第3章 网络模型分析方法

移动数据蕴含着移动个体与个体以及个体与时空的复杂交互和关联信息，因此可以建模为网络结构。我们可以从网络模型的视角，采用相应的分析方法对其进行挖掘。网络模型分析方法提供了一系列表示、统计、聚类、结构检测等方法。3.1 节介绍网络模型中基本定义及统计量等概念及其在真实网络中常见的性质，3.2 节和 3.3 节介绍用于网络社群结构分析的三角形查找算法和社群发现算法，3.4 节和 3.5 节以两个典型应用案例展示网络模型在移动数据挖掘中的具体应用。

3.1 网络模型概述

移动数据中包含的人到访空间地点、人与人之间的通话、人对地点推荐的关注等关系，都可以被抽象为节点之间的链接。自然地，这种关系可以用一种通用的网络模型来表示，在此基础上可基于网络模型的分析方法对其进行分析挖掘。本节提出网络模型中的基本定义，列举真实移动数据体现的网络统计特性并介绍分析这些网络特性的基本方法。

3.1.1 网络模型的定义

网络 $G=(V,E)$ 是一个有序二元组，其中 V 是**节点**的集合，$E\subseteq\{\{x,y\}\mid x,y\in V$ 且 $x\neq y\}$ 是**边**的集合，即一条边是两个不同节点间的链接。如图 3.1 所示，通常用一个小圆圈表示网络中的一个节点，若两个节点之间存在一条边，则使用一条线连接两个小圆圈。通常用 N 表示网络中节点的数量，$N=|V|$，用 L 表示边的数量，$L=|E|$。

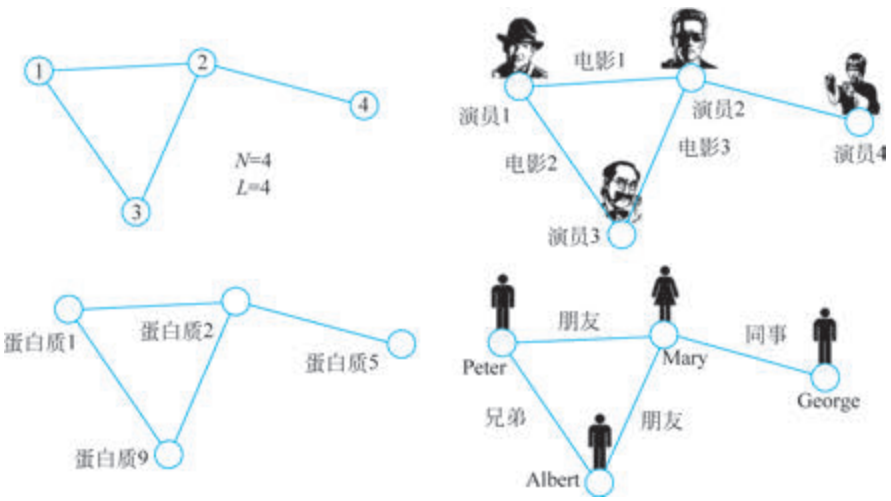


图 3.1 网络模型的抽象示意图及对应的三个实例：演员网络、蛋白质网络和社交网络

一个在数学意义上抽象的节点可以表示真实世界中的一个实体，边可以用来表示实体间的关系。如图 3.1 所示，一个有 4 个节点、4 条边的抽象网络模型可以表示一个演员网络，其中每个节点为一个演员，而在同一个电影中共同出演的演员间用一条边相连，用边来表示“共演”关系。这个网络模型还可以表示一个蛋白质网络，每个节点代表一种蛋白质，能够组装成复合物的蛋白质间通过边相连，用边表示“复合”关系。同理，该网络模

型还可以用来建模一个社交网络，其中每个节点表示一个人（个体），而有社交关系的两人间存在边相连，如兄弟、朋友、同事等。从此图中也可以看出，网络中的边根据实际建模对象可以具有多种含义。

根据网络中的边是否是节点的有序对，可以进一步将网络区分为**无向网络**和**有向网络**。在无向网络中，每条边是节点的无序对，即一条边表示的两个节点间的关系是相互的、对称的，如科学界中的共同作者关系、演员间的共同出演关系以及蛋白质间能够互相发生反应的关系。在有向网络中，边具有方向，每条边是一组节点间的有序对。有向网络中的边通常用带有方向的箭头表示，箭头的方向指示了关系的起点和终点。如图 3.2 所示，存在由 B 到 C 的边，但不存在由 C 到 B 的边，这表明有向边是不对称的。有向网络常用于表示有方向性的关系，如社交网络中的关注关系（用户 A 关注用户 B，但用户 B 未必关注用户 A）、信息流的传播方向、交通网络中的交通流向等。

