

10

Fundamentals of Graph Theory

图论入门

世间万物关系都是网状



人们思考皆，浮皮潦草，泛泛而谈；现实世界却，盘根错节，千头万绪。

We think in generalities, but we live in details.

—— 阿尔弗雷德·怀特海 (Alfred Whitehead) | 英国数学家、哲学家 | 1861 ~ 1947



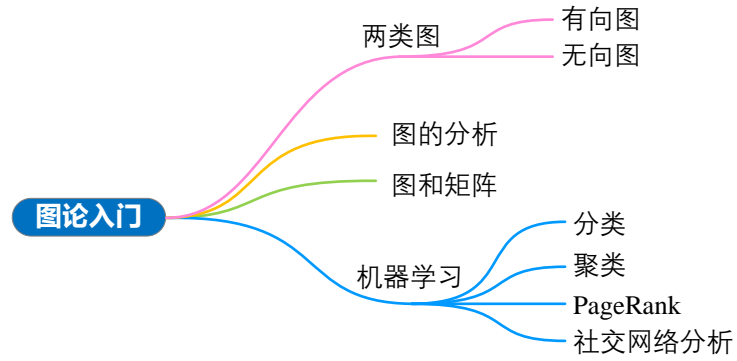
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



10.1 什么是图？

图论 (Graph Theory) 是数学的一个分支，研究的是图的性质和图之间的关系。图由节点和边组成，节点表示对象，边表示对象之间的关系。

历史上，图论起源于 18 世纪，数学家**欧拉** (Leonhard Euler) 最先提出解决了七桥问题的数学方法，开创了图论的先河。

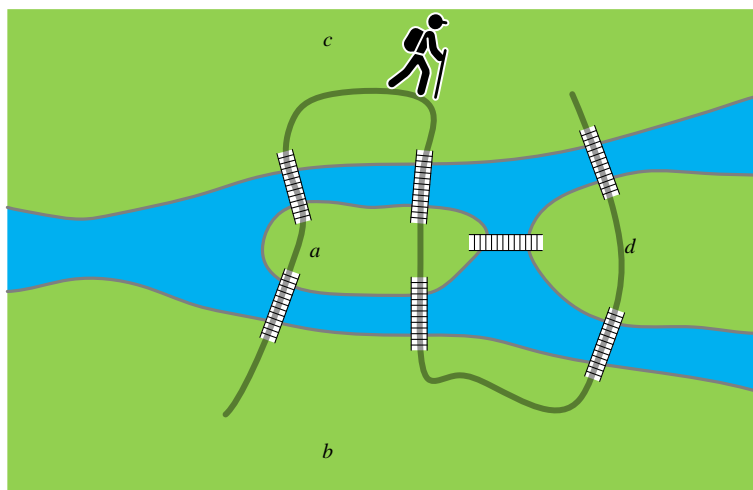


图 1. 七桥问题，走遍图中七座桥，每座桥只经过一次

柯尼斯堡七桥问题 (Seven Bridges of Königsberg)，简称七桥问题，的背景是**基尔岛** (Königsberg) 的**普雷格尔河** (Pregel River) 上有两座岛 (a、d)，有 7 座桥将两座岛和两岸 (b、c) 相连。问题是能否走遍这七座桥，每座桥只经过一次，并最终回到起点。

欧拉解决这个问题的方法是抽象化。他将问题中的地理元素简化成**节点** (nodes) 和**边** (edge) 的**图** (graph)。节点也称**顶点** (vertex 复数 vertices)。

⚠ 注意，本书一般采用“节点”这一表达，目的是和 NetworkX 统一。

每座桥成为图中的一条边，每个岸上的土地成为一个节点。这样，问题就转变成了在这个图上找一条路径，经过每条边一次且仅一次。这就是所谓的“一笔画问题”，即 Eulerian path。

欧拉将问题抽象化，引入了图论的概念，奠定了图论这一数学分支的基础。他的方法和思想对后来图论和网络理论的发展产生了深远的影响。

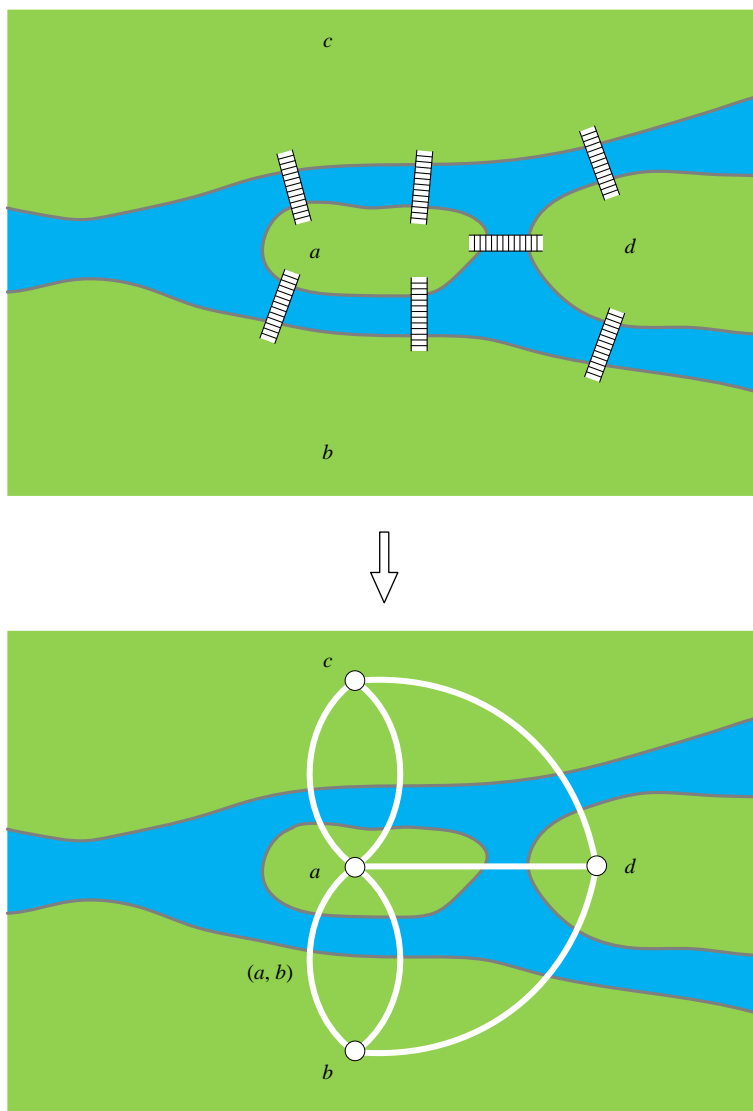


图 2. 七桥问题

无向图

无向图 (undirected graph) 是一种图，它的边没有方向。节点之间的连接是双向的，没有箭头指示方向。无向图常用于描述简单的关系，如社交网络中的朋友关系。

简单来说，无向图由节点集合和边集合构成，其中节点集合表示图中的元素，边集合定义了连接这些节点的关系。如图 3 所示，无向图就好比按特定方式布置的人行步道，任意两个节点并不限制通行方向。

在无向图中，边无权重意味着连接节点的边没有相关的数值信息。在无权重无向图中，通常使用 0 和 1 表示边的存在或不存在。具体而言，如果节点之间有边相连，则用 1 表示，否则用 0 表示。而有权重无向图的边有关联的数值，这些数值可以是距离、相似度、相关性系数等等。

