第3章 数据可视化技术

当我们准备好数据，知道数据是由什么类型的属性或字段组成以后，就可以通过相应的数据预处理技术填补数据集中的缺失值，光滑噪声，识别离群点，由基本的统计描述方法获得关于数据集属性值的中心趋势和离散趋势。但是，是否可以用相关的方法对整理好的数据集进行可视化的观察，直观地了解数据看上去如何？值如何分布？离群点分布的大概位置？数据对象之间大致的相似性如何？借助于数据可视化技术不仅能以图形化的方式观察数据，更有助于识别隐藏在杂乱数据集中的关系、趋势和偏差。不同的应用场景和数据类型使用不同的可视化方法。例如，简单的表格类型数据可以使用柱状图和折线图等，而复杂的层级结构数据可以使用树图等可视化方法。

本章从可视化技术的应用及划分开始(3.1 节)，然后学习最常用的高维数据的可视化方法(3.2 节)和网络数据的可视化方法(3.3 节)，最后通过两个竞赛案例对可视化技术的实际应用做详细介绍(3.4 节)。

3.1 可视化简介

数据可视化是一门古老的学科。早在公元 2 世纪，人们就开始用行和列管理数据。到中世纪时期，人们已开始使用包含等值线的地磁图、表示海上主要风向的天象图等。在 20 世纪，随着计算技术和显示技术的发展，数据可视化技术更是得到飞速发展。近十多年来，随着大数据技术的发展，现有可视化技术已难以应对海量、高维、多源和动态数据可视化的挑战，数据可视化技术随之迎来了新的发展机遇。

数据可视化是被许多学科和领域专家所使用的视觉传递现代技术，它包含数据可视表达的创造性和研究性两个层面。数据可视化将数据所包含的信息的综合体，包含属性和变量，抽象化为一些图表形式。数据可视化的主要目的是借助统计图、散点图和其他图的形式，准确、清晰、有效地传达数据中所包含的信息。有效的可视化更能进一步帮助用户分析数据，推论事件，寻找规律。它使得复杂数据更容易被用户理解和使用。

数据可视化处理的对象是数据，根据处理对象的不同，数据可视化可分为科学可视化和信息可视化两个分支。又由于数据可视化具有分析推理等数据分析的功能，且数据分析对理解数据和使用数据具有非常重要的作用，将可视化与数据分析相结合，又形成一个新的数据可视化方向，即可视分析。

科学可视化是一个跨学科的科学分支，也是可视化领域发展最早和最成熟的一个学

科。科学可视化主要关注三维空间数据的可视化，强调线、面、体等几何、拓扑结构的真实表达，其主要应用领域是自然科学，如物理、化学、医学、流体力学、生物学、气象学等学科。根据数据的不同类别，科学可视化可分为标量场可视化、矢量场可视化和张量场可视化。

信息可视化是以增强人的认知能力为目的的抽象数据和非结构化数据可视表达研究。与科学可视化相比，信息可视化主要关注抽象数据，不仅包括数值数据，也包括非数值数据，如文本、图像、层次结构等。统计信息可视化起源于统计图形学，与信息图形学、视觉设计等现代技术相关。信息可视化的研究融合了多个学科，如计算机科学、计算机图形学、人机交互、可视设计、心理学、商业方法等。在本章，主要关注与大数据关系最密切的信息可视化内容。

可视分析是一门辅以交互可视界面进行分析推理的科学，主要通过耦合人的分析能力与机器的计算能力智能地解决由于数据量、问题的复杂性等带来的棘手难题，可视交互界面则是耦合人与机器的介质。可视分析融合了数据表达与分析、人机交互和可视化等技术，已经并仍在推动分析决策、技术转移等技术的发展。

数据可视化是艺术也是科学。作为一门科学，可视化应该真实反映数据所包含的信息；作为一门艺术，可视化应该具有艺术美感。为了实现可视化在科学与艺术间的平衡，可视化需要达到真、善、美三种境界：

真，即真实性，指可视化结果应正确反映数据的本质；

善，即易感知，指可视化结果应有利于公众认知数据背后所蕴含的现象和规律；

美，即艺术性，指可视化结果的形式和内容应和谐统一。

在对数据进行可视化之前首先要了解数据类型，不同的数据类型都有对应的可视化方法。在3.2节中介绍针对高维数据的可视化方法，在3.3节中介绍针对网络数据的可视化方法。

3.2 高维数据可视化

在本节中主要介绍高维数据可视化的处理分析过程。在对高维数据可视化时，如果维度过高则需要对数据进行降维处理，在3.2.1节中介绍几种常见的降维处理方法，在3.2.2节中介绍几种常用的高维数据可视化方法。

无论是在日常生活中还是在科学研究中，高维数据处处可见。例如，一件简单的商品就包含型号、厂家、价格、性能、售后服务等多种属性；再如，在癌症研究中，为了找到与致癌相关的基因，需要综合分析不同患者的成百上千个基因表达；对大气、海洋、宇宙等复杂物理现象的计算模拟，也要考虑到诸如温度、压强等多个维度因素。人们一般很难直观快速地理解三维以上的数据，而将数据转化为可视的形式，就可以帮助人们理解和分析高维空间中的数据特性。高维数据可视化技术旨在用图形表现高维度的数据，并辅以交互手段，帮助人们分析和理解高维数据。

高维数据可视化主要分为降维方法和非降维方法。降维方法指将高维数据投影到低维空间，尽量保留高维空间中原有的特性和聚类关系。常见的降维方法有主成分分析(PCA)、多维尺度分析(multi-dimensional scaling，MDS)、自组织图(self-organization

map,SOM)等。这些方法通过数学方法将高维数据降维,进而在低维屏幕空间中显示。通常,数据在高维空间中的距离越近,在投影图中两点的距离也越近。高维投影图可以很好地展示高维数据间的相似度以及聚类情况等,但并不能表示数据在每个维度上的信息,也不能表现维度间的关系。高维投影图损失了数据在原始维度上的细节信息,但直观地提供了数据之间宏观的结构。

数据降维方法的分类,如图 3-1 所示。

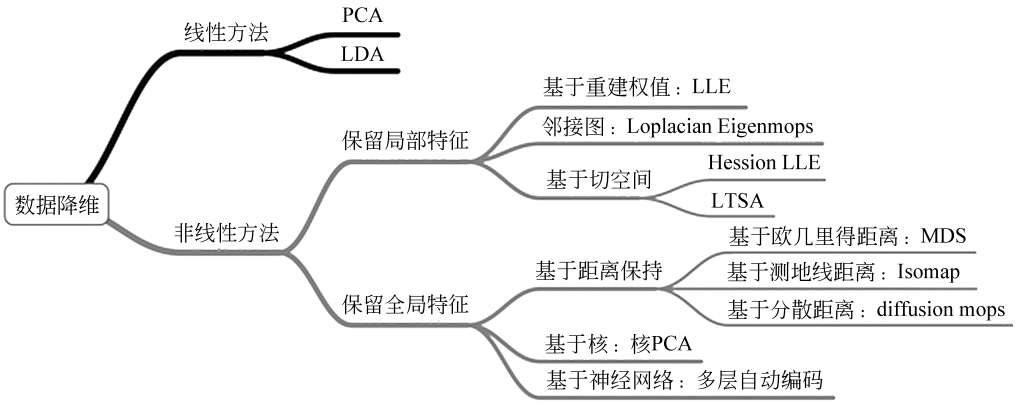


图 3-1　数据降维方法的分类

3.2.1　降维方法

1. 主成分分析(PCA)

PCA 是一种较为普遍的线性降维方法,它的目标是通过某种线性投影,将高维的数据映射到低维的空间中表示,并期望在所投影的维度上数据的方差最大,以此使用较少的数据维度,同时保留较多的原数据点的特性。

PCA 的原理就是将原来的样本数据投影到一个新的空间中,相当于矩阵分析中将一组矩阵映射到另外的坐标系下。通过一个转换坐标,也可以理解成把一组坐标转换到另外一组坐标系下,但是在新的坐标系下,表示原来的样本不需要那么多的变量,只需要原来样本的最大的一个线性无关组的特征值对应的空间的坐标即可。通俗地理解,就是如果把所有的点都映射到一起,那么几乎所有的信息(如点和点之间的距离关系)都丢失了,而如果能使映射后方差尽可能地大,那么数据点则会分散开,以此保留更多的信息。可以证明,PCA 是丢失原始数据信息最少的一种线性降维方式(实际上就是最接近原始数据,但是 PCA 并不试图去探索数据内在结构)。