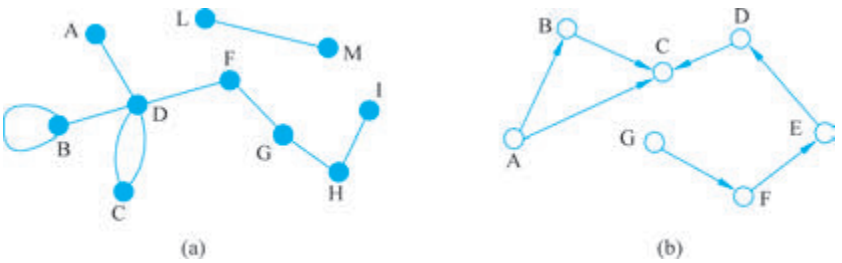


图 3.2　无向网络和有向网络

在区分边的方向之外，还可以进一步根据边是否具有权重将网络分为**无权网络**和**加权网络**。加权网络是一种用于表示具有不同强度或权重的关系的网络。如图 3.3 所示，在加权网络中与每条边（不管是否有方向）关联一个权重值，该值表示了相应边的强度或关联程度。这些权重可以代表不同的度量，如距离、相似性、关联概率等，具体取决于网络的应用和领域，如道路网络中两路口之间的车流量、通话网络中两个用户间的通话时长等。

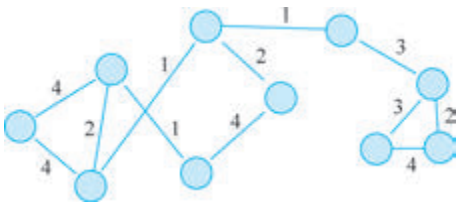


图 3.3　加权网络示意图

3.1.2　网络中的度量

在对网络模型中节点、边及基本类型进行定义后，为了刻画节点和边的一些有趣的性质，研究者提出了系列度量指标来刻画这些性质，并在此基础上发展了网络科学。从节点具有的连接性质出发，网络科学定义了节点的度及相应的度分布与节点集聚的程度。从网络中节点对之间连接能力的视角还定义了网络中的路径及距离等度量。

网络节点的**度**是指与该节点直接连接的边的数量，也称作该节点的邻居数量。节点 v 的度通常用 k_v 来表示。对于有向网络，可以进一步定义**入度**和**出度**。对于一个节点，其入度为网络中指向该节点的边的数量，其出度为从该节点出发的边的数量。一个节点 v 的入度和出度通常分别用 k_v^{in} 和 k_v^{out} 来表示。有向网络中节点的度为入度和出度之

和。如图 3.4(a)所示的无向网络中，节点 A 只与一个节点相连，其度为 1，节点 B 与 4 个节点相连，其度为 4。如图 3.4(b)所示的有向网络中，由节点 C 出发的有 1 条指向节点 B 的边，而指向节点 C 的有 2 条分别来自 A 和 D 的边。因此节点 C 的入度为 1、出度为 2、度为 3。节点的度、入度和出度是网络分析中常用的重要度量，它们可以帮助人们理解节点在网络中的作用和影响力。例如：在社交网络中，节点的度可以表示一个用户的朋友数量；入度和出度可以表示用户的粉丝和关注的人的数量；在互联网中，节点的度可以表示一个网页的入站链接数量，入度和出度可以表示网页的被引用和引用其他网页的情况。这些度量可以用于识别网络中的关键节点、分析信息传播路径等。

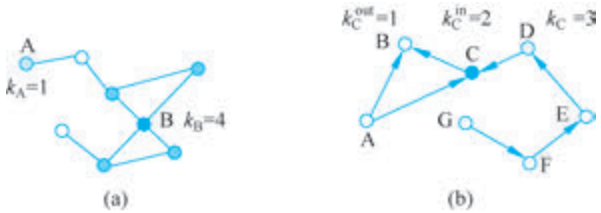


图 3.4　无向网络和有向网络中节点度的定义

节点度的一个重要统计量为网络的**平均度**，即网络中所有节点的度的平均值，记作 $\langle k \rangle$。对于无向网络，其平均度定义为

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

式中：N、L 分别为网络的节点数和边数。

对于有向网络，其平均入度和平均出度分别定义为

$$\langle k^{\text{in}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{in}}$$

$$\langle k^{\text{out}} \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^{\text{out}}$$

且有 $\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = \frac{2L}{N}$，即网络的平均入度和平均出度相同。

