

图分析与可视化

在关联数据中发现商业机会

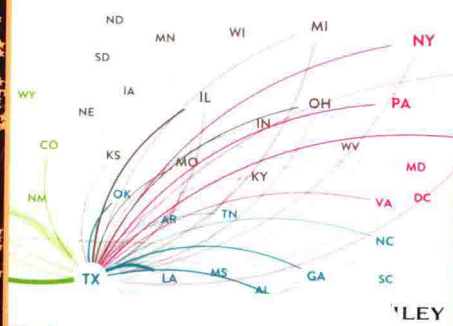
[美] 理查德·布莱斯 (Richard Brath) 著
大卫·琼克 (David Jonker)

赵利通 译

Graph Analysis and Visualization

Discovering Business Opportunity in Linked Data

Richard Brath and David Jonker



GRAPH ANALYSIS AND VISUALIZATION

Discovering Business Opportunity in Linked Data



机械工业出版社
China Machine Press

数据科学与工程技术丛书

GRAPH ANALYSIS AND VISUALIZATION

Discovering Business Opportunity in Linked Data

图分析与可视化

在关联数据中发现商业机会

[美] 理查德·布莱斯 (Richard Brath) 著
大卫·琼克 (David Jonker)

赵利通 译



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

图分析与可视化：在关联数据中发现商业机会 / (美) 布莱斯 (Brath, R.), (美) 琼克 (Jonker, D.) 著; 赵利通译. —北京: 机械工业出版社, 2016.1
(数据科学与工程丛书)

书名原文: Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data

ISBN 978-7-111-52692-6

I. 图… II. ① 布… ② 琼… ③ 赵… III. 商业信息 - 数据管理 IV. F713.51

中国版本图书馆 CIP 数据核字 (2016) 第 015256 号

本书版权登记号: 图字: 01-2015-2818

Copyright © 2015 by John Wiley & Sons, Inc., Indianapolis, Indiana

All Rights Reserved. This translation published under license. Authorized translation from the English language edition, entitled Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data, ISBN 978-1-118-84584-4/1118845846, by Richard Brath | David Jonker, Published by John Wiley & Sons. No part of this book may be reproduced in any form without the written permission of the original copyrights holder.

本书中文简体字版由约翰·威利父子公司授权机械工业出版社独家出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

本书封底贴有 Wiley 防伪标签, 无标签者不得销售。

图分析与可视化：在关联数据中发现商业机会

出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)	
责任编辑: 刘诗灏	责任校对: 董纪丽
印刷: 中国电影出版社印刷厂	版次: 2016 年 3 月第 1 版第 1 次印刷
开本: 185mm × 260mm 1/16	印张: 21
书号: ISBN 978-7-111-52692-6	定价: 119.00 元
凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换	
客服热线: (010) 88379426 88361066	投稿热线: (010) 88379604
购书热线: (010) 68326294 88379649 68995259	读者信箱: hzit@hzbook.com

版权所有·侵权必究
封底无防伪标均为盗版
本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

前言

本书将介绍如何把图的可视化与分析应用到商业中。图的应用是一种独特而宝贵的资源，可用于从数据中发现有价值的信息。近年来，世界上一些最具创新力公司内部的分析人员开始积极探索基于图的方法，以更深入地理解他们工作的动态，同时发现可以提高业绩的机会和策略。

随着可用数据的量、种类和速度都在增长，对帮助理解数据的方法和技术的需求也在增长。各种组织已经强烈感受到简单的仪表板风格图表的局限性。仪表板擅长显示指标和趋势，可以告诉你公司哪些部门在什么时候比其他部门表现得更好或更差，但是不能告诉你为什么会这样，而理解“为什么”是采取有效行动的关键。

图的作用是表示两种事物之间的连接，揭示数据关系的结构和本质。关系是理解事物的“为什么”以及“如何做到”的基础，这也是图分析和可视化具有巨大价值潜力的原因之一。

本书作者回顾过去 20 多年为商业和情报分析人员设计与构建新应用的经历，意识到图已经在许多解决方案中扮演了一种角色。如今，我们的一些最重要的研究和软件开发工作在本质上都是基于图的。

然而，尽管图十分有用，但在科学界以外却很少有图的应用，关于图设计的作品就更少了。随着开源图工具和库的能力在近期不断发展，图已经可被每个商业分析师使用，但是关于图的分析与可视化的有效原则与技术的知识，仍然只有少数人知道。我们撰写这本书的目的就是为了帮助改变这种情况。