设 n 维向量 $\boldsymbol{w}$ 为目标子空间的一个坐标轴方向(称为映射向量)最大化数据映射后的方差,有

$$\max_{w} \frac{1}{m-1} \sum_{i=1}^{m} \left[\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right]^2$$

其中,m 是数据实例的个数;$\boldsymbol{x}_i$ 是数据实例 i 的向量表达;$\bar{\boldsymbol{x}}$ 是所有数据实例的平均向量。

定义 $\boldsymbol{W}$ 为包含所有映射向量为列向量的矩阵，经过线性代数变换，可以得到如下优化目标函数：

$$\min_{w} \mathrm{tr}(\boldsymbol{W}^{\mathrm{T}} A \boldsymbol{W}), \text{s.t.} \boldsymbol{W}^{\mathrm{T}} \boldsymbol{W} = \boldsymbol{I}$$

其中，tr 表示矩阵的迹；A 是数据协方差矩阵：

$$A = \frac{1}{m-1} \sum_{i=1}^{m} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}$$

容易得到最优的 $\boldsymbol{W}$ 是由数据协方差矩阵前 k 个最大的特征值对应的特征向量作为列向量构成的。这些特征向量形成一组正交基并且最好地保留了数据中的信息。

PCA 追求的是在降维之后能够最大化保持数据的内在信息，并通过衡量在投影方向上的数据方差的大小衡量该方向的重要性。但是这样投影以后对数据的区分作用并不大，反而可能使得数据点糅杂在一起无法区分。这也是 PCA 存在的最大一个问题，这导致使用 PCA 在很多情况下的分类效果并不好。如图 3-2 所示，若使用 PCA 将数据点投影至一维空间上时，PCA 会选择 2 轴（ϕ_2），这使得原本很容易区分的两簇点被糅杂在一起变得无法区分；而这时若选择 1 轴（ϕ_1）将会得到很好的区分结果。

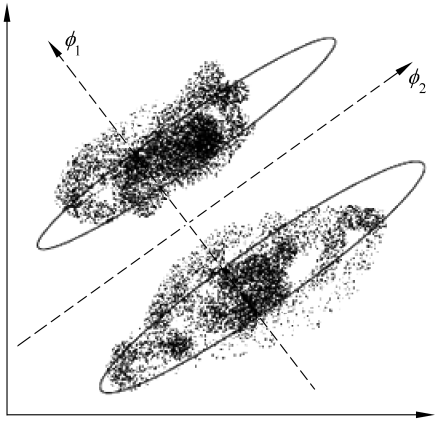


图 3-2 PCA 分析图

比如，原来的样本的维度是 30×1 000 000，就是说有 30 个样本，每个样本有 1 000 000 个特征点，这个特征点太多了，需要对这些样本的特征点进行降维。在降维的时候会计算一个原来样本矩阵的协方差矩阵，这里就是 1 000 000×1 000 000，当然，这个矩阵太大了，计算的时候会用其他的方式进行处理，这里只是讲解基本的原理，然后通过这个 1 000 000×1 000 000 的协方差矩阵计算它的特征值和特征向量，最后获得具有最大特征值的特征向量构成转换矩阵。例如，前 29 个特征值已经能够占到所有特征值的 99%以上，那么只需要提取前 29 个特征值对应的特征向量即可。这样就构成了一个 1 000 000×29 的转换矩阵，然后用原来的样本乘以这个转换矩阵，就可以得到原来的样本数据在新的特征空间的对应的坐标。将样本的维数降为 30×29，这样原来的训练样本中每个样本的特征值的个数就降到 29 个。

一般来说，PCA 降维后的每个样本的特征的维数不会超过训练样本的个数，因为超

出的特征是没有意义的。虽然目前比较火的图像、视频、文字等多媒体数据大多是在一个非线性的流形上或者附近,PCA 作为一种线性的降维方法可能很少使用,但是 PCA 依然是不可替代的,比如,2015 年 IEEE VIS 年会中的可视分析(VAST2015)最佳论文就颁给一个使用 PCA 做动态网络分析的工作。

2. 多维尺度分析(MDS)

MDS 是分析研究对象的相似性或差异性的一种多元统计分析方法。MDS 主要考虑的是成对样本间的相似性,主要思想是用成对样本的相似性构建合适的低维空间,使得样本在低维空间的距离和高位空间中的距离尽可能保持一致。根据样本是否可计量,可分为 Metric MDS 和 Nonmetric MDS。目标函数是每对低维空间中的欧几里得距离与高维空间中相似度差的平方和,最小化目标函数,用一些数值优化方法得到最优解。

3.2.2 非降维方法

非降维方法保留了高维数据在每个维度上的信息,可以展示数据的所有维度。各种非降维方法的主要区别在于如何对不同的维度进行数据到图像属性的映射。当维度数量较少时,可以直接通过与位置、颜色、形状等多种视觉属性相结合的方式对高维数据进行编码,例如,在形状、大小、颜色上映射数据维度的小图标方法,或用不同角度映射数据维度呈放射形状的星形图(star plots)。但当维度数量增多,数据量变大,或对数据呈现精度的需求提高时,这些方法往往难以满足需要。在处理科学、社会研究和应用中的复杂高维数据时,需要伸缩性(scalability)更强的高维数据可视化方法,包括散点图矩阵(scatterplot matrix)和平行坐标(parallel coordinates)等。

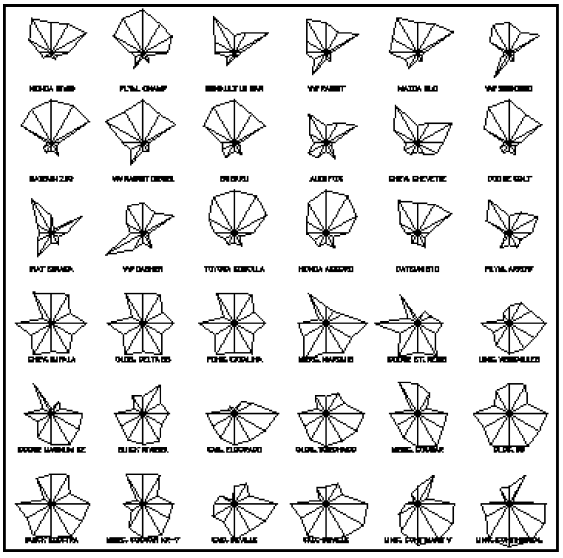
1. 星形图

星形图,又称为雷达图(radar chart)。星形图可以看成平行坐标的极坐标版本。多元数据的每个属性由一个坐标轴表示,所有坐标轴连接到共同的原点(圆心),其布局沿圆周等角度分布,每个坐标轴上的点的位置由数据对象的值与该属性最大值的比例决定,用折线连接所有坐标轴上的点,围成一个星形区域。星形区域的形状和大小反映了数据对象的属性。

星形图提供了一种比较紧凑的数据可视化。随着数据维度的增加,可视化所占的圆形区域内需要显示更多的坐标,但是,其总面积并不变。由于人类视觉识别对形状和大小的敏感性,星形图能使得不同数据对象之间的比较更加容易和高效。

图 3-3 左边展现了汽车数据 12 个数字变量的星形图,这 12 个变量按照图 3-3 右边的变量赋值键所示进行排列。图 3-3 右边的为星形图的变量赋值键;其侧面和底部的变量与大小相关联,其他变量与价格和性能相关联。图 3-3 中每颗星代表一种汽车模型;星中的每条射线长度与各个变量值成正比。对于汽车数据而言,一个变量有较大的值(即较长的射线)则可能意味着这是一辆好车,如 PRICE、TURN 和 GRATIO 等变量。在图 3-3 中最显著的标示就是:在顶部的星形图中顶部的价格和性能(price and performance)变量射线较长,并且底部的尺寸(size)变量射线较短;但反过来这样理解也是对的:最重的

模型在星形图的底部一行。



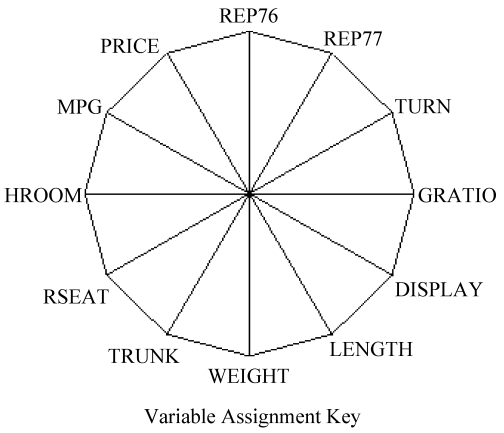


图 3-3 汽车数据星形图

图片来源：https://www.webpages.uidaho.edu/~stevel/519/R/Statistical%20Graphics%20for%20Multivariate%20Data.pdf