网络的平均度由总边数和总节点数直接决定，体现了整个网络的连接密度。平均度越高，表示网络中的节点之间连接越密集，平均度越低，表示节点之间的连接越稀疏。这有助于判断网络的稠密程度，从而影响信息传播、网络效率等方面的分析和决策。

仅用平均度这一单一度量无法体现网络中各节点间的差异，更无法精细地体现网络的拓扑结构。因此，引入网络的**度分布**，即不同度的节点在网络中的分布情况，记作 $P(k)$ 或 P_k。度分布描述了网络中节点的度数分布模式，通常以概率分布或直方图的形式表示。如图 3.5 所示，一个有 $N=10$ 个节点的无向网络中，共有 6 个度为 1 的节点、1 个度为 2 的节点、2 个度为 3 的节点和 1 个度为 4 的节点。因此其度分布如直方图所示，$P(k=1)=0.6$，$P(k=2)=0.1$，$P(k=3)=0.2$，$P(k=4)=0.1$。度分布可以帮助人们了解网络的拓扑结构和复杂性，在更复杂的网络拓扑下，度分布会呈现出更复杂的情况。常见的度分布包括幂律分布、指数分布等，不同的度分布对网络的性质和行为产生重要影响。

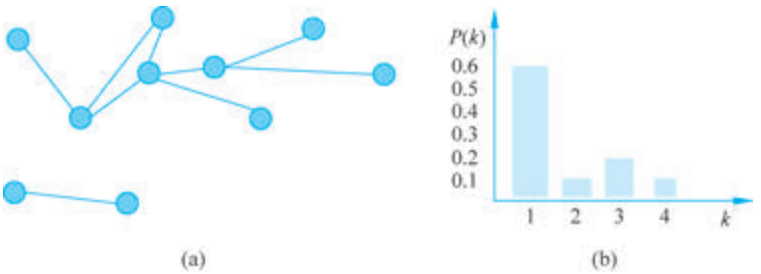


图 3.5　网络度分布 $P(k)$示意图

除从节点的角度关注度的均值和分布，还需要关注网络中边的连接情况。在大多数真实世界的网络中，特别是在社交网络中，节点往往以相对高密度的联系为特征，形成紧密的群体，且这种可能性往往大于两个节点之间随机建立平局的平均概率。因此，引入网络中的**集聚系数**来刻画网络中节点之间的连接紧密程度。无向网络中节点 i 的局部集聚系数定义为

$$C_i = \frac{2e_i}{k_i(k_i - 1)}$$

式中：k_i 为节点 i 的度；e_i 为节点 i 的 k_i 个邻居节点间连接的边的个数。

集聚系数通常取为 0～1。如图 3.6 所示，若深色节点的所有 3 个邻居节点都两两相连，即 $e_i=3$，集聚系数 $C_i=1$。若 3 个邻居节点间仅存在一条边，即 $e_i=1$，集聚系数 $C_i=1/3$。若邻居节点间均不相连，则集聚系数为 0。可以看出，局部集聚系数可以用于衡量节点周围的小社群或子图的聚类程度。一个节点的局部集聚系数越高，表示该节点的邻居之间连接越紧密，形成了一个聚类。基于集聚系数，可以揭示网络的局部结构、评估网络的聚类性质、检测网络中的三角形结构（将于 3.2 节中讨论）以及帮助识别网络中重要的枢纽节点。

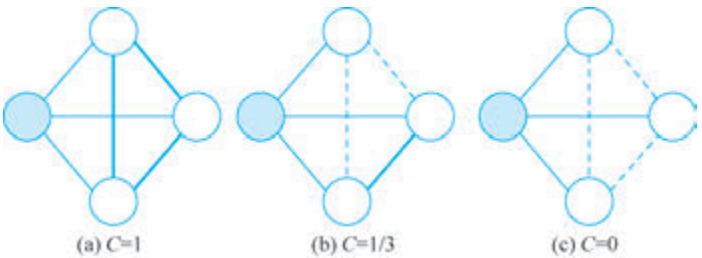


图 3.6　网络中节点的集聚系数示意图

注：实线表示节点间有边相连，虚线表示没有边相连。

集聚系数刻画了网络局部的聚类程度，网络的另一个特征是连接可能在空间中距离较远的节点。在网络模型中，**路径**是网络中连接两个节点的一系列边的序列，描述了从一个节点到另一个节点的路线或路由。路径可以是有向的（考虑边的方向）或无向的（不考虑边的方向）。如图 3.7 所示，节点 I 到 K 间存在 I—G—K 的路径。一个网络中可以有多个路径连接两个节点，如 I 到 K 还可以由 I—G—H—K 连接。若网络中的任意两个节点之间都存在至少一条路径，即可以通过网络中的边从一个节点到达另一个节点，那