本书目标读者

本书针对的是希望知道如何将图分析应用到决策相关问题的数据科学家和分析人员。本书中的示例取自商界，但是使用的原则与技术也可用于政府机构和非营利组织。

读者不需要具有关于图论及其实践的知识。新接触图分析的读者可以从头到尾阅读本

书，这样更有帮助。有经验的读者可以选择跳到第Ⅲ部分中感兴趣的主体，该部分详细讨论了分析主题。

本书的一些例子包含少量的编程，但是大部分示例应用都使用鼠标点击类工具。对于这两种情况，都需要有一定程度的技术能力。

本书结构

本书包含 4 个部分。第一部分对图的主题进行了概述。剩余章节逐渐讲解更加具体或者高级的主题。第 3~10 章由 Richard Brath 撰写，其他章节由 David Jonker 撰写。

- 第 I 部分：在本书的第一部分中，作者概述了图在商业中的应用，并介绍了各种类型的图（第 3 章进行了详细描述）。
- 第 II 部分：本书的第二部分全面探讨了图的可视化与分析过程的主要步骤。
- 第 III 部分：本书的第三部分讲解了不同的分析主题及与之相关的图类型与技术。
- 第 IV 部分：本书的第四部分关注高级主题（仍在不断研究中的领域），以及根本的设计原则。

下载材料

本书为各章的示例提供了在线的数据文件、源代码包和图可视化文件，并按章将这些补充材料组织起来^①。查看或者运行这些文件所需的软件在每章的示例中进行了描述。下载文件中包含以下内容：

- **数据文件**：大多数数据文件以通用格式提供，例如文本（.txt）或逗号分隔值（.csv），可以直接读入图软件或者被程序使用。在一些情况中，会有两个文件，一个是节点文件，另一个是边（即节点之间的连接）文件。在其他情况中，以图特定的文件格式来提供图数据文件，例如 .gdf 或 .graphml。这些是许多图工具能够直接导入的格式。
- **Excel 文件**：有一些文件是扩展名为 .xls 或 .xlsx 的 Excel 电子表格示例。这些文件需要使用 Microsoft Excel 运行。
- **图可视化文件**：一些示例还包含图可视化文件，例如 .gephi 或 .cys。这些文件与特定的图可视化软件关联，例如这两种文件分别与 Gephi 和 Cytoscape 关联。要查看这些文件，必须首先下载并安装免费的图可视化软件包。具体细节下一节将进行介绍。
- **Python 代码**：编程示例使用了 Python 语言。这些程序文件的扩展名为 .py。Python 示

①（在线资源请登录 <http://as.wiley.com/wileyCDA/wileyTitle/productCd-1118845846.html>，同时可以登录华章网站 www.hzbook.com 下载）

例中使用的是 Python 3.x 版本，要求下载并安装 Python。具体细节下一节将进行介绍。

- **HTML 和 JavaScript**：使用 JavaScript 的示例通常是包含 JavaScript 的网页文件，扩展名为 .html。这些文件在标准的现代 Web 浏览器中就可以运行，例如最新版本的 Chrome 或 Firefox。

示例中用到的工具

本书使用了众多工具来处理数据或可视化数据。为了使用前面列出的数据文件，需要有下列软件。

- **Gephi**：Gephi (<https://gephi.github.io/>) 是终端用户使用的一个免费的鼠标点击类软件，本书中的许多图可视化示例都用到了这个工具。许多数据文件都可以导入到 Gephi 中进行分析 and 可视化。第 7 章以第 3 章～第 6 章描述的基本图分析过程为基础，讨论了 Gephi 的一些功能。
- **Cytoscape**：Cytoscape (www.cytoscape.org/index.html) 是另外一个免费的、供终端用户使用的图分析软件工具，也用在了本书的许多示例中。许多数据文件也可以导入到 Cytoscape 中进行分析 and 可视化。第 7 章讨论了 Cytoscape 的一些功能，并说明了 Gephi 与 Cytoscape 之间的一些区别。
- **yEd**：yEd (www.yworks.com/en/products/yfiles/yed/) 也是一个免费的、供终端用户使用的鼠标点击类软件产品，由 yWorks 开发，用于图的分析与可视化。
- **Excel**：有几个示例中用到了 Microsoft Excel (<http://products.office.com/en-us/excel>) 电子表格。Excel 不是免费的，但是大部分读者应该已经安装了该软件，而 Microsoft 也允许下载该软件，并评估试用一段时间。有几个例子还使用了 Excel 的 NodeXL 插件。
- **NodeXL**：Excel 允许开发人员创建插件来访问并增强 Excel 的功能。NodeXL (<http://nodexl.codeplex.com/>) 为社交网络数据获取提供了图功能，还提供了图的分析与可视化功能。
- **Python**：为了通过编程操纵数据，一些示例中使用了 Python 3 (<https://www.python.org/>) 编程语言。Python 可以免费获取。
- **一个现代浏览器**：虽然任何现代的 Web 浏览器都应该能够查看 JavaScript/HTML 示例，不过作者们使用的浏览器是 Chrome (https://www.google.com/intl/en_us/chrome/browser/)。
- **D3.js**：D3 (<http://d3js.org/>) 是用于在浏览器中创建多种交互式数据可视化的一个 JavaScript 库，第 8 章等地方就使用了 D3。
- **Aperture JS**：Aperture JS (<http://aperturejs.com/>) 是本书后半部分（例如第 12 章）的