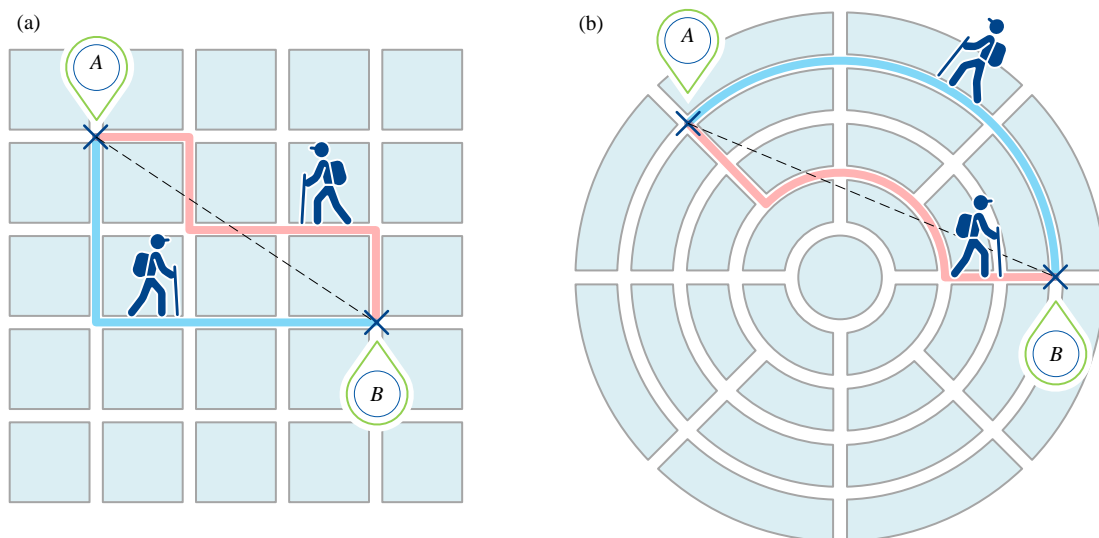


图 3. 不同方法布置的步道

图 4 所示为 5757 个 5 个字母的单词上生成一个无向图；如果两个单词在一个字母上不同，它们之间就会有一条边。

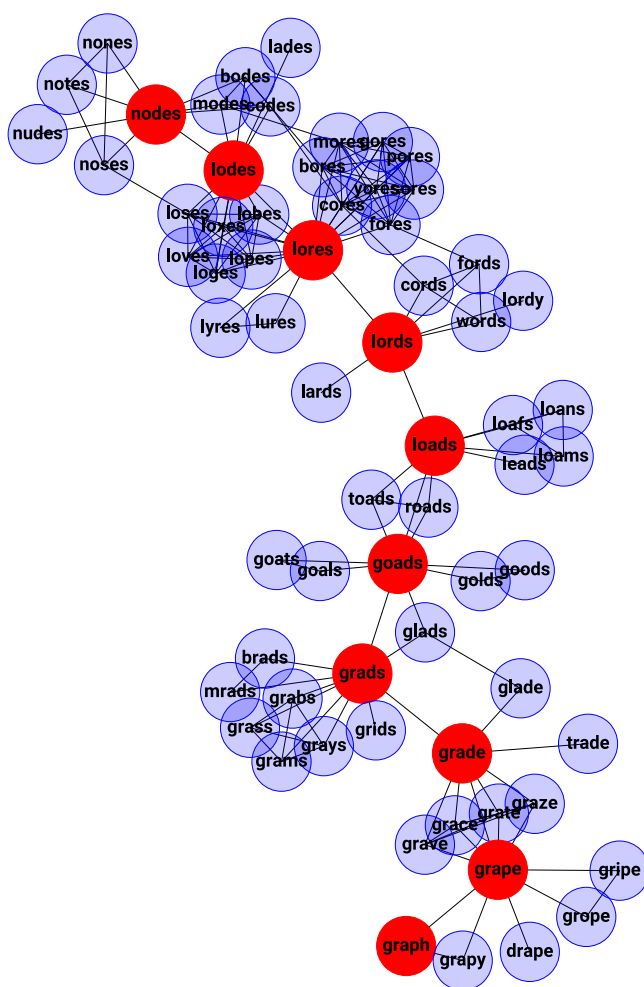


图 4. 5757 个 5 个字母的单词上生成一个无向图

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

无向图还可以用来呈现图 5 所示的这种**社交网络** (social network)。社交网络中的用户关系可以被建模成一个图，其中节点表示用户，边表示用户之间的连接。这种图结构有助于分析信息传播、社交网络分析等问题。

图 5 所示的是图论中一个经典数据集——空手道俱乐部人员关系图。

如图 5 (a) 所示，这个空手道俱乐部一共有 34 名成员，编号从 0 到 33；图中每个节点代表一个成员。节点之间如果存在一条边（黑色线），就代表两个成员存在好友关系。

图 5 (a) 似乎已经告诉我们这个俱乐部存在两个“中心人物”——0 和 33。

将图 5 (a) 布置成图 5 (b)，并且想办法根据每个节点（成员）的“中心性”大小分配不同颜色。越偏向暖色系，说明该成员越居于中心；越偏向冷色系，说明该成员越边缘。

图 5 (b) 显然地告诉我们 0 和 33 是这个空手道俱乐部的“灵魂人物”。有意思的是，这个俱乐部后来因为这两个人的矛盾一分为二，也印证了起初的分析。

本书后续将介绍量化图 5 (b) 所示的这种“中心性”的不同方法。

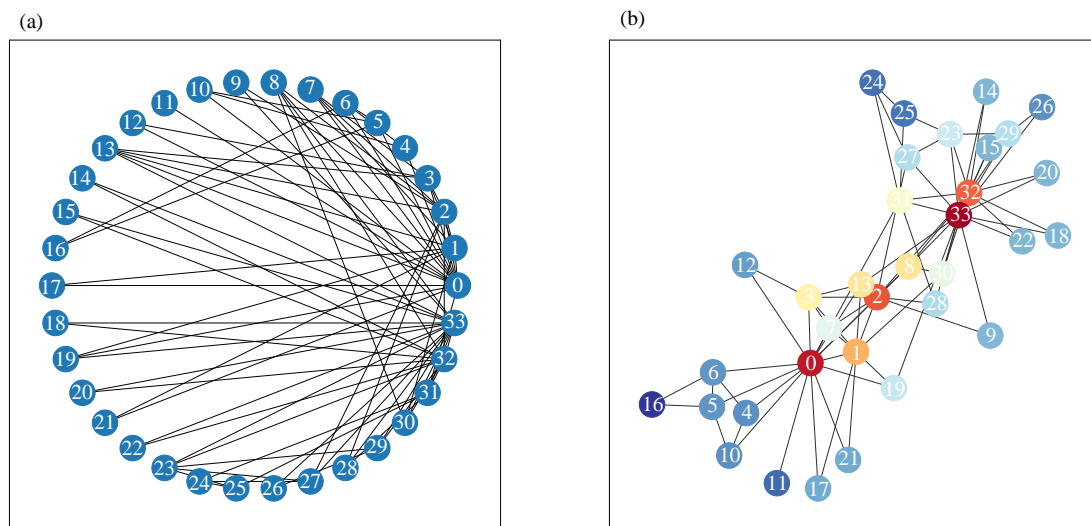


图 5. 空手道俱乐部人员关系图

有向图

有向图 (directed graph) 则是边有方向的图，每条边从一个节点指向另一个节点。有向图常用于描述有向关系，例如网页之间的链接、任务执行的顺序等。

顾名思义，边有方向的图就是**有向图** (directed graph, digraph)。在图 3 的步道中任意两个节点规定通行方向，我们便得到了有向图。

生活中，有向图无处不在。

图 6 所示的陆地物质能量流动链条就可以抽象成有向图。分析这幅图，我们可以知道陆地生物链的能量流动模式。

图 7 所示的多地之间航班信息也可以抽象为一幅有向图。有向图中节点代表城市，有向边代表航班。有向边权重(颜色渲染)代表航班载客量。分析这幅有向图，可以得到不同城市机场的重要性，可以设计新航线，优化航班配置资源。