么称这一网络为**连通网络**。连通网络中的节点之间存在通信或传输信息的能力，而在非连通网络中存在孤立的节点或多个不相互连接的子网络。图 3.7(a)展示了一个非连通的网络，D、E 两个节点和其他节点间无法经过边到达。

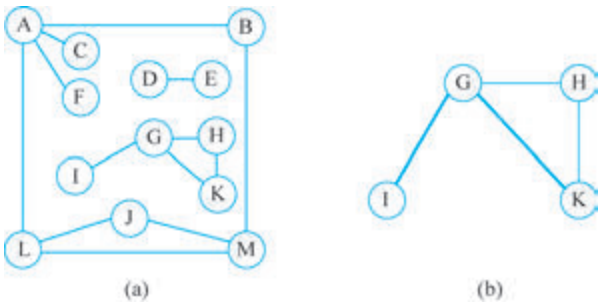


图 3.7　非连通网络和连通网络示意图

注：深色边表示由节点 I 到节点 K 的一条路径。

在一个连通网络上可以定义网络的**直径**，即网络中最长路径的长度，或者说连接两个节点的最短路径的最大长度，记作 $L_{\max}$。直径反映了网络中最远的节点之间的距离，通常用于衡量网络的全局规模或距离。

还可以定义网络的**平均路径长度**，计算如下：

$$\langle d \rangle = \frac{1}{2L_{\max}} \sum_{i \neq j} d_{ij}$$

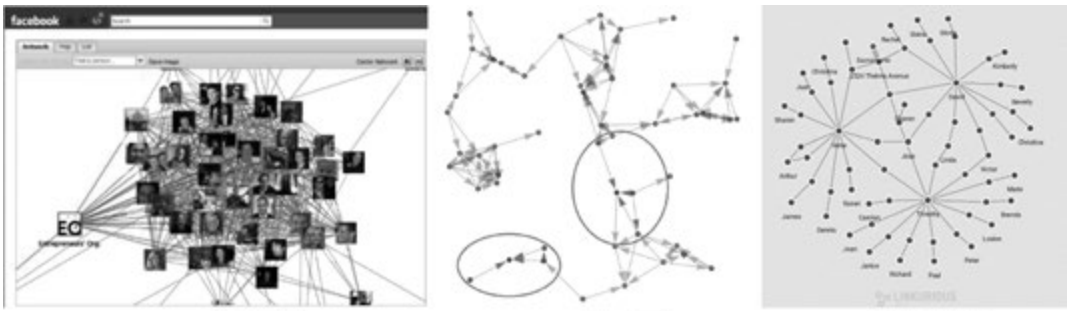
式中：d_{ij} 为从节点 i 到节点 j 的路径的长度(经过的边的数量)。

平均路径长度是网络中所有节点对之间的平均最短路径长度。它表示了网络中节点之间的平均距离，通常用于衡量网络的整体连接性。较短的平均路径长度通常意味着网络中的信息传输更加高效。这些概念对于分析网络的结构、传播过程、信息流动以及网络的整体性能具有重要意义。

3.1.3　真实网络的性质

在对网络模型的抽象数学定义及度量基础上，可以衡量真实世界复杂关系所呈现出的网络多样化性质。常被研究的一类网络是由社交关系组成的社交网络，它可以是有向的，也可以是无向的。如图 3.8 所示，一种社交网络为微信、Meta 等平台的“好友”网络，在此网络中，节点表示单个用户，边连接两个互为好友的个体。由于关系是相互的，因此好友网络是一个无向网络。而微博、Twitter 等社交媒体的关注者网络是一种有向的社交网络，节点表示单个用户，边从一个用户指向其关注的所有其他用户。而由电话通话构建的社交网络，既可以视为从拨打者到接收者存在有向边的有向网络，也可以视为由无向边连接拨打者和接收者的无向网络。

在历来对社交网络的研究中发现了很多有趣的属性，其中著名的属性是小世界属性，即在非常大的网络中任意两个节点之间的最大距离(网络的直径)相对较小的现象。或者说，大多数节点并不直接连接在一起，但给定节点的邻居很可能彼此相连。因此，大多数相邻节点之间可以在相对少的步骤或跳数内相互到达。与小世界属性相关的一个



(a) Meta好友网络　(b) Twitter关注者网络　(c) 电话通话网络

图 3.8　一些真实社交网络示意图[1]

