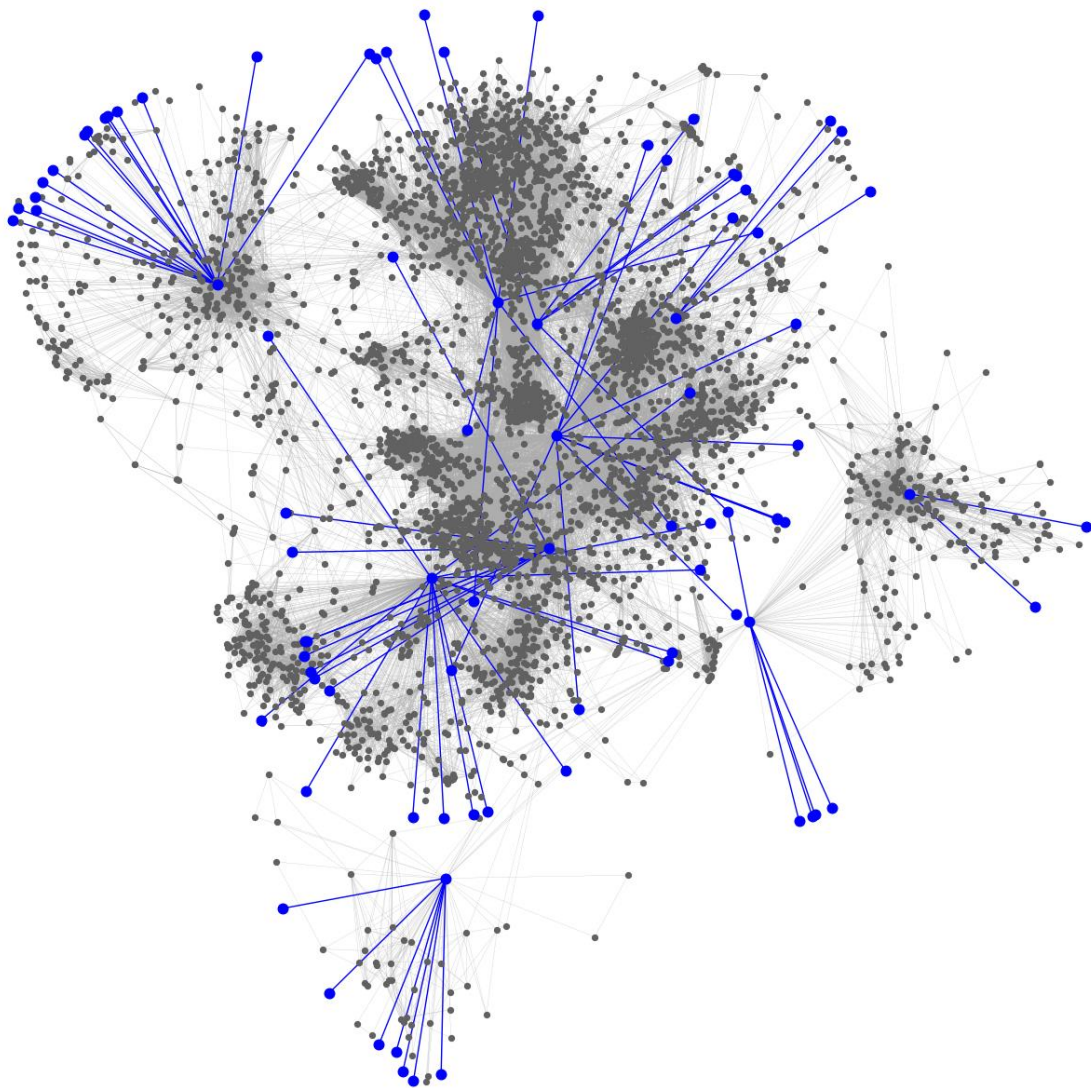


图 20. 社交网络图中的 78 座局部桥

排版时，请替换为矢量图，见附件 SVG 文件

本书前文介绍过，所有桥都是局部桥，但不是所有局部桥都是桥。桥的定义涉及到图的全局结构，而局部桥的定义主要关注节点的局部邻域。

图 21 所示为利用标签传播 (label propagation) 完成社区划分。

标签传播是一种简单而高效的社区检测算法，其基本原理如下：

- ▶ 初始化标签：将每个节点初始化为一个唯一的标签。
- ▶ 标签传播：在每一轮中，节点会将其当前标签传播给邻居节点。具体来说，节点选择其邻居中标签数最多的标签，并将自己的标签更新为这个最多的标签。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

- ▶ 迭代：重复进行标签传播过程，直到网络中的节点标签趋于稳定。这是一个迭代的过程，每一轮都涉及节点的标签更新。
- ▶ 社区形成：当标签传播稳定后，具有相同标签的节点被认为属于同一个社区。

这个算法不需要预先知道社区的数量，并且在大型网络中具有较好的扩展性。然而，标签传播算法的结果可能对初始节点标签敏感。



图 21. 社区划分，标签传播

社交网络分析利用图论中的数学工具来研究社交结构通过节点（个体）和边（关系）的模式。

度分析关注节点的直接联系数量，揭示影响力或活跃程度。

图距离度量时节点间最短路径，图距离相关概念（平均距离、图距离矩阵、离心率、直径、半径）有助于理解信息流动的效率。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

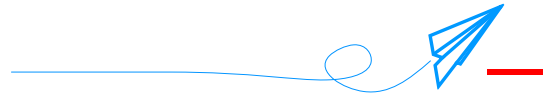
本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

中心性分析，如度中心性、介数中心性、紧密中心性、特征向量中心性等，评估节点在网络中的重要性，识别关键影响者。

社区结构分析通过识别紧密连接的节点群组，揭示网络内的自然分层或团体，有助于理解网络的细分结构和功能。

这些方法共同提供了深入理解社交网络动态和结构特性的手段，对于社会科学、市场营销和信息技术等领域至关重要。



《数据有道》几易其稿。稿件不断大修大改的过程中，笔者不断问自己，《数据有道》怎么写才把鸢尾花书之前五本书的内容融合在一起，又能用数据视角扩展知识网络，还能帮助大家铺平学习第 7 册《机器学习》的道路？

想来想去，想到一个办法——以数据为视角，强调实践中可能出现的数据相关工具。

《数据有道》中大家看到前五本书介绍的各种编程、可视化、数学工具在数据实践相关的应用，同时又拓展讲解了时间序列、图这两种有趣的数据形式。

图和网络是《数据有道》的一大特色，从图论入门、图与矩阵，到图论实践，图占据了本册大半。特别是通过图的各种应用场景，我们还回顾了线性代数中常用的数学工具。

学完本书，希望大家特别记住这句话——图就是矩阵，矩阵就是图。

让我们在《机器学习》再见！