一些示例中使用的一个 JavaScript 框架库。

📦 **Titan**：第 14 章的几个大数据示例中使用了 Titan (<http://thinkaurelius.github.io/titan/>) 图数据库。

要使用这些软件库和工具，需要自己下载并安装它们，不过 JavaScript 库 (D3.js 和 Aperture JS) 是例外，它们已经与下载示例打包在一起，可从前面提到的本书配套网站上下载。

注意事项

本书的各个章节使用案例分析来演示图的各种应用与形式，以及如何使用图。在可能的地方，演示使用了真实的工具和真实的数据。对于这些情况，有几点需要牢记在心。

虽然作者使用的是开源工具，任何人都可以免费获得这些工具，但是其中的许多工具仍然处在开发当中，因而缺少最终成品的一些光彩与健壮性。需要知道，格外耐心有时候是早期采用一个产品所要付出的代价。将本书中与工具相关的步骤视为一个过程的一般指导原则。如果用户界面看上去与书中的描述不完全相同，则要在更新的界面中找到对应的选项。如果找不到，快速地在网上搜索通常足以帮助你找到你要寻找的东西。

另外要记住的一点与要分析的数据有关。像本书这类图书依赖于公共数据集。虽然近年来将公司数据集开放给公众，以发展分析与可视化的艺术与科学的行动有了巨大的进展，但是私有的数据集始终更加庞大、更加丰富。虽然本书中的分析对于使用的数据是正确的，但是很多时候这些数据只是公司网络内的数据的样本。将本书的分析当做一种模板方法，在你的全部数据处理中可以照用它们。

约定

为了帮助你最大程度地理解文字内容，并跟上内容进度，本书中采用了一些约定。

警告 警告框中包含重要的、不能忘记的信息，这些信息与警告框周围的内容有直接关系。

注意 注意框指出了一些注意事项、提示、暗示、技巧或者题外话。

提示 提示框提供了能够帮助掌握所讨论信息的提示或者技巧。

作者简介

Richard Brath 是数据可视化的积极实践者和先行者，其视觉分析的研究、设计与开发不仅涉及研究领域还用于商业领域。他创建的解决方案范围很广，从用于移动设备中丰富的交互式可视化，到用于商业应用的多点触控、多屏幕装置以及基于 Web 的可视化分析，涉及的应用领域也很广，如贸易、职业体育和广播电视等，每天都有成千上万的人使用。

David Jonker 是 Uncharted (原来的 Oculus Info Inc) 公司的联合创始人和高级合伙人。他是一名设计师和开发人员，为基于 Web 的、分布式的、移动的应用设计可视化分析工具和平台。他在过去 20 多年做了大量可视化工作，其中包括位于时代广场 NASDAQ MarketSite 实时广播中心的可视化系统。目前，他是 DARPA XDATA 项目的带头人。Jonker 和 Brath 是商业合作伙伴，两个人也经常领先的行业及研究论坛上发表演讲，进行展示。

技术编辑简介

Scott Langevin 是 Uncharted 的一位主管和研究人員，拥有超过 12 年的行业和学术界经验。他在南卡罗来纳大学获得了计算机科学的博士学位，方向是机器学习、面向服务计算和软件工程。Langevin 的研究兴趣包括概率图建模、大规模可视化分析和适应性用户界面。

Peter MacMurchy 是拥有超过 15 年经验的专业软件开发人员，他关注 UX、UI 和交互式数据可视化工具。在卡尔加里大学读计算机科学学位研究计算机图形学时，课程作业激发了他对信息可视化的强烈兴趣。自那之后，他就一直为金融、电影、能源等行业开发可视化和交互软件。

目 录

前言

作者简介

第I部分 概述

第1章 为什么使用图 2

- 1.1 商业中的可视化 3
- 1.2 商业中的图 4
 - 1.2.1 找出反常现象 5
 - 1.2.2 管理网络和供应链 7
 - 1.2.3 辨别风险模式 9
 - 1.2.4 优化资产组合 11
 - 1.2.5 绘制社会等级分层图 13
 - 1.2.6 发现社区 15
- 1.3 图的现状 16
- 1.4 小结 17