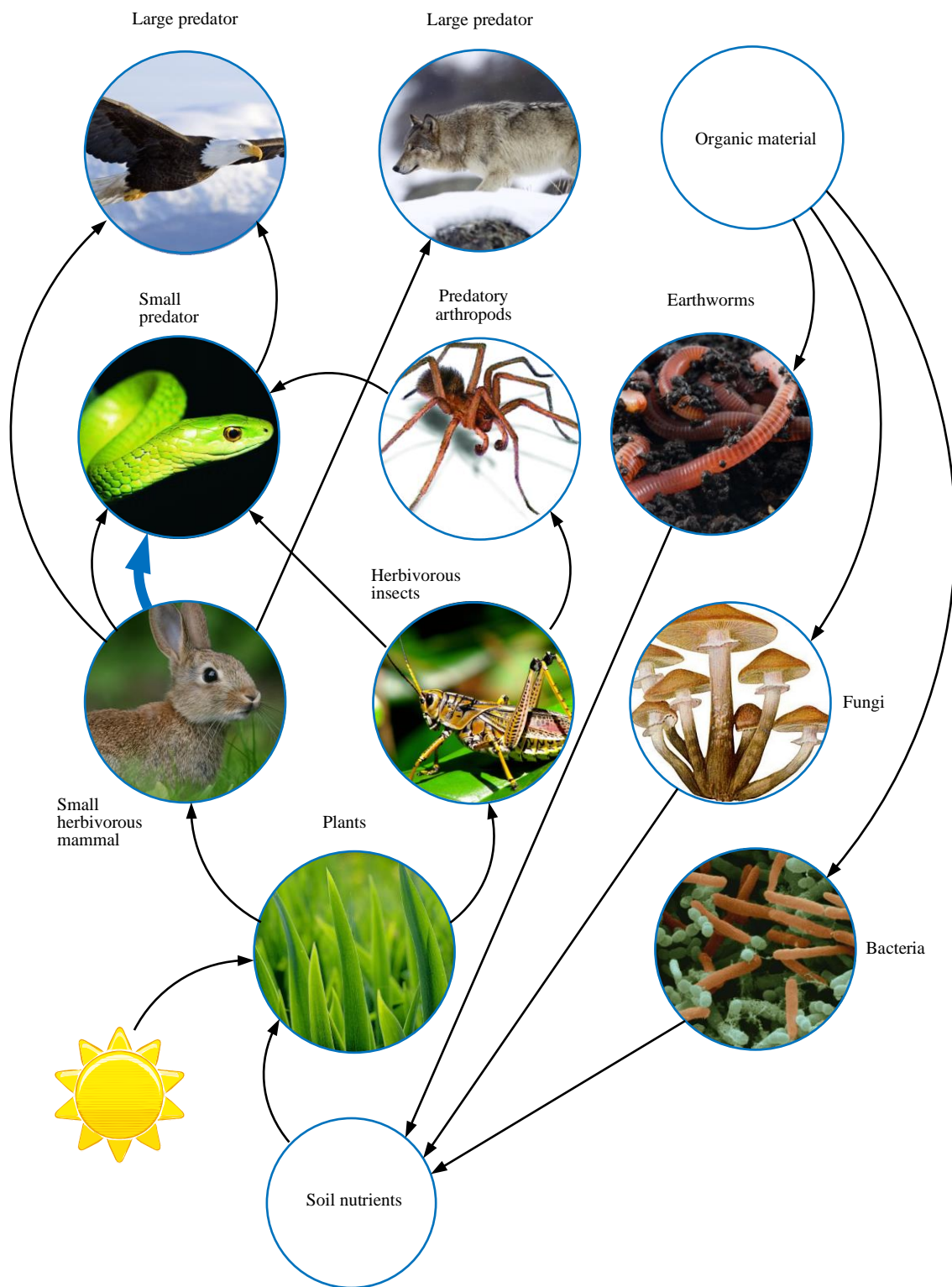


图 6. 食物链中物质能量流动链条具有方向性

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

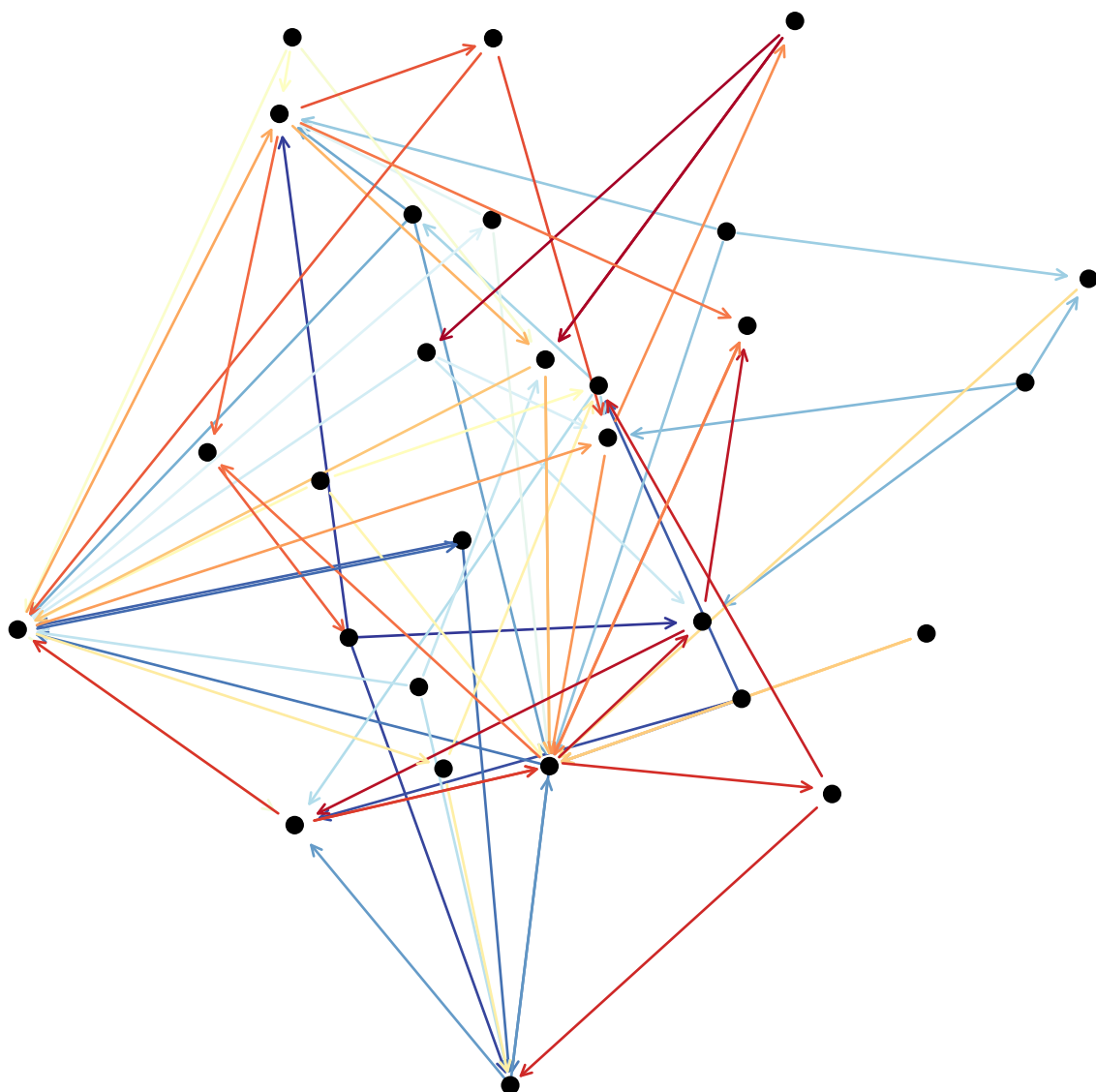
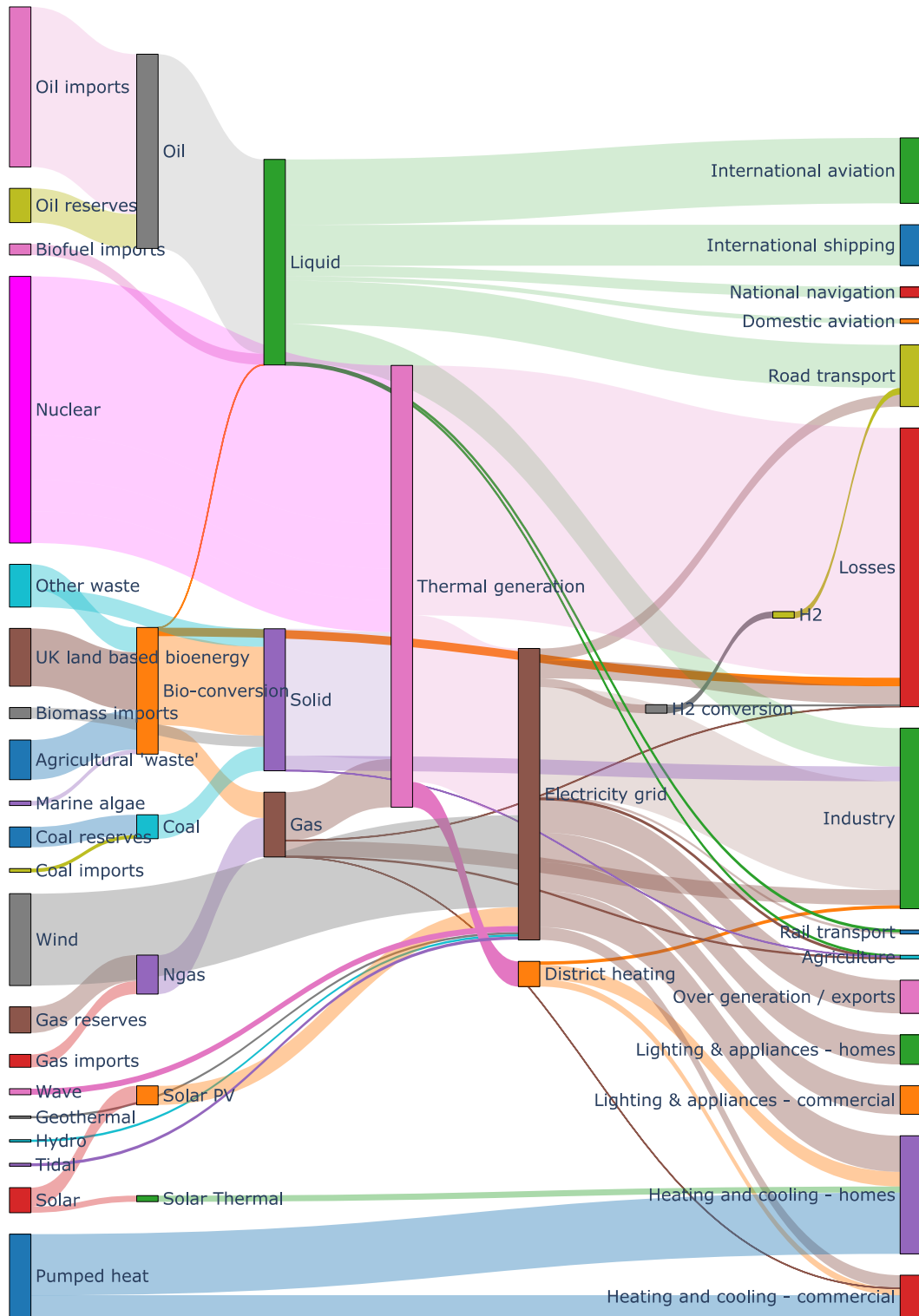


图 7. 航班具有方向性

图 8 用**桑基图** (Sankey diagram) 可视化未来能源流向。我们可以从图论和网络分析的角度理解这幅图。图中的节点代表能源系统中的各种实体，例如，能源来源、转换过程、最终消费者等，而边表示能源从一个实体流向另一个实体的路径。每条边都有一个与之关联的权重，这个权重代表能源流的大小或者比例。这种有向图的表达形式非常直观地揭示了能源如何在不同的实体之间转移和转化。

图 8. 2050 年能源预测，来源 <https://plotly.com/python/sankey-diagram/>