图 3-4 展现了美国 50 个州以及首都华盛顿哥伦比亚特区的犯罪率。在右图的范例中，可以看到在星形图中 7 个坐标分别表示 7 种类型的犯罪（车辆失窃、盗窃、抢劫、谋杀和强奸等）。范例中展示的乔治亚州除了强奸发生率低之外，其余各种犯罪率都较高。在左图中，代表所有 50 个州和华盛顿特区的星形图按照顺序依次排列，用户可以通过比较

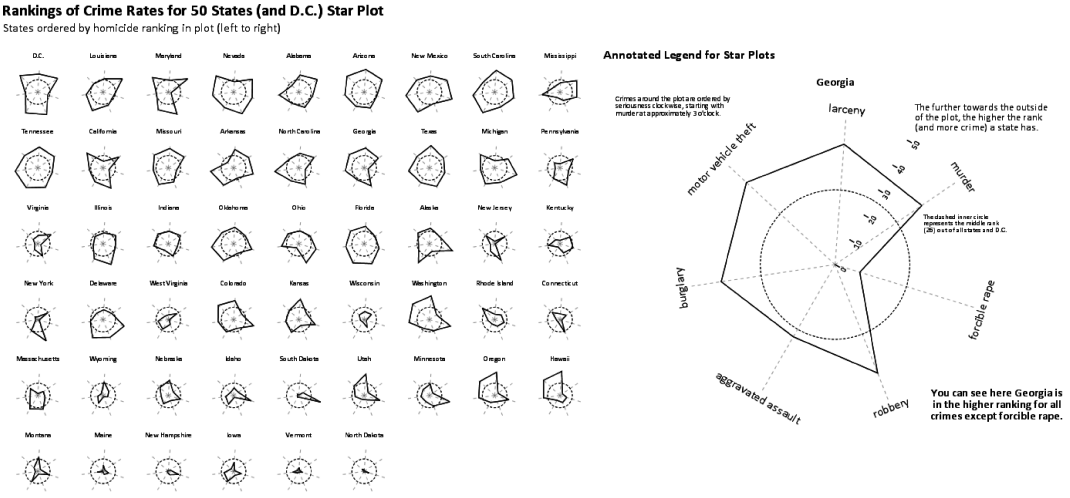


图 3-4 美国 50 个州以及首都华盛顿特区的犯罪率星形图

图片来源：https://www.webpages.uidaho.edu/~stevel/519/R/Statistical%20Graphics%20for%20Multivariate%20Data.pdf

不同星形区域的大小和形状,了解各州的犯罪情况。例如,北达科他州是各类犯罪最少的州,而南达科他州除了强奸发生率较高之外,其余犯罪率都较低。

2. 切尔诺夫脸谱图

切尔诺夫脸谱图(Chernoff Faces)是由美国统计学家切尔诺夫在1976年率先提出的,用脸谱分析多维度数据,即将 P 个维度的数据用人脸部位的形状或大小表示。此方法和星形图类似,也采用图标表示单个多元数据对象,不同的是,切尔诺夫脸谱图采用模拟人脸的图标表示数据对象,它可以把多元数据用二维的人脸的方式整体表现出来。各类数据变量经过编码后,转变为脸型、眉毛、眼睛、鼻子、嘴、下巴等面部特征,数据整体就是一张表情各异的人脸。例如,图3-5展示了使用切尔诺夫脸谱图可视化同样的美国各州犯罪率数据的例子,其中脸的长度表示谋杀案的发生率,脸的宽度表示强奸案的发生率,从图中可以明显地观察到阿拉斯加州(Alaska)的强奸案发生率较高,华盛顿哥伦比亚特区(Washington D.C.)的谋杀案发生率较高。

切尔诺夫脸谱图的灵感来源于,当人们面对错综复杂的信息时,人脑会自动过滤掉无用信息,保留有用信息。人脑通常可以察觉到一些非常细微甚至难于测量的变化,然后对其做出反应,同时,人脑区分脸谱时,这种优越性更加明显,因为无论是脸的胖瘦还是五官的大小和位置,都极易给人留下深刻的印象,因而易于区别。但人们对于人脸上各个部分或者特征的感知程度不同,所以需要根据数据分析的目的和属性的优先级别选择合适的属性与某个人脸特征之间形成映射。

3. 散点图

散点图(Scatterplot)的本质是将抽象的数据对象映射到二维的直角坐标系表示的空间。数据对象在坐标系的位置反映了其分布特征,直观、有效地揭示两个属性之间的关系。面向多元数据,散点图的思想可泛化为:采用不同的空间映射方法将多元数据对象布局在二维平面空间中,数据对象在空间中的位置反映了其属性及相互之间的关联,而整个数据集在空间中的分布则反映了各个维度之间的关系及数据集的整体特性。图3-6表达的是世界各国预期寿命与人均国内生产总值的散点图可视化。每个数据对象(圆点)代表一个国家,圆点的大小表示国家人口,圆点的颜色表示国家所在的洲。

散点图矩阵是散点图的高维扩展,它从一定程度上克服了在平面上展示高维数据的困难,在展示多维数据的两两关系时有着不可替代的作用。散点图矩阵是由所有两两维度间的散点图按矩阵形式排列而成的,每个散点图都表现出两个维度间的关系。通过散点图矩阵,可以看到数据在任意两个维度间的相关特性以及聚类情况。它的缺点是不能显示各个数据在多个维度上的协同关系,同时需要很大的显示空间,需要的显示空间的面积正比于维度数目的平方。

图3-7是以鸢尾花数据为例,利用R语言graphics包中的pairs()函数绘制的散点图矩阵。图中不同颜色分别代表不同品种的鸢尾花,从图中可以观察到各类鸢尾花的花瓣、花萼长宽的大体分布以及它们两两之间的关系。