第2章 图的类型及其适用的问题 18

- 2.1 关系 18
- 2.2 分层 21
- 2.3 社区 23
- 2.4 流 27
- 2.5 空间网络 30
- 2.6 小结 32

第II部分 过程和工具

第3章 数据：收集、清洗和连接 35

- 3.1 了解目标 35
- 3.2 收集：识别数据 35
 - 3.2.1 潜在的图数据源 36
 - 3.2.2 潜在的分层数据源 41
 - 3.2.3 获取数据 43
- 3.3 清洗：准备数据 44
- 3.4 连接：组织图数据 45
 - 3.4.1 计算图 46
 - 3.4.2 图数据的文件格式 48
- 3.5 集中回顾 54
- 3.6 小结 54

第4章 统计数据和布局 55

- 4.1 基本的图统计数据 55
 - 4.1.1 大小（节点数和边数） 55
 - 4.1.2 密度 56
 - 4.1.3 成分数 56
 - 4.1.4 度和路径 56
 - 4.1.5 中心度 58
 - 4.1.6 病毒式营销示例 59
- 4.2 布局 60
 - 4.2.1 节点-连接布局 60

4.2.2 其他布局	61	5.7 小结	101
4.2.3 力导向布局	62	第 6 章 探索 and 解释	102
4.2.4 仅节点布局	66	6.1 探索、解释和导出	102
4.2.5 时间布局	67	6.2 必要的探索性交互	104
4.2.6 自顶向下和其他正交分层	68	6.2.1 缩放和摇动（以及比例缩放和旋转）	105
4.2.7 辐射状分层	71	6.2.2 识别	106
4.2.8 地理布局和地图	72	6.2.3 过滤器	107
4.2.9 弦图	74	6.2.4 隔离和重做布局	109
4.2.10 邻接矩阵	74	6.3 更多交互式探索	110
4.2.11 树图	76	6.3.1 识别邻近节点	111
4.2.12 分层饼图	76	6.3.2 路径	111
4.2.13 平行坐标	77	6.3.3 删除	112
4.3 集中回顾	79	6.3.4 分组	112
4.4 小结	79	6.3.5 迭代分析	114
第 5 章 视觉特性	80	6.4 解释	114
5.1 基本视觉特性	81	6.4.1 数据故事的顺序	115
5.2 关键的节点特性	82	6.4.2 图例	116
5.2.1 节点大小	82	6.4.3 注释	116
5.2.2 节点颜色	84	6.4.4 导出数据子集、图和图片	118
5.2.3 标签	87	6.5 集中回顾	119
5.3 关键的边特性	91	6.6 小结	120
5.3.1 边的权重	91	第 7 章 鼠标点击类图工具	121
5.3.2 边的颜色	91	7.1 Excel	121
5.3.3 边的类型	92	7.1.1 汇总连接	122
5.4 组合基本特性	93	7.1.2 提取节点	122
5.5 捆绑、形状、图片及更多	94	7.1.3 Excel 中的邻接矩阵可视化	123
5.5.1 捆绑边	94	7.2 NodeXL	125
5.5.2 形状	95	7.2.1 NodeXL 基础	125
5.5.3 节点图片	95	7.2.2 社交网络功能	127
5.5.4 节点边框	96	7.3 Gephi	129
5.5.5 更多特性	97	7.3.1 Gephi 基础	129
5.5.6 干扰与分隔	97		
5.6 集中回顾	101		

7.3.2 注意事项	131	第 10 章 分层	189
7.4 Cytoscape	133	10.1 组织结构图	189
7.4.1 Cytoscape 基础	133	10.2 树与图	191
7.4.2 将数据导入 Cytoscape	134	10.3 绘制分层	193
7.4.3 视觉特性	135	10.4 决策树	198
7.4.4 Apps 菜单	139	10.5 网站树及有效性	200
7.5 yEd	139	10.6 小结	203
7.6 小结	141	第 11 章 社区	204
第 8 章 轻量级编程	143	11.1 社区的定义特征	205
8.1 Python	143	11.2 图聚类	205
8.1.1 上手	143	11.2.1 社交网络案例分析	206
8.1.2 清洗数据	144	11.2.2 使用 NodeXL 和 Gephi 分析 社交媒体	206
8.1.3 从连接数据集中提取节点 集合	145	11.2.3 可聚类的布局	208
8.1.4 将电子邮件数据转换为图	149	11.2.4 使用颜色描述簇的特征	210
8.1.5 图数据库	154	11.2.5 社区发现	211
8.2 JavaScript 与图的可视化	155	11.2.6 使用颜色来区分簇	212
8.2.1 D3 基础	155	11.2.7 社区话题分析	214
8.2.2 D3 和图	160	11.2.8 社区情感	216
8.2.3 D3 弹簧图	169	11.3 团伙和其他组	219
8.3 小结	174	11.3.1 社交媒体中的团伙	220
第Ⅲ部分 图的可视化分析		11.3.2 使用凸包的社区组	220
第 9 章 关系	176	11.4 小结	223
9.1 连接和关系	176	第 12 章 流	224
9.1.1 诈骗索赔中的相似性	177	12.1 桑基图	225
9.1.2 网络安全	179	12.2 构造一个桑基图	229
9.2 电子邮件关系	181	12.2.1 创建页面结构	229
空间分隔	181	12.2.2 处理和建模数据	230
9.3 演员与电影	184	12.2.3 可视化数据	231
9.4 将连接转换为节点	186	12.2.4 高亮显示通过节点的流	233
9.5 小结	188	12.3 使用流的社区布局	235
		12.4 弦图	237