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

10.2 图和几何

图和几何的联系千丝万缕。首先，一幅图中的节点、边就自带几何属性。可以这样说，图这种数学思想就是典型的“几何化”思维。下面，让我们用几个例子让大家看到图和几何之间的联系。

《数学要素》介绍过的**柏拉图立体** (Platonic solid) 中的**正四面体** (tetrahedron) 就直接对应**正四面体图** (tetrahedral graph)，具体如图 9 所示。

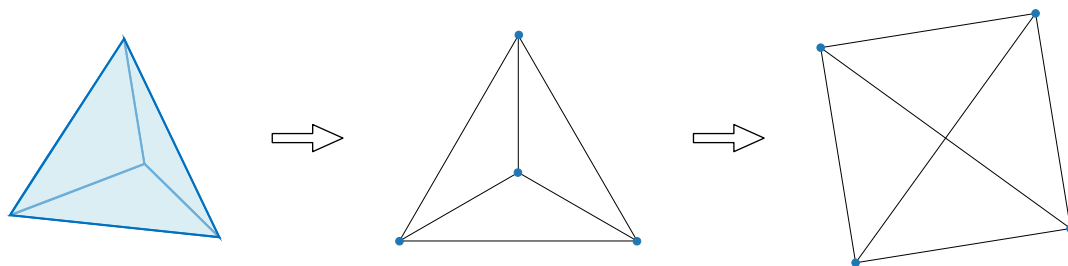


图 9. 正四面体和正四面体图

图 10 左侧散点图有 12 个点，共有 66 个成对距离。我们可以把它抽象成一个由 12 个节点、66 条边构成的无向图。边的权重用对应的欧氏距离值表示。图 10 还利用颜色映射根据欧氏距离大小对边进行渲染。冷色系的边代表距离远，暖色系的边代表距离近。