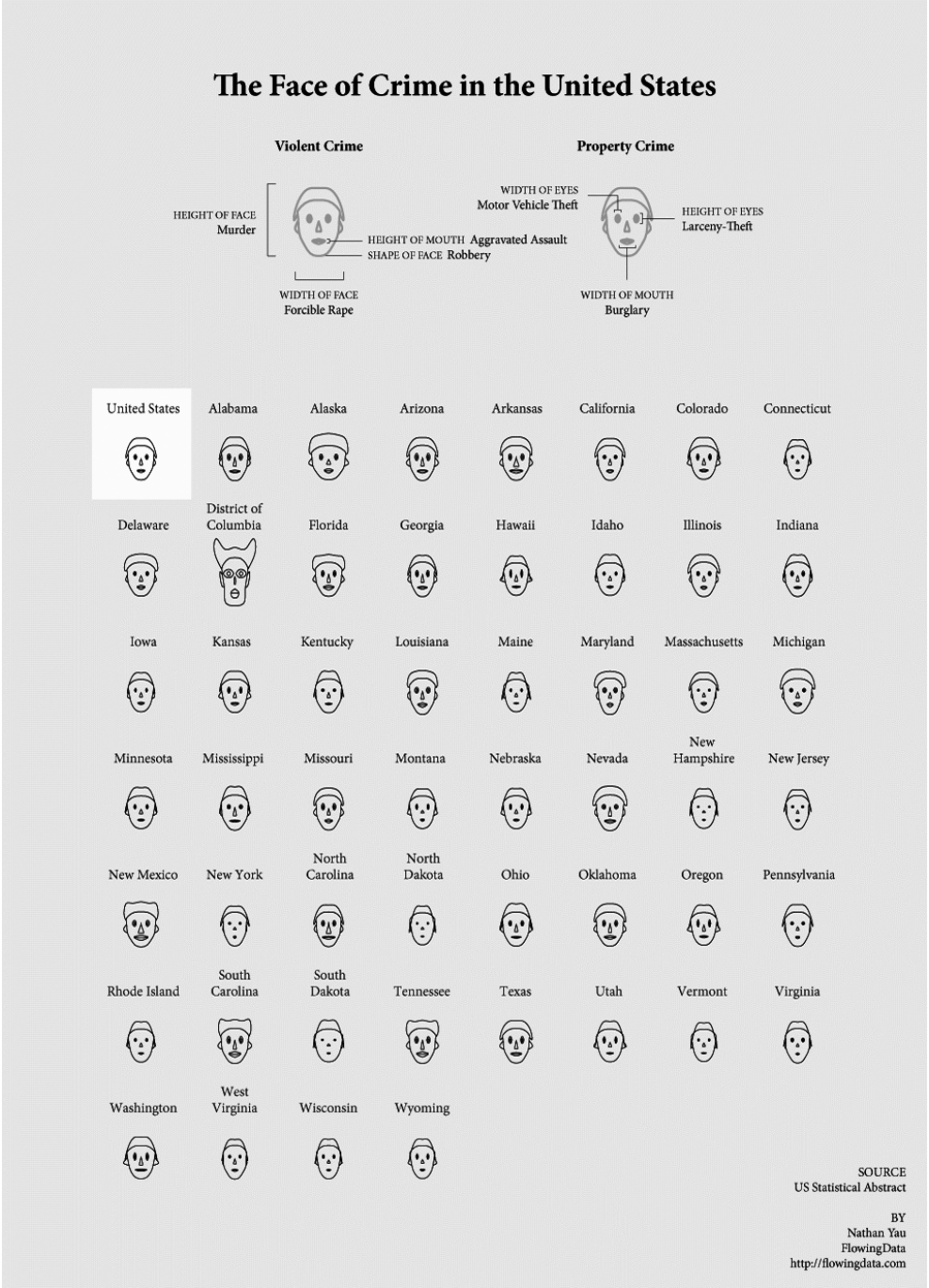


图 3-5 Chernoff Faces

图片来源：http://flowingdata.com/2010/08/31/how-to-visualize-data-with-cartoonish-faces/crime-chernoff-faces-by-state-edited-2/

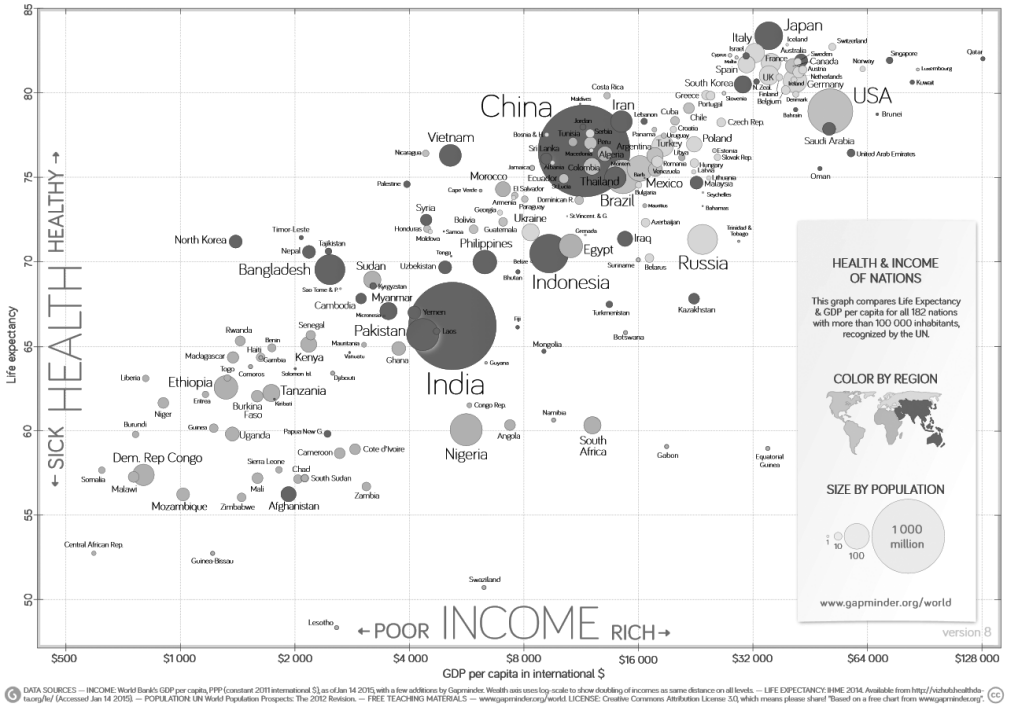


图 3-6　世界各国预期寿命与人均国内生产总值的散点图可视化

图片来源：http://www.gapminder.org(见彩图)

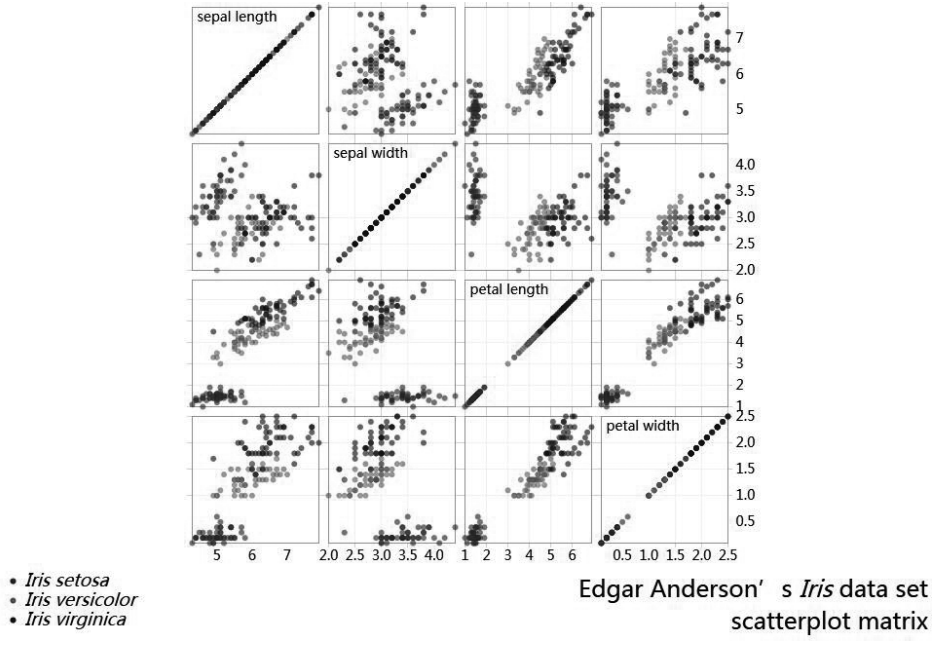


图 3-7　鸢尾花散点图矩阵

图片来源：https://github.com/d3/d3/wiki/Gallery(见彩图)

4. 平行坐标

平行坐标(Parallel Coordinates)是将高维数据的各个变量维度用一系列相互平行的坐标轴表示,变量值对应轴上的位置。将描述不同变量维度的同一数据对应各点连接成折线,代表一个数据的一条折线在平行的坐标轴上的投影就反映了变化趋势和各个变量维度间的相互关系。平行坐标能够帮助分析数据在多个维度上的分布和多个维度之间的关系,且平行坐标需要的显示面积仅正比于维度的数目。由于每个数据点都表现为一条折线,所以平行坐标在两个维度之间关系的表现不如散点图清楚,并且当数据量增大时更容易受到图元堆叠的影响。

图 3-7 中的例子使用的鸢尾花数据也可以用平行坐标系表示,如图 3-8 所示。从图 3-8 中可以看到,鸢尾花的 Iris setosa 类花萼长度主要分布于 4.3～5.8cm,花萼宽度主要分布于 2.9～4.4cm,但也有一个例外,其花萼宽度为 2.3cm,这种情况可以猜测到是植物中的突变或者异常;花瓣长度主要分布于 1.0～1.9cm,花瓣宽度主要分布于 0.1～0.6cm;其他两种鸢尾花的情况也可以从图中观察到。除此之外,还可以从图中观察出它们两两之间更多的联系。