12.5	构造一个弦图	238	14.3	分析邻域	281
12.5.1	准备数据	238	14.4	绘制网络活动	287
12.5.2	创建页面结构	239	14.5	社区可视化	289
12.5.3	处理和建模数据	240	14.6	小结	290
12.5.4	可视化数据	243			
12.5.5	根据需要显示交互细节	247	第 15 章 动态图	291	
12.6	行为因子树	248	15.1	图的变化	291
12.7	小结	249	15.1.1	有机动画	292
			15.1.2	完整时间跨度布局	293
第 13 章 空间网络	250		15.1.3	重影	295
13.1	示意图布局	250	15.1.4	淡出	296
13.2	小世界分组	255	15.1.5	社区演化	297
13.3	连接玫瑰汇总	255	15.2	交易图	298
13.4	路线模式	263	15.2.1	聚类交易分析	299
13.4.1	可视化路线段	264	15.2.2	空间交易分析	304
13.4.2	轨迹聚合	267	15.3	小结	305
13.5	小结	268			
			第 16 章 设计	307	
			16.1	节点	307
第 IV 部分 高级技术			16.1.1	节点的形状	308
第 14 章 大数据	270		16.1.2	节点大小	313
14.1	图数据库	271	16.1.3	节点标签	314
14.1.1	产品营销示例	271	16.2	连接	314
14.1.2	创建和填充一个图数据库	273	16.3	颜色	318
14.2	图查询语言	275	16.4	小结	320
14.2.1	使用 Gremlin 进行图查询	276			
14.2.2	使用图查询来提取邻域	278	图论术语表	322	

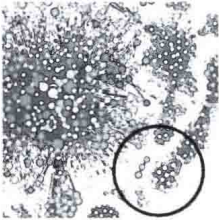
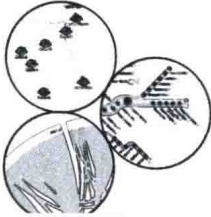
概 述

本书第 I 部分介绍了图这个主题，并回答了两个重要的问题：为什么对于商业分析来说图很有价值？可以用来发现什么样的机会？本部分借鉴历史发展和现实经验，讨论了一系列技术和应用。还给出了一些案例，以说明图的价值。

在本书第 II 部分讨论图分析的过程之前，本部分的概述可帮助读者感受图的类型有多少种，以及图可以在多少个领域提供价值（即使在一个企业中也可能存在多个这样的潜在价值领域）。本部分的引用内容作为学习第 III 部分的指导，该部分的各个章节将详细讨论图的各种类型，并采用教程风格讲解了图分析的应用。

表介绍了第 1 章和第 2 章的主题。

概述

主题		描述
为什么使用图？（第 1 章）		什么是图？图为什么对于商业分析师很有用？第 1 章介绍了图的概念，并定义了本书中用到的几个关键术语。本章遴选的历史和现代轶事叙述了图的分析 and 可视化在商业中的应用，并记录了当今庞大而复杂的数据带来的挑战，以及在这种挑战的推动下，图的重要性如何稳定提升。
每个问题一个图（第 2 章）		第 2 章系统地概述了图的广泛类型，以及它们用来解决的问题的类型。本章的讨论以一个例子开始，说明以其他方式揭示的关系也可以用节点和连接来表达。后续的主题描述了用于获得商业见解的图技术，涉及分层、社区、流和空间网络。本章还包含一些引用内容，后面章节中将详细介绍这些内容。

第1章

为什么使用图

本书介绍图以及如何使用图来帮助解决商业问题。当听到“图”这个词时，许多人想到的是条形图表或折线图表，当然这是没有问题的，因为这些图表有时候也叫做条形图或折线图。本书介绍的不是图表，而是节点 - 连接图这类图。