进一步观察，我们可以发现将偏冷色系的边删除，我们似乎可以把这 12 个散点分成两簇。这就是图论在机器学习领域另外一个重要应用——聚类。

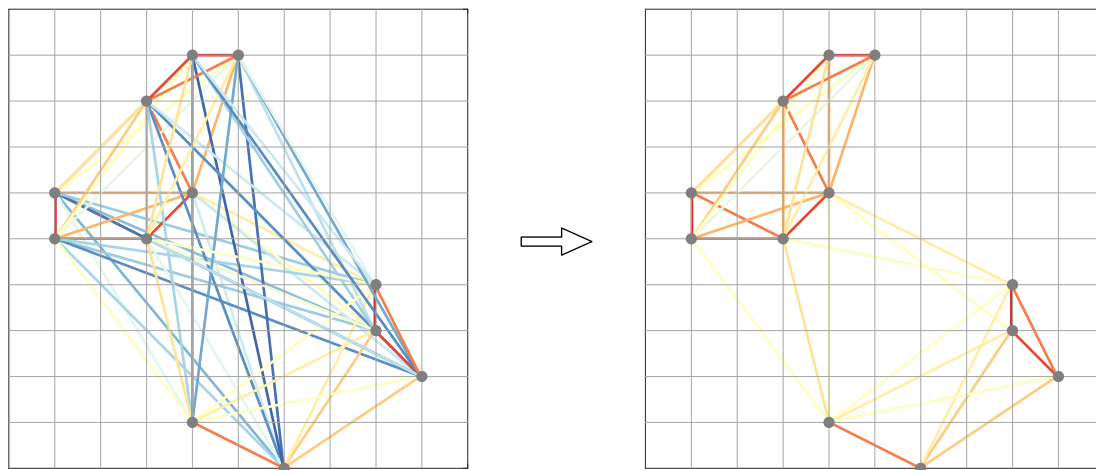


图 10. 散点两两欧氏距离

图 11 所示的**推销员问题** (Traveling Salesman Problem, TSP) 是经典的路径问题之一。简单来说，给定一系列城市 (图 11 中蓝点) 和每对城市之间的距离 (图 11 中图的边长度)，推销员问题求解访问每一座城市一次并回到起始城市的最短回路。图 11 中红色回路就是我们要找的最优化解。如果图中的边权重代表两个城市飞机机票价格，推销员问题也可以求解访问所有城市路费最低的回路。

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

本书还会介绍其他几种常见的路径问题。

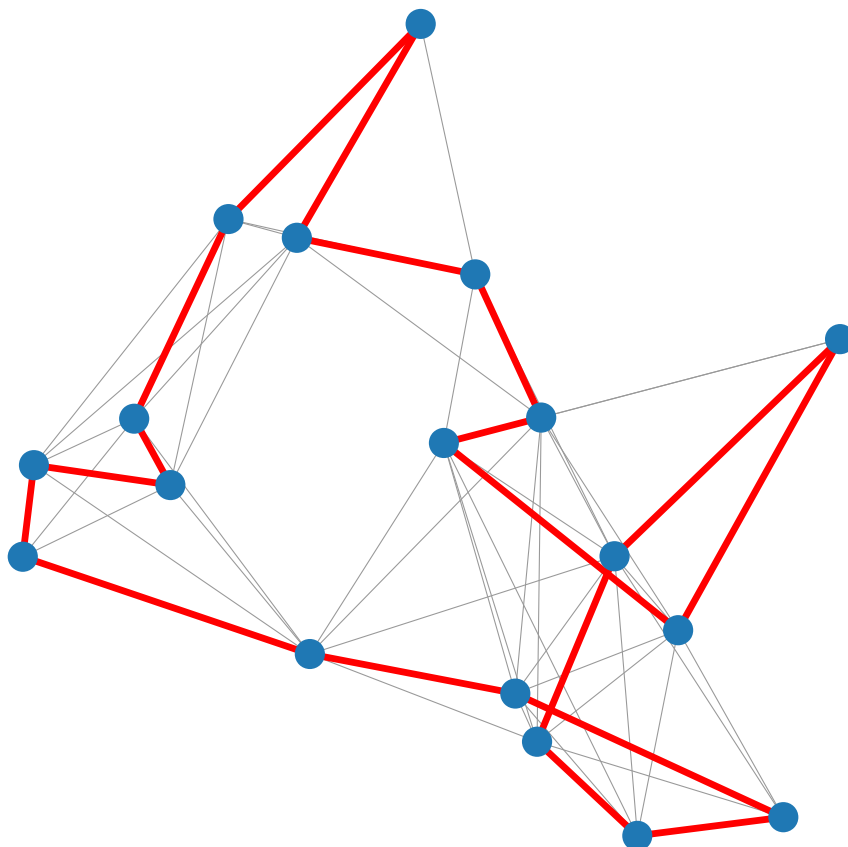


图 11. 推销员问题

10.3 图和矩阵

图就是矩阵，矩阵就是图！请大家在遇到任何一幅图，或者看到任何一个矩阵的时候，要多一层图和矩阵联系思考。

图 12 所示的就是无向图和**邻接矩阵** (adjacency matrix) 之间的有趣关系。简单来说，对于简单图，如果两个节点之间有一条边，邻接矩阵相应位置就为 1；如果不存在边的话，邻接矩阵相应位置便为 0。