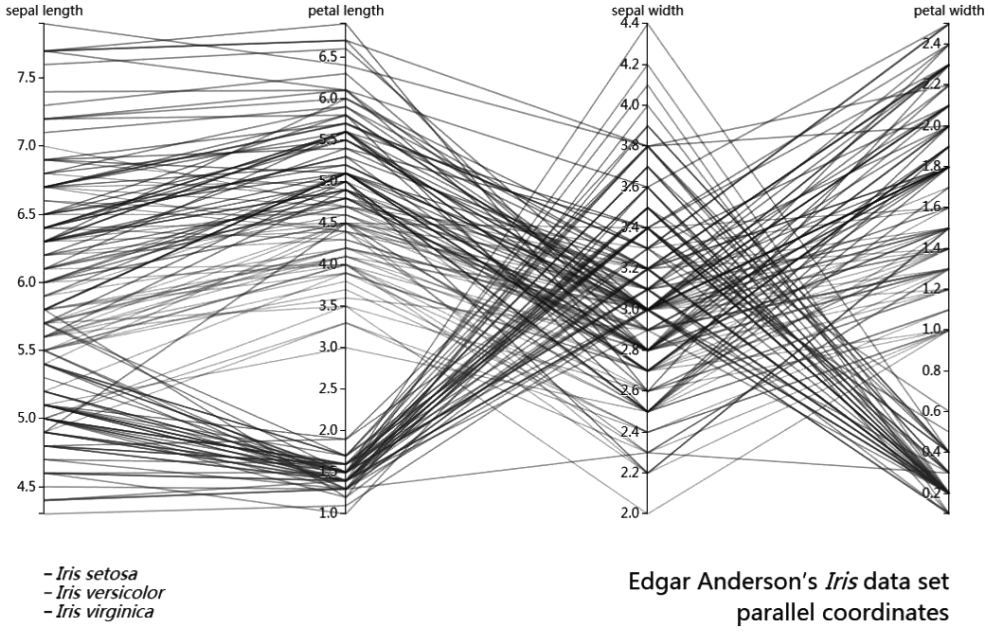


图 3-8 鸢尾花平行坐标系

图片来源:https://github.com/d3/d3/wiki/Gallery

图 3-9 为在平行坐标中结合散点图。图中的两个属性 Horsepower 和 Displacement 的关系在两个轴之间用散点图表示。

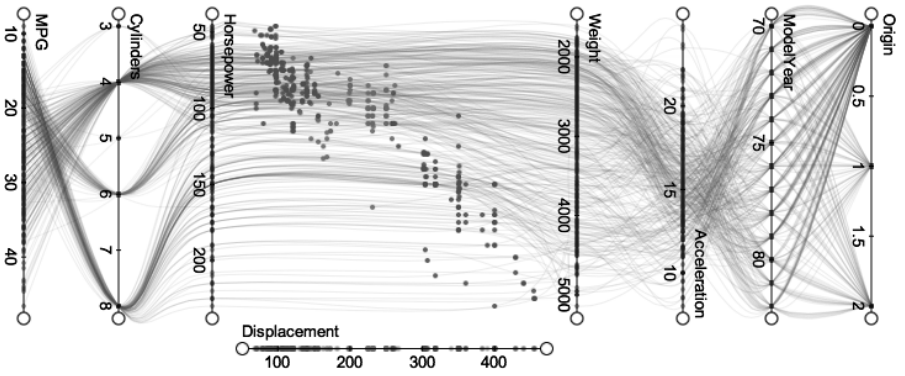


图 3-9　平行坐标结合散点图

图片来源：http://vis.pku.edu.cn/mddv/val/gallery(见彩图)

3.3　网络数据可视化

网络可视化充分利用视觉感知系统，将网络数据以图形化的形式展现出来，快速直观地解释及概览网络结构数据。一方面辅助用户认识网络的内部结构；另一方面有助于挖掘隐藏在网络内部中的有价值信息。本节针对网络拓扑类型数据介绍几种对应的可视化方法，在 3.3.1 节中介绍常见的结点-链接法，在 3.3.2 节中介绍相邻矩阵布局，在 3.3.3 节中介绍多种布局的混合应用。

互联网已经成为信息的主要载体之一，基于网络的研究在数据获取方面越来越容易，网络数据分析是伴随着现代网络技术的应用出现的一个较新的研究领域，它具有复杂性、多维度和时变性等特点。人们已经习惯在网络上分享信息，结识新的朋友等；基于微信、微博等社交 App，可以迅速发布身边正在发生的事情，因此社交媒体相对传统媒体(如报纸、电视新闻等)具有很好的竞争力。企业可以对巨大的网络数据进行挖掘分析并发现潜在的商业价值，甚至能通过基于网络的各种平台直接影响客户，客户同样可以从网络数据中获取信息了解公司的方方面面，以达到指导和决定投资的目的。

相较于树型数据中明显的层次结构，网络数据并不具有自底向上或自顶向下的层次结构，表达的关系更加复杂和自由。图的绘制包括 3 方面：网络布局、网络属性可视化和用户交互，其中最核心的要素是网络布局决定图的结构关系。最常用的网络布局方法有结点-链接法和相邻矩阵两类。

3.3.1　结点-链接法

在可视化布局中最自然的表达就是用结点表示对象，用线或者边表示关系的结点-链接布局(node-link)。它容易被用户理解、接受，帮助人们快速建立事物与事物之间的联系，显式地表达事物之间的关系，例如关系型数据库的模式表达、地铁线路图的表达，因而是网络数据可视化的首要选择。图 3-10 就是结点-链接法的一个例子。

结点-链接法的优点有：比较直观地反映网络关系；能够表现图的总体结构、簇、路

径;灵活,有许多的变种。结点-链接法的局限性有:几乎所有直观算法的复杂度均大于$O(N^2)$;对于密集(尤其是关系密集)的图不是很适用。

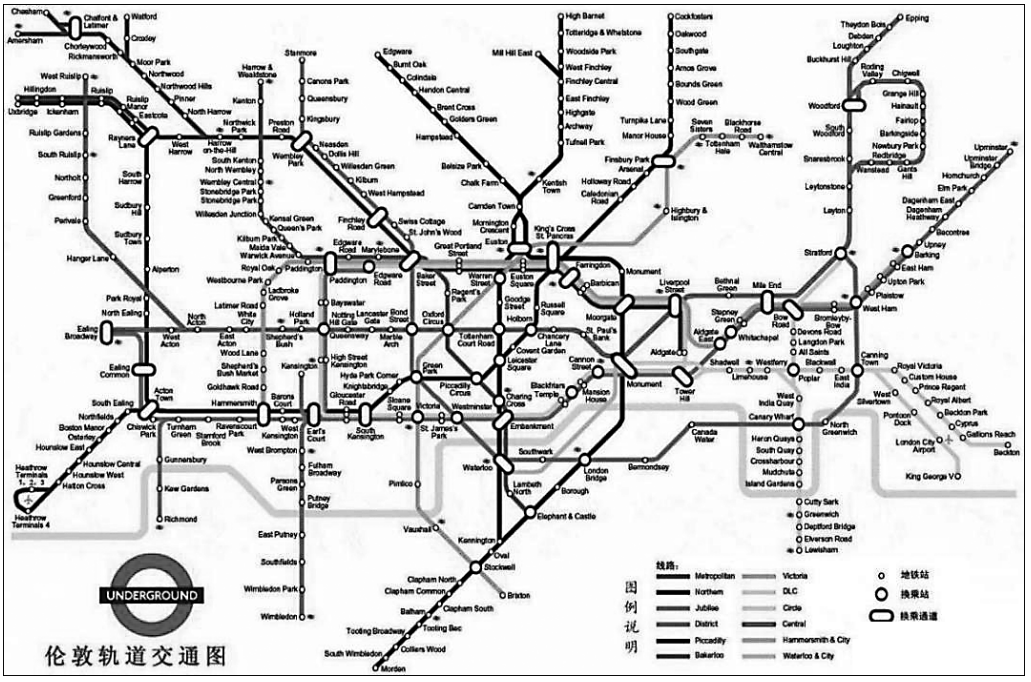


图 3-10 伦敦地铁图

1. 力引导布局

力引导布局最早由 Peter Eades 在 1984 年的《启发式画图算法》一文中提出,目的是减少布局中边的交叉,尽量保持边的长度一致。此方法借用弹簧模型模拟布局过程:用弹簧模拟两个点之间的关系,受到弹力的作用后,过近的点会被弹开,而过远的点被拉近;通过不断的迭代,使整个布局达到动态平衡,趋于稳定。其后,“力引导”的概念被提出,演化成力引导布局算法(Fruchterman-Reingold 算法,简称 FR 算法),即一种用于丰富两点之间关系的物理模型,加入点之间的静电力,通过计算系统的总能量并使得能量最小化,从而达到布局的目的。这种改进的能量模型可看成弹簧模型的一般化。