本质上，“图”是相互连接的事物及其关系的一种结构化的表示。在后面的章节中将会看到，图能够表示复杂的数据，并帮助分析人员理解这些数据。

图在数学中由来已久，所以讨论图的分析 and 可视化常常会包含许多深奥的、令人不知所云的术语，例如“边”(edge)和“度”(degree)。相关的研究领域一般叫做“图论”。

在本书的讨论中，我们尽可能地使用更加容易被大众理解、更加清晰的术语。例如，“连接”(link)是“节点”之间的关系，通常绘制为线条。节点是实体（本质上就是“事物”），由连接线连在一起。节点在可视化表示中通常绘制为圆圈。

在图论中，边是另外一个表示连接的词。如果熟悉六度空间理论(six degree of separation)（同名的歌曲和电影使之流行起来），那么术语“度”的含义会稍微清晰一点。但只是清晰一点而已，因为“度”不只可以表示连接实体之间分隔的最少步骤数，还可以表示节点拥有的连接数。

注意 本书末尾的“图论术语表”通过一系列术语解释了图论，所以如果不熟悉图论，可以参考这一部分。

一些圈子里仍然把图视为抽象而难以理解的结构，主要由头发乱糟糟的科学家使用。虽然图在科学界确实由来已久，但是实际上如果能够恰当地设计和创建图，那么图是分析信息最直观的方式之一。如果曾经在笔记本或者白板上通过绘图的方式来思索或解释概念（这其实是可视化的一种形式），那么其实就使用过图表示了。

更重要的是，图能够用来从数据中获得高度独特而有价值的见解。图分析能够揭示复杂的关系，从而帮助有效地制定决策。可视化是此过程的核心。能够以可视化的方式看到关系对于理解关系十分关键，不管这些关系是原始数据的特征，还是图分析揭示出来的具体特性。

信息可视化存在的唯一目的就是在更少的时间里理解更多信息。大脑的工作方式决定了人们以形象化的方式察觉和理解失误。“读”是一个耗时的、顺序的过程，要求阅读者在脑中将信息串联起来形成理解。图片能够即时传达信息，以易于消化的方式揭示复杂的模式和离群值。

有一段时间，可视化是通过手绘完成的，而在绘制之前，还要经历辛苦的数据收集过

程。如今，计算机系统能够在几毫秒的时间内收集大量数据并将其转换为图形，使分析人员能够立即理解并处理信息。几乎所有企业都可以从可视化中受益，因而，可视化已经成为全世界各个行业的系统的核心。但是，图是仍然未被充分利用的可视化形式之一。就是说，有一段时间，所有的信息可视化形式在企业中都没有得到充分的利用。

1.1 商业中的可视化

在企业的决策制定中使用计算机呈现的可视化相对而言是近期出现的一种现象。20年前，我们刚刚从滑铁卢大学建筑学院毕业，受到当时新兴的广阔虚拟景观新世界的吸引，我们决定放弃设计物理景观。我们中的一位作者花了几年时间研究 3D 建模软件，后来我们与其他同事合作，看看是否能用类似的技术，为金融和其他行业的成功的决策制定者解决显示大量抽象信息的问题。这次合作最终催生了长期的合作伙伴关系，这些合作伙伴中包括 William Wright 和另外一个年轻的建筑师 Thomas Kapler。

在那次合作探索商业可视化的早期，图表（即使是最简单的图表）的价值也还没有被财富 500 强公司广泛理解或接受。我们一开始用最基本的价值主张——可视化自身价值——向公司的决策制定者推销我们的观念。我们首先用一张幻灯片显示一个包含数字的小表格，然后让房间中的决策者们描述模式。下一张幻灯片用折线图的形式显示了相同的数字。将数字可视化后，模式显而易见。在表格中，模式则很难看出。以这个基本原则作为基础，可以推断出要从庞大许多、也复杂许多的数据中获取有用的见解，可视化会更加关键。

当时，使用计算机来进行基本绘图的方法还只是处于新生阶段，而以可视化方式分析商业数据的行业总体来看尚未出现。当时进行的一些先进的研究工作局限在少数几个公司的研究实验室和新兴公司中。商业世界是一个还不大运用图表的世界。

在早期商业世界在采用可视化时面临着一些阻碍，其中之一就是当时的计算机系统的图形处理能力有限。当 Edward Tufte 的著作 *Envisioning Information* (Cheshire, CT : Graphics Press, 1990) 出版时，业界的最佳实践仍然基于打印，他的这本具有开创意义的设计图书中选用的案例分析也不例外。普通计算机的显示质量仍然十分落后。