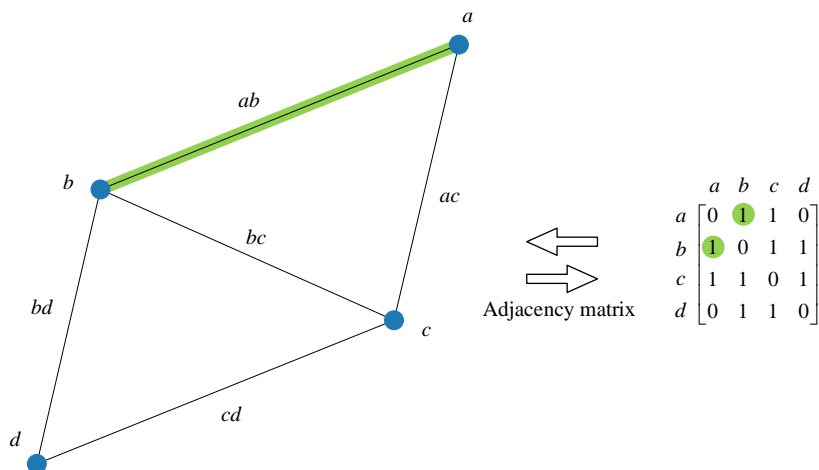


图 12. 无向图和邻接矩阵

类似地，有向图也有对应的邻接矩阵 (如图 13 所示)，相关内容请大家参考本书第 18 章。

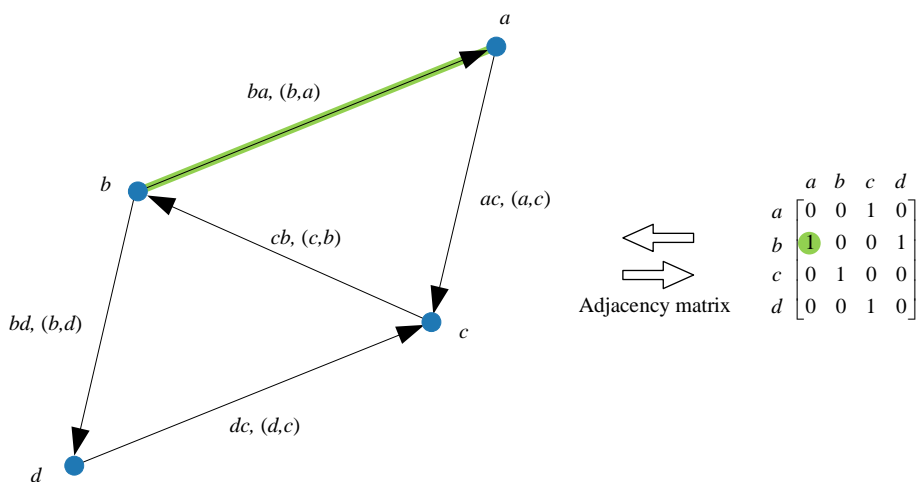


图 13. 有向图和邻接矩阵

换个角度来看图 10，图中散点之间的成对欧氏距离矩阵本身就是一幅图！

如图 14 所示，欧氏距离矩阵可以“抽象”为一幅无向图，反之亦然。

进一步拓展思维，我们可以发现成对亲近度矩阵、成对余弦距离、协方差矩阵、相关性系数矩阵等等都可以看成是图。

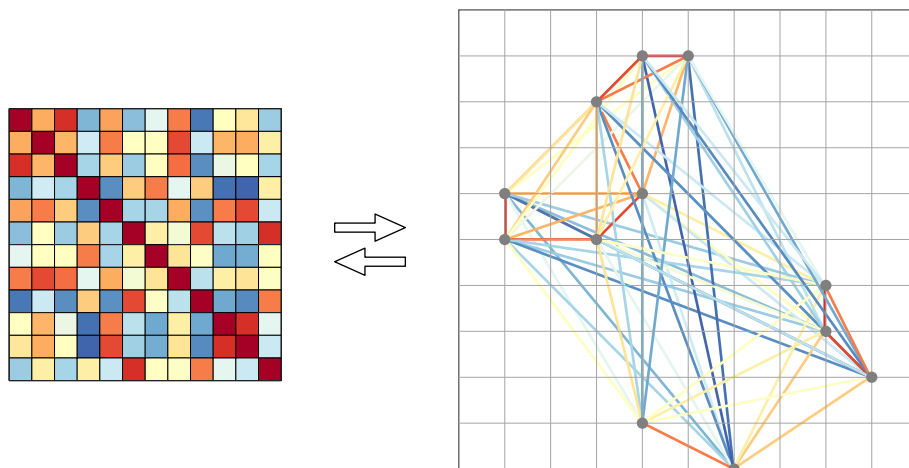


图 14. 成对欧氏距离矩阵

大家是否回忆起《数学要素》在最后介绍的鸡兔同笼三部曲中“鸡兔互变”？

图 15 左图实际上就是一幅有向图；而**转移矩阵** (transition matrix) T ，就是有向图的一种矩阵表达。这也告诉我们图、条件概率 (图 16)、马尔科夫链、随机过程这些数学板块之间的联系也盘根交错。

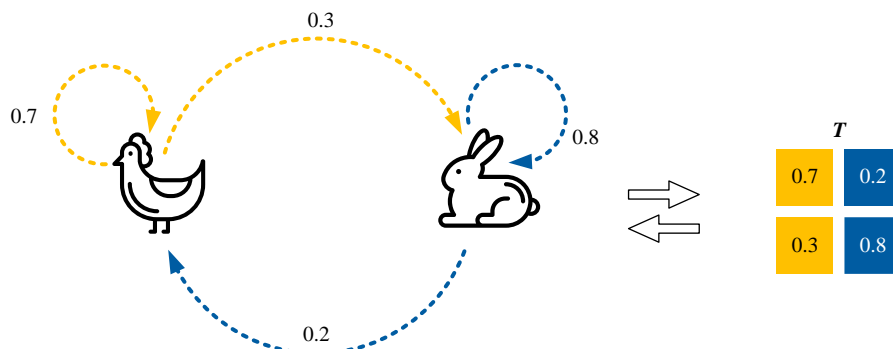


图 15. 鸡兔同笼三部曲中“鸡兔互变”，图片来自本系列丛书《数学要素》第 25 章

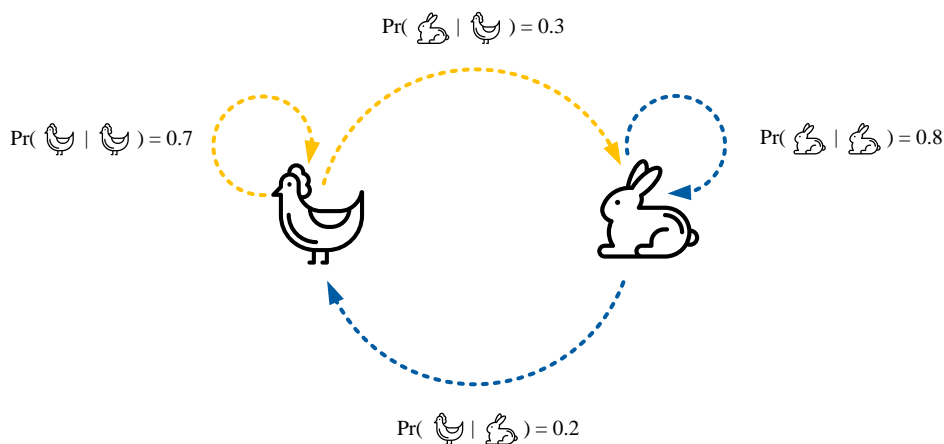


图 16. “鸡兔互变”中的条件概率

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

图和矩阵的关系不止如此，本书后续大家还会接触到**关联矩阵** (incidence matrix)、**度矩阵** (degree matrix)、**拉普拉斯矩阵** (Laplacian matrix) 等概念。

10.4 图和机器学习

在机器学习中，图论常用于表示和分析数据之间的关系。图模型可以用来建模复杂的关联关系，尤其在结构化数据和网络数据方面。

从具体算法分类角度来看，图论可以用来**分类** (classification)、**聚类** (clustering)。举几个例子，图的特殊形态——树——在机器学习算法中应用很多，比如最近邻 (k -Nearest Neighbor, k -NN) 算法中的 kd 树，**决策树** (decision tree)，**层次聚类** (hierarchical clustering)，等等。

图 17 ~ 图 20 所示为层次聚类用在股票聚类。图 17 所示为 17 只股票日收益率的相关性系数矩阵。图 18 所示为将相关性系数矩阵转化成的距离矩阵；简单来说，相关性系数越大，距离越近，越靠近 0；反之，相关性系数越小，距离越远，越靠近 1。