无论是弹簧模型还是能量模型,其算法的本质是要解决能量优化问题,区别在于优化函数的组成不同。优化对象包括引力和斥力部分,不同算法对引力和斥力的表达方式不同。对于平面上的两个结点 i 和 j,用 $d(i,j)$表示两个点的欧几里得距离,$s(i,j)$表示弹簧的自然长度,k 是弹力系数,r 表示两个点之间的静电力常数,w 是两个点之间的权重。

弹簧模型为

$$E_s = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} k[d(i,j) - s(i,j)]^2$$

能量模型为

$$E = E_s + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{rw_i w_j}{d(i,j)^2}$$

力引导布局易于理解和实现,可以用于大多数网络数据集,而且实现的效果具有较好的对称性和局部聚合性,因此比较美观。然而,力引导布局只能达到局部优化,而不能达到全局优化,并且初始位置对最后优化结果的影响较大。

算法 3-1 力引导布局的伪代码

设置结点的初始速度为(0,0);设置结点的初始位置为任意但不重叠的位置。

1: 开始循环

2: 总动能 GE=0 //所有粒子的总动能为 0

3: 对于每个结点 i 净力 f=(0,0)

4: 对于除该结点外的每个结点 j,净力 f=净力 f + j 结点对应 i 结点的库仑斥力

5: 下一个结点 j+1

6: 对于该结点上的每个弹簧 s,净力 f=净力 f+弹簧对该结点的胡克弹力

7: 下一个弹簧 s+1

8: //如果没有阻尼衰减,整个系统将一直运动下去

该结点速度=(该结点速度 + 步长 * 净力) * 阻尼

该结点位置= 该结点位置 + 步长 * 该结点速度

总动能:=总动能 + 该结点质量 *(该结点速度)^2

9: 下一个结点 i+1

一般来说,整个算法的时间复杂度为 $O(n^3)$,迭代次数与步长有关,一般认为是 $O(n)$,每次迭代都要两两计算点之间的力和能量,复杂度是 $O(n^3)$。为了避免达到动态平衡后反复振荡,也可以在迭代后期将步长调到一个比较小的值。

例如,图 3-11 展示了一个小型的社交网络。社交网络是一个由个人或社区组成的点

图 3-11 力引导布局-小型社交网络

图片来源: http://bl.ocks.org/mbostock/950642

状网络拓扑结构。其中每个点（Node）代表一个个体，可以是个人，也可以是一个团队或是一个社区，个体与个体之间可能存在各种相互依赖的社会关系，在拓扑网络中以点与点之间的边（Tie）表示。而社交网络分析关心的正是点与边之间依存的社会关系。随着个体数量的增加以及个体间社会关系的复杂化，最后形成的整个社交网络结构可能会非常复杂。图 3-12 为法国作家维多克·雨果的小说《悲惨世界》的人物谱图（力引导布局）。结点用颜色对通过子群划分算法计算的人物分类类别进行编码，边的粗细表示两个结点代表的人物之间共同出现的频率。

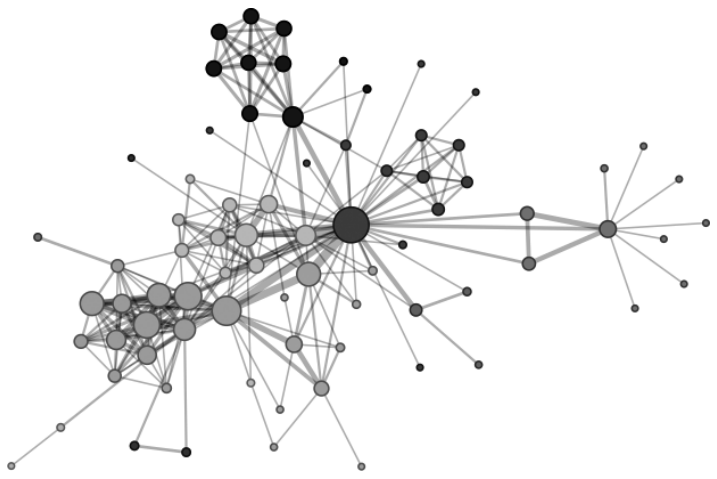


图 3-12 法国作家维多克·雨果的小说《悲惨世界》的人物谱图（力引导布局）

图片来源：http://homes.cs.washington.edu/~jheer///files/zoo/

2. 多维尺度分析布局

多维尺度分析（MDS）是一种探索性数据分析技术，主要是用适当的降维方法将多个变量通过坐标定位在低维空间（二维或三维）中，变量之间的欧几里得距离就可以反映它们之间的差异性和相似性。MDS 布局的出现是为了弥补引导布局的局限性。它针对高维数据，用降维方法将数据从高维空间降到低维空间，力求保持数据之间的相对位置不变，同时也保持布局效果的美观性。力引导布局方法的局部优化使得在局部点与点之间的距离能够比较真实地表达内部关系，但却难以保持局部与局部之间的关系。相对地，MDS 是一种全局控制，目标是要保持整体的偏离最小，这使得 MDS 的输出结果更加符合原始数据的特性。

MDS 根据数据集特征分为不考虑个体差异 MDS 模型和考虑个体差异 MDS 模型。MDS 模型允许多种类型的数据输入，并且在实际应用中，也有多种测量相似性或差异性的方法，根据分析数据的类型分为以下两种：

（1）度量化 MDS 模型。也称为古典 MDS 模型，输入的数据是直接反映变量间差异或相似的距离或比率，例如，城市间的距离就是现成的反映差异的数据。

（2）非度量化 MDS 模型。输入的数据不是直接反映变量间的差异，而是通过对其属

性的评分,间接地反映变量间的差异或相似性。

多维尺度分析的步骤如下:

(1) 界定问题。明确研究的问题和范畴,确定相关的变量种类和数量。

(2) 获取数据。根据实际情况获取分析数据。

(3) 选择 MDS 模型。根据获得的数据类型,选择相应的 MDS 模型。

(4) 确定维度。MDS 模型是为了生成一个用尽可能小的维度对数据进行最佳拟合的空间感知图,因此要确定一个合适的维度,维度太高不易于解读,维度太低会影响拟合度,通常采用二维或三维。

(5) 模型评价。考察应力系数 Stress 和拟合指数 RSQ,应力系数越小越好,拟合指数越大越好。

(6) 解读图表。多维尺度分析最重要的结果是感知图,图中各点之间的距离直接反映各变量的相似或差异程度,除了查看差异程度之外,如果要对图表进行整体的分析解读,还需要对每个维度进行解释。

3. 弧长链接图

弧长链接图(Arc Diagram)是结点-链接法的一个变种。它将结点沿某个线性轴或环状排列,圆弧表达结点之间的链接关系,例如,图 3-13 是法国作家维多克·雨果的小说《悲惨世界》的人物谱图(弧长链接图)。图 3-14 是 2011 年年末欧债危机时 BBC 新闻制作的各国之间错综复杂的借贷关系的可视化。各个国家被放置在圆环布局上,曲线的箭头表示两国之间的债务关系,箭头的粗细表示债务的多少,颜色标识了债务危机的严重程度,每个国家外债的数额决定了它们在圆环上所占的大小。由于债务关系太复杂,所以只有用户单击选中国家的债务关系才被标注出来。欧洲最严重的是希腊,外债达到国民生

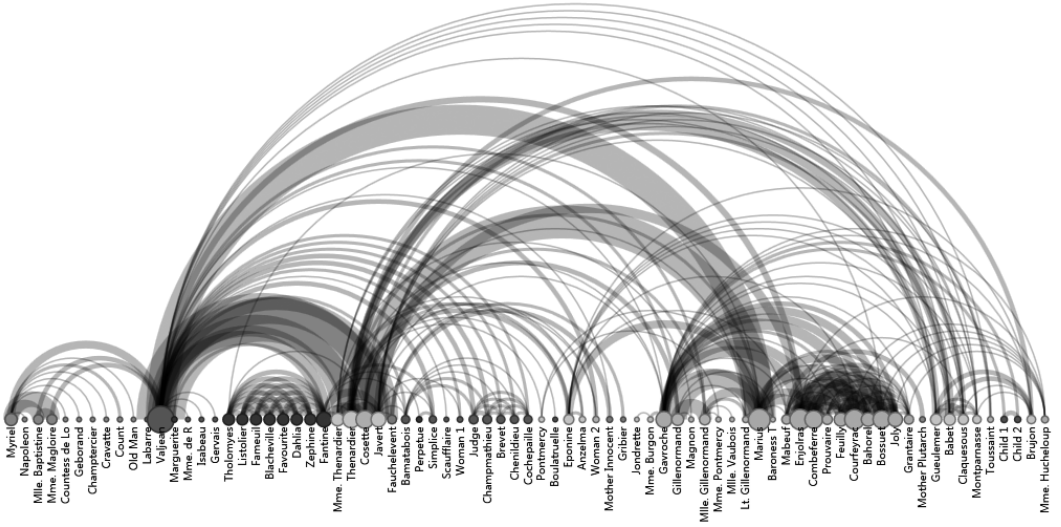


图 3-13　法国作家维多克·雨果的小说《悲惨世界》的人物谱图(弧长链接图)