20 世纪 90 年代初，我们带着新颖的交互式 3D 演示来到纽约，把它们展示给金融分析师和交易员。当时，他们使用着专用的几百磅重的硬件。支持一个系统需要笨重的 Silicon Graphics Inc. (SGI) 计算机和监视器。不断地把设备搬入搬出出租车的后备箱，并把设备放在快散架的折叠式手推车上沿着人行道运送，导致不久以后新机器就用上了强力胶布。

更大的一个问题是，当时华尔街（甚至整个商业界）还没有谁有一台 SGI 机器。每个用户的新机器和操作系统的价位达到五位数，但是不能运行他们其他的应用程序，这使得交互式可视化软件系统很难让人接受。我们把许多高调的公司列到一个单子里，为他们制作了有针对性的原型，但是广泛的采用仍然很难实现。

后来 Microsoft Windows 计算机问世，提供了更好的图形 API 和显卡，也改变了整个局势。由于能够在大部分台式机上使用更高质量的图形能力，所以就不再需要昂贵的专用机器，这代表了企业在广泛采用先进的可视化方面的一大进步。到 20 世纪 90 年代中后期，被广泛部署的高性能的分析客户端平台（如 Bloomberg Terminal）就运行在 PC 上。甚至高度专业化、要求极高的系统（如 NASDAQ MarketSite 广播墙）也运行在商用 Windows 计算机上。

随着硬件的图形处理能力开始成熟，人们也越来越意识到可视化的价值。及时、准确而

迅速地察觉事件和趋势，对于在交易大厅或者其他需要不断监控系统和事件的地方快速做出决策至关重要。在业务分析中，通过以图形化方式呈现信息来辅助获取深入见解和支持战略级决策制定，这种方法的价值在各行各业快速得到了认可和接受。

面对一个快速增长的市场和一个尚未被图形占领的世界，我们在整个世界的新鲜而又令人兴奋的领域找到了我们一展拳脚的机会。例如，当 NASDAQ MarketSite 从市中心的一间办公室搬到时代广场的一个公共的工作室时，需要重新构建其软件基础设施，于是委托我们设计和构建可视化系统和内容。新工作室打算在千禧年前夕启用，它包含一个 40 英尺[⊖]长的广播墙，上面有大约 100 个显示器，并且在其七层楼的外墙嵌上了一个电子显示屏。按照记者和公众的需求，广播墙能够以可视化方式实时显示超过 6000 个股票和指数。

从那以后（其实之前就已经开始），我们有幸通过设计和技术开发，在幕后帮助世界上最具创新性的公司和组织以可视化的方式解决棘手的信息问题。在这个过程中，我们有机会在几近 100 家企业内部见证了行业的变迁，这些企业分属于各种数据密集的行业。随着时间不断前进，可用数据的数量越来越多，从数据中能够获取的信息的潜力也越来越大。数据现在是在无处不在，等待人们加以利用以获取可以指导行动的见解。

随着人们逐渐认识到需要利用可视化来深入理解数据，人们也开始意识到可视化系统必须具备高度的可交互性。简单地绘制并查看数据是不够的，正如简单地计算并显示结果不能满足要求一样。“分析”是进行快速查询、解答和探索的一个交互式过程，涉及计算过程、视觉显示和视觉上的操纵。在 21 世纪早期，很多人认为可视化就是一种输出方法，由于对这种观念感到不满，研究社区创造了“可视化分析”（visual analytics）这个术语，以更好地表达和宣传这种以交互方式分析的形式。

随着企业中信息问题的大小和复杂度逐渐增加，人们也开始认识到，基本的折线图、条形图和饼图常常无法表达全部可用的有价值信息，将这些信息运用在决策制定中。还需要有更丰富的形式和形式组合。事实证明，图是最有价值的形式之一。

1.2 商业中的图

在大约 25 年的时间里，我们一直在帮助各种组织进行可视化和对图进行分析。图出现的时间要早得多。最早的与图有关的问题之一由莱昂哈德·欧拉提出，这个问题乍看上去很简单：有没有一条路线，使得普鲁士哥尼斯堡（如今俄罗斯的加里宁格勒）的七座桥中的每一座只被通过一次（图 1-1 的左图）？欧拉将这个问题简化为一个图，如图 1-1 的右图所示。

自那之后，无论是在商业界还是科学界，更多的问题被明显地作为图来分析。其中许多问题都与地理位置有关，就像欧拉的问题一样。

我们最早创建的图可视化方案中，有一个也是地理图问题。在供应链优化中，要完成的任务是优化工厂和仓库之间的产品运送，以降低成本。如图 1-2 所示，我们的可视化方案描绘了各个场所的位置，并用图标指示一些特性，如类型、库存、容量和使用情况，另外还用一些较大的连接指示平均成本。