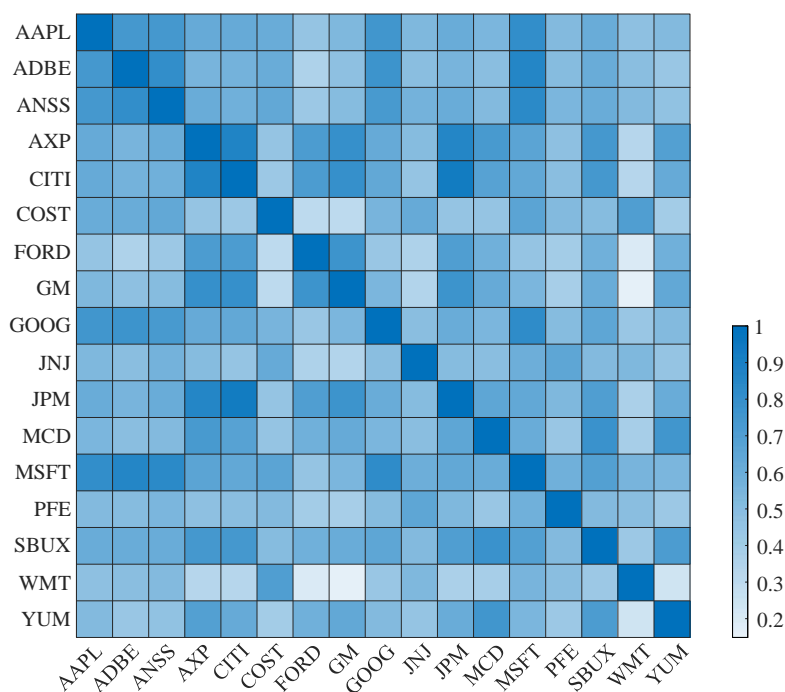


图 17. 相关性矩阵

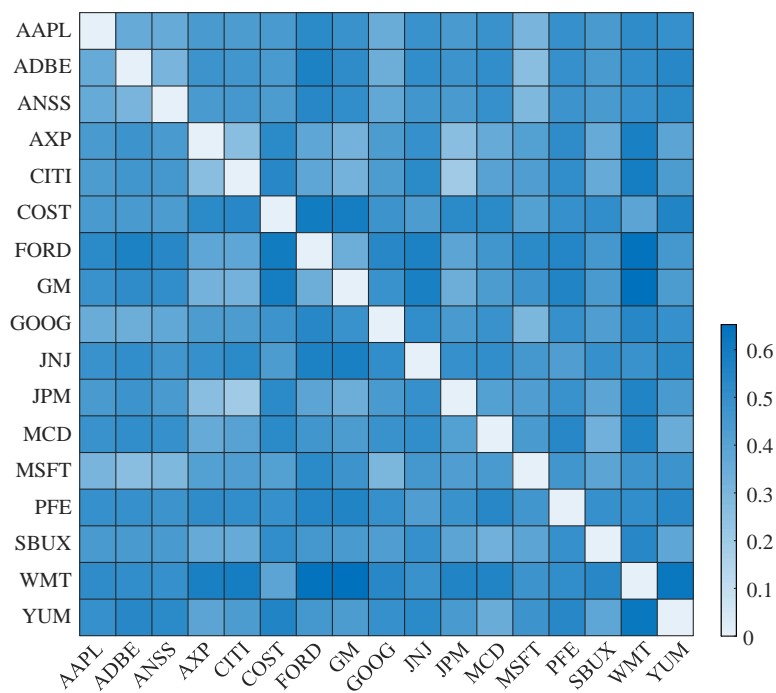


图 18. 距离矩阵

图 19 展示图 18 样本数据的树形图。树形图横轴对应股票，纵轴对应数据点间距离和簇间距离。图 20 所示为根据层次聚类重新排布的相关性矩阵。容易发现，同一行业的个股距离很近，因此被分为一簇，比如这几簇：(CITI、JPM 和 AXP)，(FORD 和 GM)，(MCD、SBUX 和 YUM)，(AAPL、ADBE、MSFT、ANSS 和 GOOG)，(COST 和 WMT)和 (JNJ 和 PFE)。

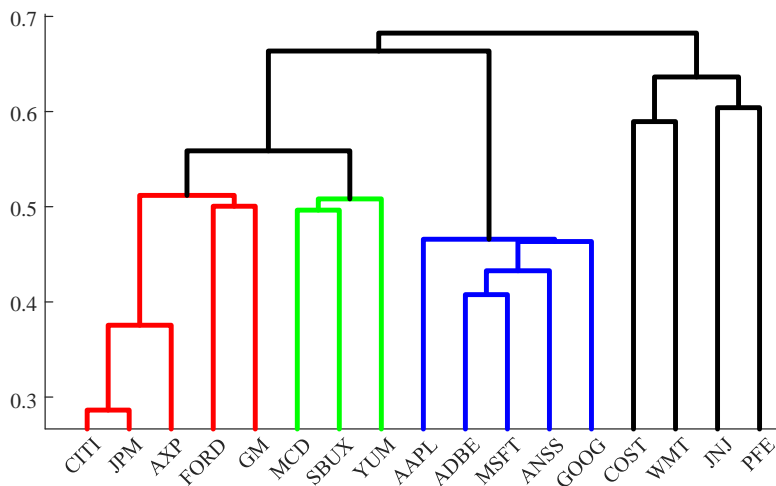


图 19. 距离矩阵数据树形图

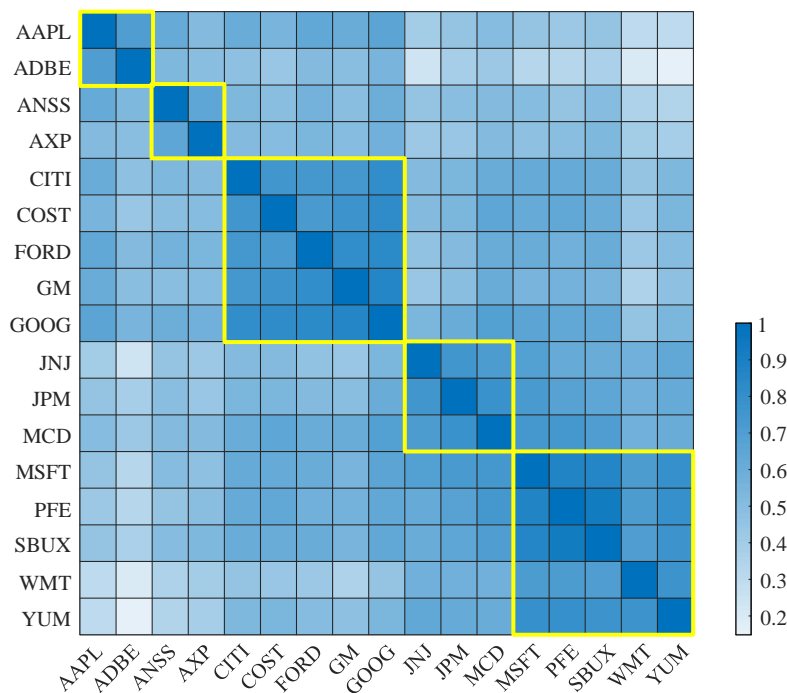


图 20. 根据层次聚类重新排布的相关性矩阵

本书后续还会介绍图论的其他几个应用，比如 PageRank 算法、**社交网络分析** (Social Network Analysis, SNA)。

生活中，我们会发现没有人是一座座孤岛，世界是一张极其错综复杂的网络。简单来说，人以类聚，物以群分，通过分析社交网络关系，我们可以在网络中发现隐藏的组织结构、社区群体、信息流动等等信息。在图 21 这个社交网络中，我们可以很容易发现 3 个社区。