很有名的理论是“六度分离理论”，该理论认为地球上的任何两个人之间平均只隔着六个中间人，即通过六个或更少的人都可以与世界上的任何其他人建立联系。而在万维网中，大多数网页之间都可以通过 12 个或更少的链接到达，这体现了互联网的小世界属性。如图 3.9 所示，在 7 个真实网络中，节点间路径的典型距离 l 与节点个数 N 的关系呈现出对数关系，即 $l \approx \dfrac{\log N}{\log\langle k\rangle}$。

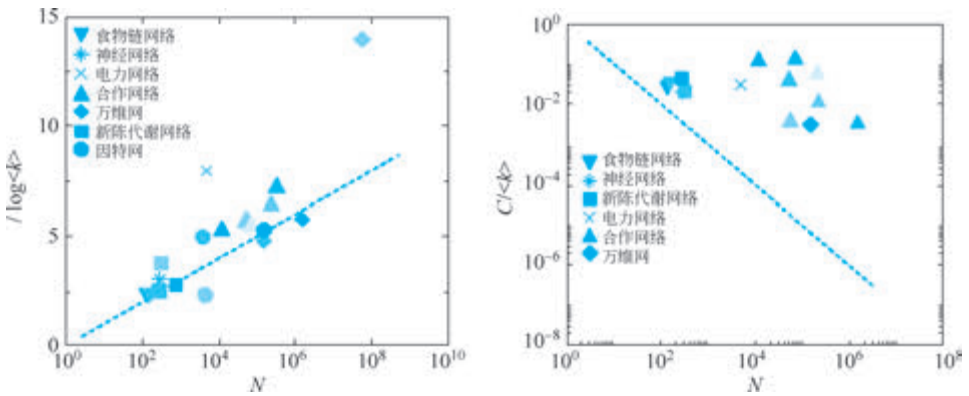


图 3.9　7 个真实网络的中节点典型距离及平均集聚系数和网络大小的关系[2]

真实网络的另一个特性是局部性，指的是网络中的节点或连接往往倾向于聚集在一起形成局部集群或社区。这意味着，如果一个节点与另外两个节点相连，那么这两个相连的节点之间也有更高的概率相互连接，而不仅仅是与起始节点相连。局部性在网络中是一种常见的现象，它表现为网络的一部分节点在空间或功能上聚集在一起，形成了紧密相连的小组。例如，在微信等社交平台，如果用户 y 与用户 x 和 z 都是朋友，那么用户 x 和 z 也是朋友的可能性很高，表现出局部性。在一个具有局部性的网络中，一些节点会形成高度连接的集群，而另一些节点可能与集群外的节点连接较少。这种不均匀的连接模式会导致集聚系数的分布存在一定的差异。但一个网络节点的平均集聚系数和节点平均度的比例往往是固定值，如图 3.9 所示，在 7 个具有不同节点数的真实网络中，集聚系数与平均度的比例均维持在 0.1 附近。

局部性在网络中有重要意义，因为它影响了信息传播、病毒传播、社交网络中的互动等各种网络过程。网络中的局部性可以导致社交网络中的朋友关系、互联网中的连接模式以及其他网络中的特定模式，特别是网络中的社群，即一组紧密相互连接的节点，它们之间的边的密度比网络的其余部分要高得多。在 3.3 节中将进一步探讨网络中社群结构的发现算法。

在网络科学的发展历史上，有很多科学家曾试图提出简单的模型来生成具有特定性质的真实网络结构。其中一种重要的拓扑结构特性就是网络的无标度性。无标度性指的是在网络中一些节点具有非常高的度，而大多数节点具有相对较低的度。换句话说，网络中存在一些“枢纽节点”或“超级节点”，它们连接了大量的其他节点，而大多数节点只连接了少数节点。这将导致网络的度呈现幂律分布，呈现出长尾的形状。在对万维网的研究中，如果将存在超链接的两个网页间连上相应的边构成一个网络（图 3.10(a)），那么此网络的节点度就呈现图 3.10(b)所示的幂律分布，在双对数坐标下为一条直线。而如果网络间任意两个点随机连接，网络的度将呈现图 3.10(c)所示的二项分布形式。显然，真实网络并非随机连接，而是服从特定的规则，因此有一些理论模型被提出以复现万维网的各种性质。