使用这种供应链可视化方案可以完成多种类型的分析，例如从检查单独的路线到精简工厂和仓库的总体数量。一个值得注意的观察结论是，特定两个工厂之间的成本在三月、六

⊖ 1 英尺 = 0.3048 米。——编辑注

月、九月和十二月会翻倍。调查后发现，在每个季度末一条路线的运送成本会明显增长。进一步调查显示这条路线是从陆运改为更快（但更贵）的空运。一些问询揭示出这种变化是由高层目标推动的，以便满足每季度的目标。因为这种模式在每个季度重复出现，分析人员意识到，整个季度中在两个工厂之间进行更好的规划和协调能够实现更好的运送安排，从而在季度的最后一个月份中降低运送成本。类似的，在其他供应链网络的分析和优化中也可以使用图的分析 and 可视化。

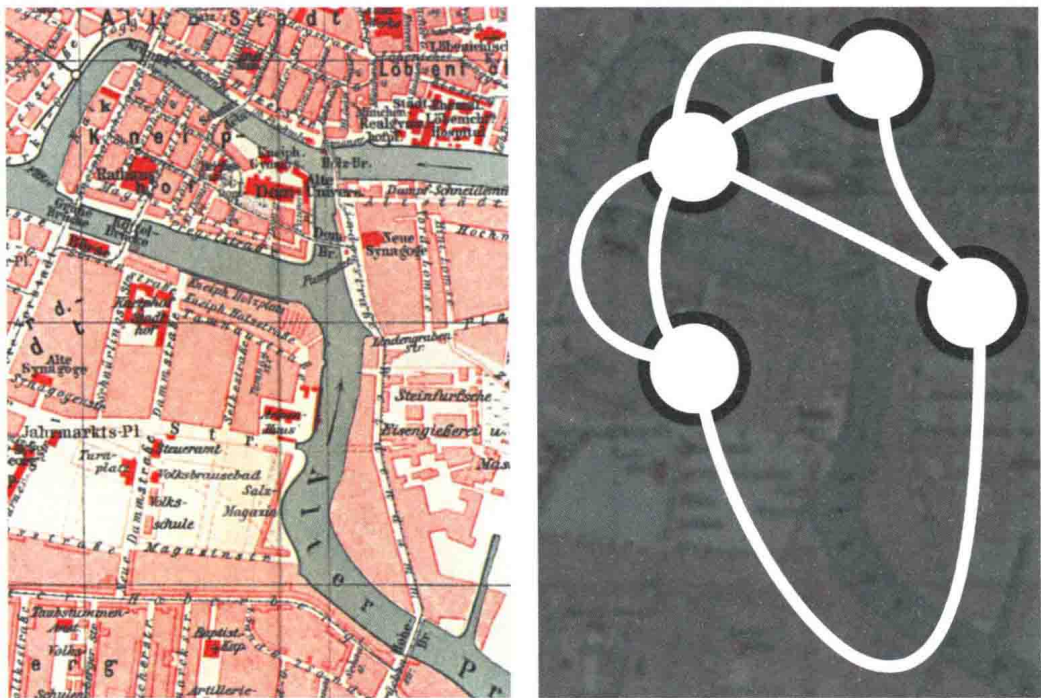


图 1-1 在哥尼斯堡七桥问题中，莱昂哈德·欧拉研究每座桥是否能够只被通过一次。左侧是显示了七座桥位置的地图，右侧是欧拉简化后的图

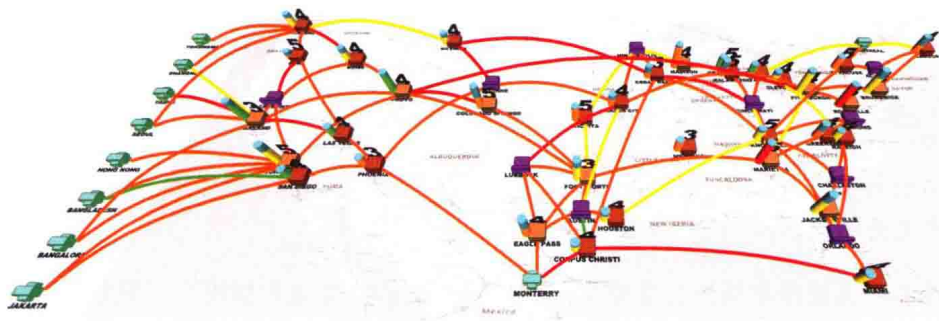


图 1-2 作者最早创建的可视化方案之一描述了一个制造和配送供应链网络

注意 第 9 章将更详细地讨论基本的图 and 关系。

1.2.1 找出反常现象

空间图（Spatial graph）常用于分析商品在公司中或者全世界的流动情况。这种图的一