图片来源: http://homes.cs.washington.edu/~jheer///files/zoo/

产总值的两倍。美国外债最多，主要的债权国是法国。弧长链接图不能像二维布局那样表达图的全局结构，但在结点良好排序后可清晰地呈现环和桥的结构。对结点的排序优化问题又称为序列化，在可视化、统计等领域有广泛的应用。

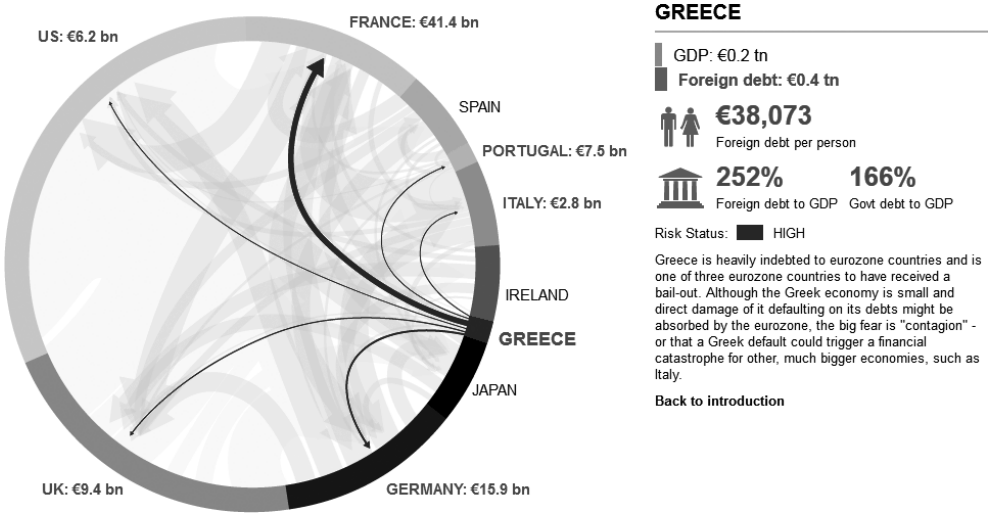


图 3-14 2011 年年末欧债危机时 BBC 新闻制作的各国之间错综复杂的借贷关系的可视化

图片来源：http://www.bbc.co.uk/news/business-15748696

3.3.2 相邻矩阵布局

相邻矩阵(Adjacency Matrix)指代表 N 个结点之间关系的 $N\times N$ 的矩阵，矩阵内的位置(i,j)表达了第 i 个结点和第 j 个结点之间的关系。对于无权重的关系网络，用 0-1 矩阵(Binary Matrix)表达两个结点之间的关系是否存在；对于带权重的关系网络，相邻矩阵则可用(i,j)位置上的值代表其关系紧密程度；对于无向关系网络，相邻矩阵是一个对角线对称矩阵；对于有向关系网络，相邻矩阵的对角线表达结点与自己的关系。相邻矩阵关系图的示例如图 3-15 所示。

相邻矩阵布局能够完全规避边的交叉，所以非常适用于密集的图。其视觉伸缩性强；能更好地表达两两关联的网络数据。但同时它的可视化结果比较抽象，难以呈现网络的拓扑结构。另外，它不能直观地表达关系传递性和网络中心性，难以跟踪出路径。

相邻矩阵的表达简单易用，可以用数值矩阵，也可以将数值映射到色彩空间表达。但是，从相邻矩阵中挖掘出隐藏的信息并不容易，需要结合人机交互。最关键的两种交互是排序和路径搜索，前者使具有相似模式的结点靠得更近，而后者则用于探索结点之间的传递关系。图 3-16 为相邻矩阵法实例，是城市交易相邻矩阵，横向和纵向分别表示收货和发货的关系。饱和度越低，颜色越白，表示交易值越小；颜色越深，表示交易值越大。

相邻矩阵排序的意义是凸显网络关系中存在的模式。类似于弧长链接图，这个问题也称为序列化问题。一个 $N\times N$ 的相邻矩阵共有 $n!$ 种排序方式，在这 $n!$ 种组合中找到使代价函数最小的排序方式称为最小化线性排列，是一个 N-P 难度的问题。在实际应

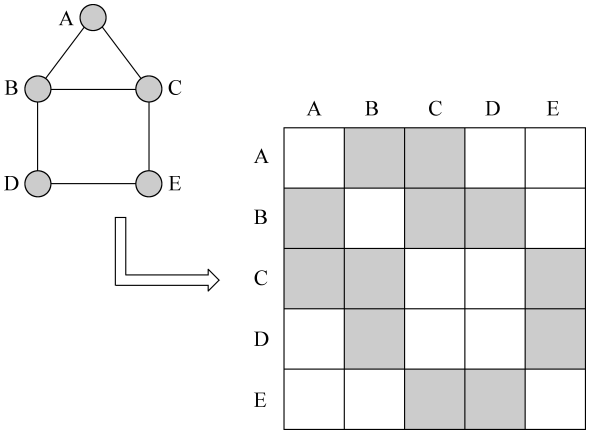


图 3-15　相邻矩阵关系图

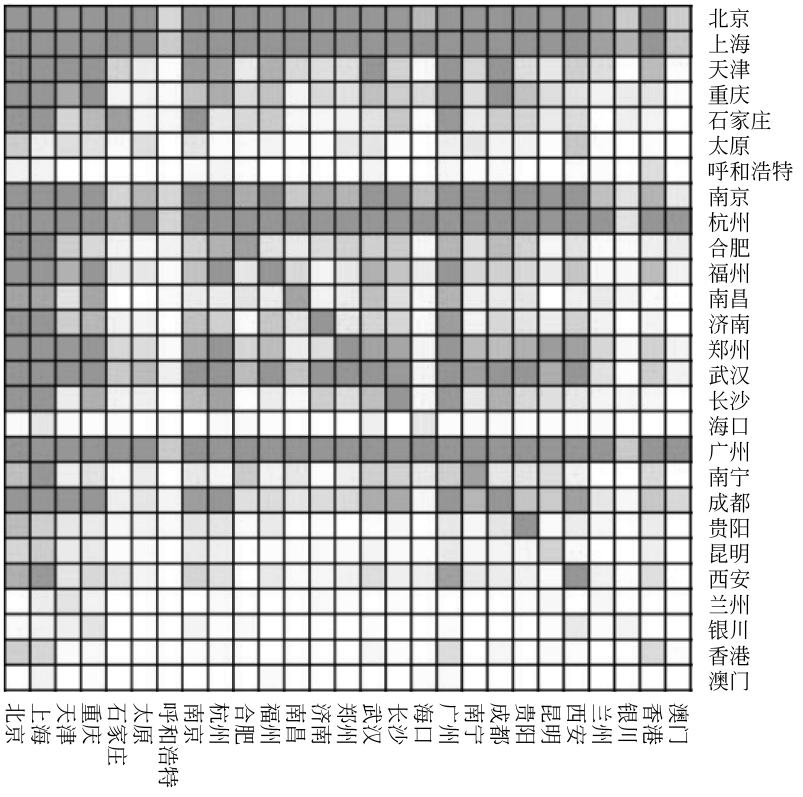


图 3-16　城市交易相邻矩阵

用中,通常采用启发式算法,不求达到最优。

常规的排序方法依据网络数据的某一数值(矩阵值或结点的度)的大小执行。选择不同的排序项产生不同的矩阵排序结果。图 3-17 为法国作家维多克·雨果的小说《悲惨世界》的人物谱图(相邻矩阵)。图中采用子群聚类算法获得的人物分类结果对相邻矩阵的