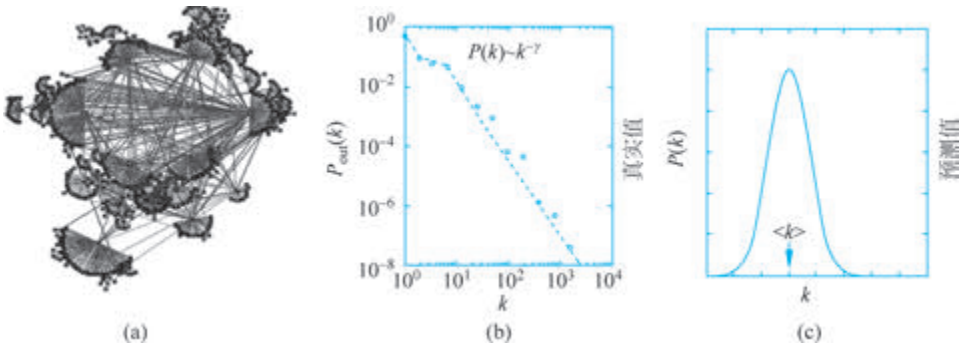


图 3.10　万维网网页链接网络示意图及节点度的真实分布（幂律分布）和预期分布（二项分布）[3]

经典的随机网络理论为 ER（Erdös-Rényi）网络[4]，由匈牙利数学家 Paul Erdös 和 Alfréd Rényi 于 1959 年提出。ER 网络通常用 $G(n, p)$ 表示，其中 n 为节点的数量，p 为每对节点之间边的连接概率。ER 网络的主要特点是随机性和均匀性连接。如图 3.11 所示，在 ER 网络中，每对节点之间是否存在一条边是根据概率 p 独立随机决定的。这意味着，ER 网络中的边分布是均匀的，每条边都有相同的概率被创建，节点度数也具有较小的方差。

ER 网络通常不具备小世界特性，因为平均最短路径长度通常较大。节点之间的连接是随机的，因此存在较大的平均路径长度。ER 网络的局部聚类系数通常较低，因为边的连接是独立随机的，不太容易形成高度聚类的社群。此外，ER 网络通常不具备无标度特性，因为节点度数的分布近似于泊松分布，与幂律分布不同，没有出现具有非常高度连接度的“枢纽节点”。因此，ER 网络模型并非一个合适的对真实网络的建模方式。

为了解决 ER 网络无法产生小世界属性的缺点，Duncan J. Watts 和 Steven H. Strogatz

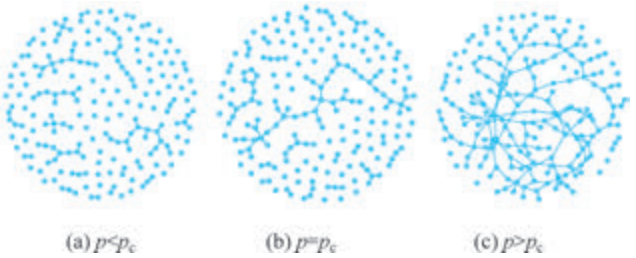


图 3.11　ER 随机网络示意图[4]

注：p 为任意两个节点间连边的概率。

在 ER 网络思想的基础上于 1998 年提出了 WS(Watts-Strogatz)网络模型[5]。WS 网络的生成过程如图 3.12 所示。在初始阶段，WS 网络模型以一个具有 N 个节点的正则图作为基础。在正则图中，每个节点都与其 k 个最近邻节点连接，其中 k 是一个参数。接下来，模型通过重新连接一些边来引入随机性。对于每个节点 i，以概率 p，它会选择将其与一个随机选择的不是它的邻居的节点 j 重新连接，以替换原来的边。这一步骤的目的是打破图的规则性，引入小世界特性。通过重复上述过程，直到所有节点都重新连接一次，就生成了 WS 小世界网络正规网络。在整个过程中 p 参数控制了边的重连概率，它是生成小世界网络的关键参数。

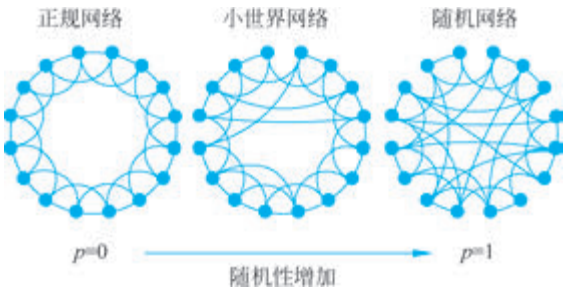


图 3.12　WS 随机网络示意图[5]

注：p 为边重连的概率。

WS 网络模型的关键特性包括小平均路径长度和高聚类系数。由于大多数节点通过少数几步或跳跃与其他节点相连，WS 网络具有较短的平均路径长度。这使得节点之间的信息传播速度很快。此外，WS 网络中的节点倾向于形成紧密的社区或集群，这导致了较高的聚类系数。还可以通过控制重连概率参数 p，在小世界网络的生成中引入不同程度的随机性。当 $p=0$ 时，网络接近于规则图；当 $p=1$ 时，网络变为随机图。WS 小世界网络模型是一种重要的网络生成模型，可用于研究小世界网络的性质，如短路径和高聚类系数，以及网络的演化和结构。这个模型在解释和模拟真实世界中的网络方面具有重要的应用。