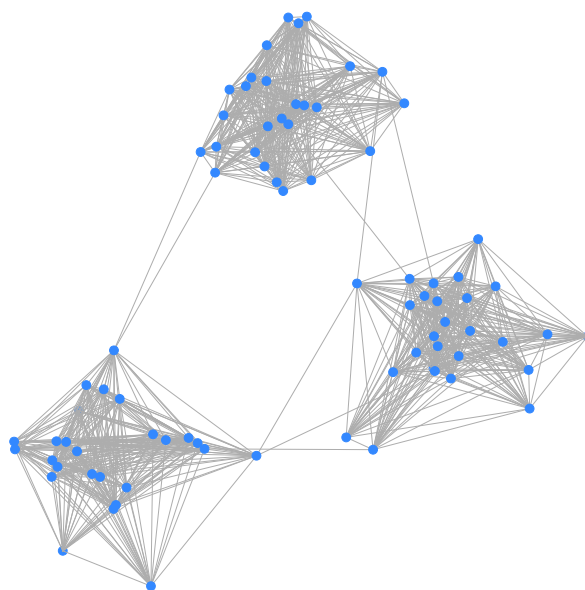


图 21. 社交网络中的社区

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

对于图 22 所示的社交网络，我们可以用社交网络分析 SNA 发现其中“影响力”更大的节点（用户）。当然，我们也可以分析得到其中隐藏的社区结构，这是本书最后一章要介绍的案例。

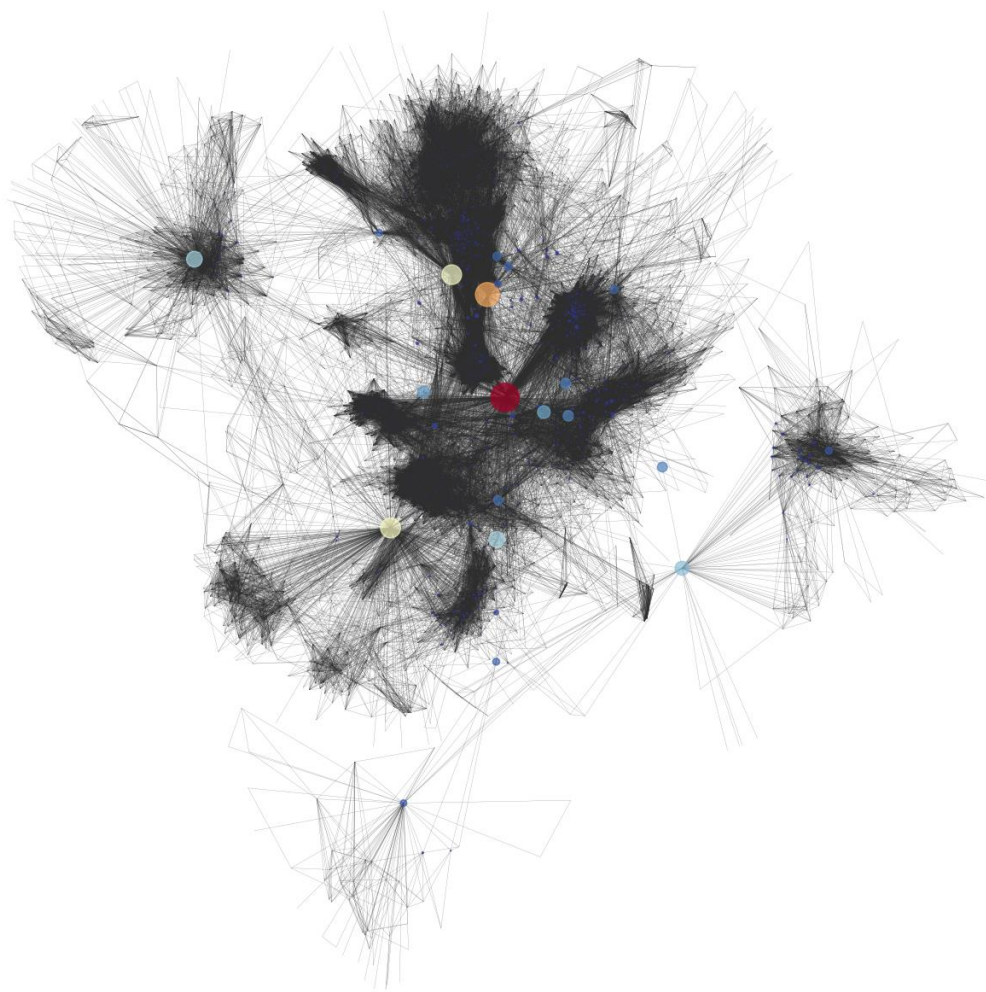


图 22. 社交网络图，发现其中“影响力”较大的节点

排版时，请替换为矢量图，见附件 SVG 文件

此外，在深度学习中，**图神经网络** (Graph Neural Networks, GNNs) 是图论的一种扩展，专门用于处理图结构数据。GNNs 能够在图上进行节点和边的信息传递，使得模型能够理解节点之间的复杂关系。

至于**大语言模型** (Large Language Model, LLM)，图论在自然语言处理中的应用主要体现在语言结构的建模上。语言结构可以被视为一个图，其中单词或子词是节点，语法和语义关系则是边。通过图模型，大语言模型可以更好地理解语言的层次结构和关联关系，从而提高对文本理解和生成的能力。

10.5 NetworkX

NetworkX 是一个用 Python 编写的图论和复杂网络分析的开源软件包。它提供了创建、操作和研究复杂网络结构的工具。以下是一些 NetworkX 的主要特点和用途。

NetworkX 允许用户轻松创建各种类型的图，包括无向图、有向图、加权图等。它提供了丰富的图操作和算法，使用户能够对图进行修改、查询和分析。

NetworkX 支持图的可视化，可以使用各种布局算法将图形绘制成可视化图形。这有助于直观地理解和展示复杂网络的结构。

NetworkX 包含许多图算法，涵盖了图的各个方面，如最短路径、连通性、中心性度量等。用户可以利用这些算法来分析和研究图的特性。

除了基本的图操作和算法外，NetworkX 还提供了用于复杂网络分析的工具。这包括社区检测、小世界网络、度分布等分析方法。

NetworkX 支持从多种数据源导入图数据，也可以将图数据导出为不同的格式，如 GML、GraphML、JSON 等。

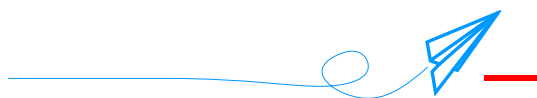
由于 NetworkX 是用 Python 编写的，因此具有很高的灵活性和可扩展性。用户可以方便地自定义算法和功能，以满足特定的需求。

本书下面会结合 NetworkX 介绍图论基础内容，并用 NetworkX 构造并求解一些和图论相关的常见数学问题。

NetworkX 提供大量有趣的应用案例，本书会经常结合 NetworkX 案例扩展讲解图论中常用数学概念和工具。强烈推荐大家练习如下 NetworkX 给出的案例。

https://networkx.org/documentation/stable/auto_examples/index.html

《可视之美》最后一章展示的很多网络可视化方案都会在本书展开讲解，也就是大家不但要知其然，也会知其所以然。



本册超过一半的内容和图论有关，但是请大家注意本书毕竟不是一本图论教科书；因此，本书对图论体系不会面面俱到，更强调图论的理论结合实践。

本书图论相关内容如果能够帮读者达成如下学习目标，笔者便心满意足：1) 了解图论基础入门知识，同时不觉得图论无聊，甚至有兴趣继续深入学习；2) 将图论、矩阵、概率统计、几何、随机过程等数学板块联系起来，并且了解图论在机器学习算法中的应用；3) 用 NetworkX 完成常见图论问题的实践。

下面让我们一起开始本书图与网络之旅。