行和列进行排序。

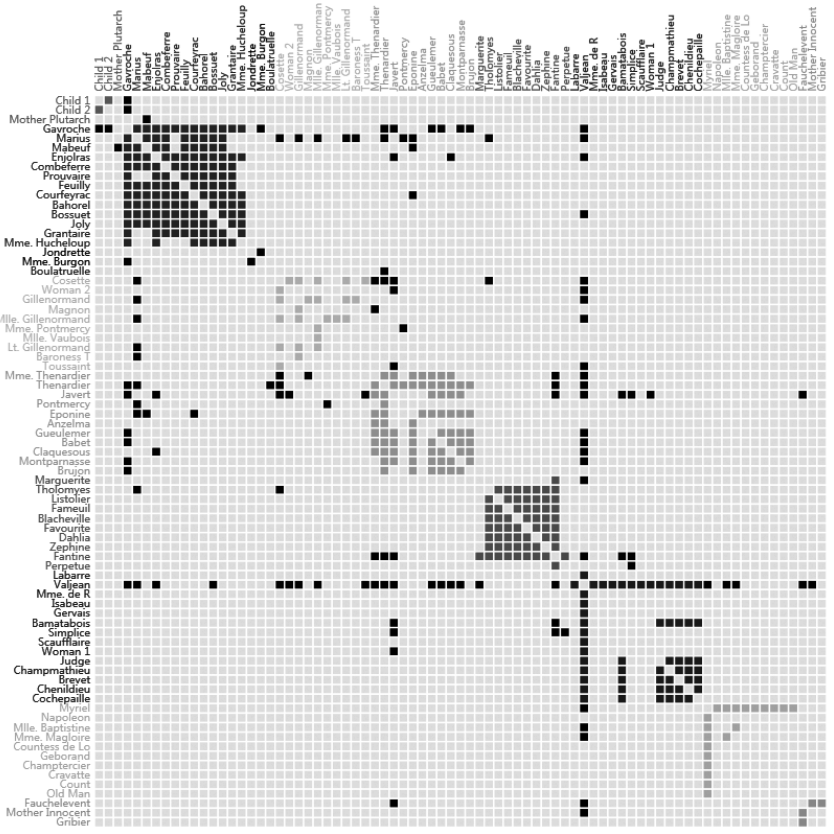


图 3-17 法国作家维多克·雨果的小说《悲惨世界》的人物谱图(相邻矩阵)

图片来源：http://homes.cs.washington.edu/~jheer///files/zoo/

3.3.3 混合布局

结点-链接布局适用于结点规模大但边关系较为简单，并且能从布局中看出图的拓扑结构的网络数据；而相邻矩阵恰恰相反，适用于结点规模较小，但边关系复杂，甚至是两两结点之间都存在关系的数据。这两种数据的特点是用户选择布局的首要区分原则。对于部分稀疏、部分稠密的数据，单独采用任何一种布局都不能良好地表达数据，这时便可以混合两者的布局来设计。一些人认为不同的布局各有所长，取长补短的混合布局永远优于单一布局的选择，但实际上顾此失彼的例子常常见到。是否需要混合布局、如何设计混合布局必须经过仔细思考。

是否一种布局已经足以表达数据的模式？无须为了混合而混合，尤其不应为了花哨的可视效果而混合。可视化是一个科学严谨的过程，如果增加布局并不能表达重要的额外信息特征，或者另一种布局的可视化效果重叠，那么混合布局不是一种好的选择。

多种布局如何混合成一种布局？简单的多种布局罗列称为仪表盘(dashboard)或多

视图模式,即将不同的布局结果放置于同一个页面,在视图之间实现可视交互的同步。根据可视化布局组合研究,除并列式组合外还有载入式、嵌套式、主从式、结合式4种组合模式,对可视化布局更丰富的组合能大大提高可视化结果的分析效率。

3.4 可视化案例分析

在本节中将结合两个实际案例介绍可视分析的实际应用。在3.4.1节中介绍在完成China VIS 2015比赛中所采用的一些可视化方法以及分析过程。在3.4.2节中介绍在完成VAST Challenge 2016比赛中所采用的一些可视化方法以及分析过程。

3.4.1 案例一:China VIS 2015竞赛题

1. 简介

本题关注对正常网络流量日志数据的分析。TSZNet公司将提供一周的tcpflow日志数据,希望参赛者找到公司内部网络的客户端和服务器,分析公司内网的网络体系结构,总结公司内网的正常网络通信模式有哪些。

2. 数据

一周的tcpflow(TCP协议层的数据传输记录)日志,正常流量。tcpflow日志字段包括time(时间)、sip(源IP)、dip(目的IP)、proto(协议)、dport(目的端口)、uplen(上行字节数)、downlen(下行字节数)。

3. 问题

(1) 找出TSZNet公司内部网络中的客户端与服务器,并给出该公司网络的体系结构拓扑图。

(2) 对TSZNet公司内部网络中的服务器进行分类,分类标准不限,如按照功能、时间特点、行为特点、流量特点等。

(3) 说明TSZNet公司内部网络的客户端-服务器、客户端-客户端、服务器-服务器之间的常规通信模式。

4. 解答

(1) 通过对一周tcpflow数据的可视分析,找出TSZNet公司内部网络中的客户端与服务器,并给出该公司网络体系结构拓扑图。

首先,根据日志记录中内网与内网之间的通信数据,用力导向图表示内网之间的通信(如图3-18所示),每一个结点代表一个主机。用连线表示两台主机之间有通信,连线的宽度代表两个结点之间的流量大小,其中颜色表示使用最多的通信协议,结点的大小表示流量。从图3-18可以看出,有一些主机与很多台主机都有连接关系,并且流量大于大多数主机,那么可以认为该主机可能为服务器。接着依次找出可能是服务器的主机,如图中

黑框内所示，然后进行下一次筛选。

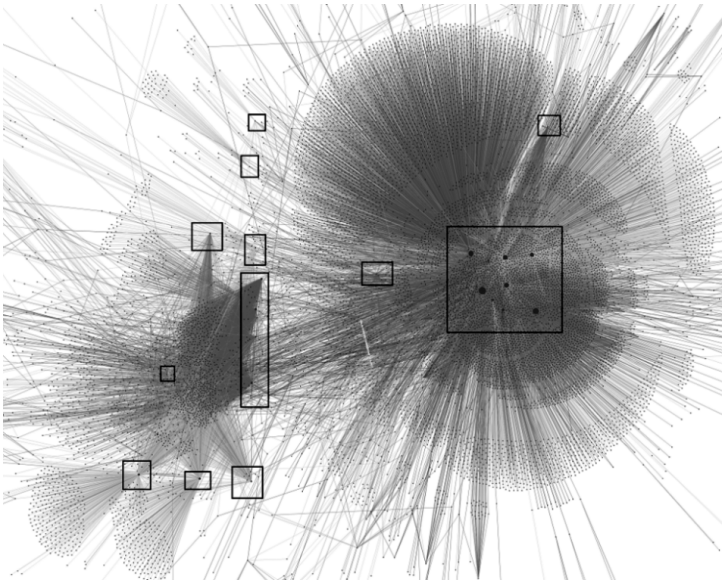


图 3-18 内网通信的力导向图

其次，根据第一次筛选结果再次画出通信网络图（如图 3-19 所示），然后从中去除孤立的点和度小于平均度的结点，剩下的基本可以确定为服务器。

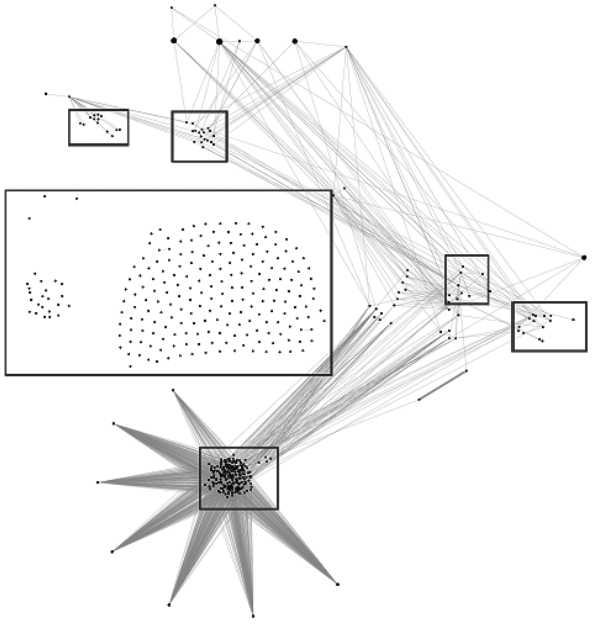


图 3-19 通信网络的力导向图

图 3-20 表示内网主机与服务器之间的通信关系，其中白色的结点表示客户端的 IP