然而由于最初的正则图具有固定的节点度，即每个节点都与其 k 个最近邻节点连接。通过随机地重连一些边，引入了一些随机性，但整体上的度分布不会呈现如真实网络一般的幂律分布，而是指数分布。在此基础上，科学家 Albert-László Barabási 和 Réka

Albert 在 1999 年提出了基于增长和优先依附的 BA(Barabási-Albert)网络模型[6]。与 ER 网络模型和 WS 网络模型不同,BA 网络模型采用逐步增长的方式生成网络。如图 3.13 所示,起初网络只包含少量节点(通常是两个或三个节点)。每当新节点加入网络时,该节点都将与 m 个已有节点相连,并且连接的概率与已有节点的度数成正比,即 $p_i=\frac{k_i}{\sum_j k_j}$。这意味着度较高的节点更有可能获得新连接,也就是"优先依附"的关键特性。随着网络不断增长,BA 网络的度呈现出幂律分布,可以证明节点度分布服从 $P(k)\sim k^{-3}$。图 3.13 展示了一个节点数为 200000,$m=2$ 的 BA 网络的度分布,其服从指数约为 −2.78 的幂律分布。这证实了网络中存在少数节点具有极高的度,而大多数节点具有较低的度,这赋予了网络无标度性。

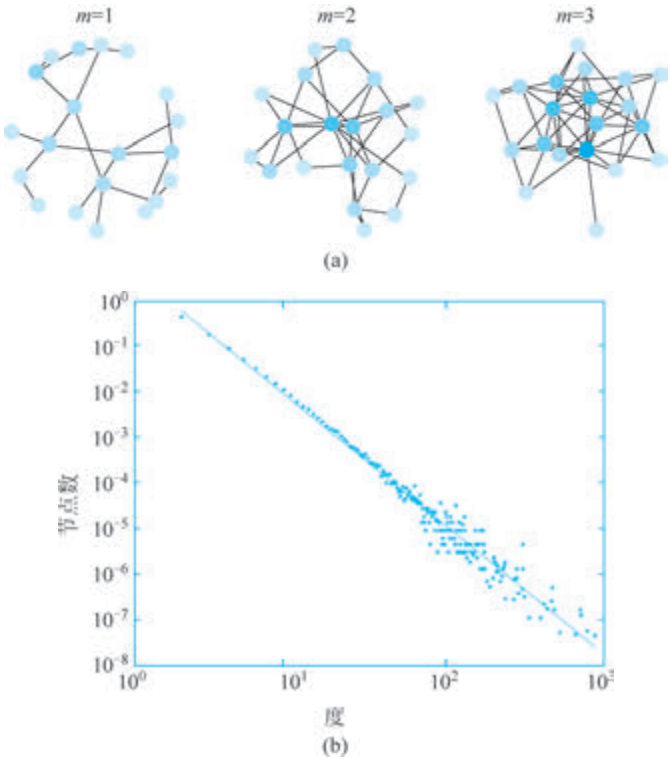


图 3.13 BA 网络示意图

注:m 为每次加入新节点的度。图(b)为具有 20 万个节点的 BA 网络的度分布。

BA 网络中蕴含的优先依附思想意味着节点的连接越多,接收新连接的可能性就越大。度数越高的节点抓取添加到网络中的新连接的能力越强。直观上,如果从连接人们的社交网络的角度来思考,从 A 到 B 的链接意味着 A"认识"或"熟悉"B。重度连接的节点代表有很多关系的知名人士。当新人进入社区时,他们更有可能熟悉那些更引人注目的人,而不是相对不知名的人。这种能够反映现实世界中优先依附现象的能力使得 BA 模型能够很好地刻画现实网络的性质。后来,BB(Bianconi-Barabási)模型通过引入"适应

度”参数来进一步刻画优先依附现象。该模型认为优先依附是正反馈循环的一个例子，其中最初的随机变化(一个节点最初拥有更多链接或比另一个节点更早开始积累链接)被自动加强，从而极大地放大了差异。这有时也称为马太(Matthew)效应，或者“富者愈富”。

BA 模型的提出和研究揭示了许多真实世界复杂网络(如互联网、社交网络、科学引文网络等)的无标度性质，这一发现在网络科学领域具有革命性意义。BA 模型还启发了许多后续研究，包括复杂网络的演化、网络动力学、信息传播、社会系统建模等方面的研究[7]。目前，后续研究提出了各种改进的 BA 模型，以考虑更多复杂的情况，如有向网络、多层网络、节点属性等。

本节简述了网络模型的概念、真实网络的重要性质以及网络科学研究中对建模真实网络特殊性质的多个模型。基于这些概述可以将网络模型应用于解析和利用移动数据中的信息传播、社交网络、位置轨迹等信息，以揭示移动数据中的隐藏规律和模式。同时，这也将有助于改进移动数据挖掘方法，进一步推动对移动行为和社交网络的深入理解。

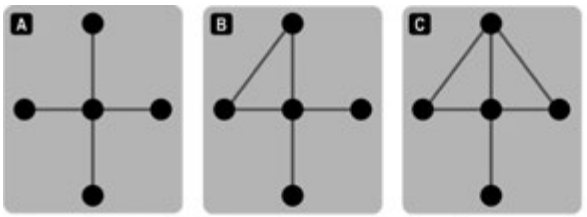
3.2 三角形计数算法

3.2.1 网络中的三角形

网络结构出现在包括但不限于社交、通信和信息科学的众多领域，尽管这些领域中的网络在节点组成方面有所不同，但某些拓扑结构，特别是三角形(网络中三个互相相邻的节点的集合)，在不同领域的网络中都大量存在。在现实生活中，三角形的丰富性激发了科学家发明了集聚系数和传递性比率等度量来表征和分析网络。社交网络中存在三角形也已从各种社会科学理论中进行了研究和解释，如同类亲和力和传递性。这些研究的一个关键计算任务是计算网络中的三角形数量，这是本节在网络分析部分要介绍的重点问题之一。

对网络中三角形计数有许多现实应用，其中著名的是计算网络的传递性比率，如图 3.14 所示，它定义为网络中三角形和含有长度为 2 的路径的三元组(可能产生三角形的三元组)的数量之比。鉴于可以简单地从网络的顶点度数计算三元组的数量，传递性计算实际上等同于计数三角形的任务。集聚系数是另一个类似的度量：对于节点 u，其集聚系数衡量 u 的邻居节点中有多少是彼此的邻居。集聚系数和传递性比率都已用作网络分析和网络演化模型的关键度量。三角形计数还用于其他一些有趣的应用。如使用网络中三角形分布来检测互联网垃圾邮件，因为垃圾邮件主机的三角形频率分布与非垃圾邮件主机的分布明显不同。三角形分布也用于揭示万维网中的隐藏主题结构，三角形密集的区域通常代表一个共同的主题。重叠的三角形结构还可以用于社群发现算法(见 3.3 节)。

尽管从算法上看三角形计数似乎是一项简单的任务，但多年来已吸引了来自包括数据挖掘和图论等不同领域的科学家的诸多贡献。早期的作品主要关注渐进计算复杂性，而近期的研究则更关注实际执行时间，其动机来自实际网络的巨大规模，其中顶点数量



	实际三元组数(t_0)	可行三元组数(t_1)	传递性(t_0/t_1)
A	0	4	0
B	1	4	0.25
C	2	4	0.5
D	3	4	0.75
E	4	4	1.0

图 3.14　网络的传递性比率

在数百万到数十亿。为了实现效率，通过抽样来进行三角形计数已经成为许多近期作品中非常活跃的方向。此外，研究人员还试图通过在多核或分布式环境中运行的算法来实现效率。一些三角形计数算法的变体也受到了数据访问约束的启发。例如，已经提出了针对与传统随机内存访问不同的各种数据访问场景的三角形计数算法。

三角形计数算法的计算复杂性是其效率的良好指标，但在实际生活中，即使两个算法具有相同的计算复杂性，它们的执行时间也可能差异很大。这一事实的主要原因是计算复杂性的隐藏常数，它取决于输入图的各种属性。稀疏性是其中一种属性。大型现实网络非常稀疏，其中边的数量通常是顶点数量的常数因子，换句话说，顶点的平均度是恒定的。另一个重要属性是现实网络的度分布是倾斜的。尽管网络的平均度是常数，但总会存在一些具有非常大度的顶点。这种现象通常称为幂律度分布，它显著影响了三角形计数算法的性能。

从以上重要应用中可以看出，三角形计数在网络分析中具有重要的应用价值，它可以帮助我们测量社区的成熟度、理解社区的发展趋势，并应用最优连接计算理论来优化查询处理，从而深入研究网络结构和关系。这对于各个领域，包括社交网络、社区发现、数据库管理和数据分析，都具有重要意义。本节将对网络中的三角形计数基础及一些进阶算法作简要介绍。

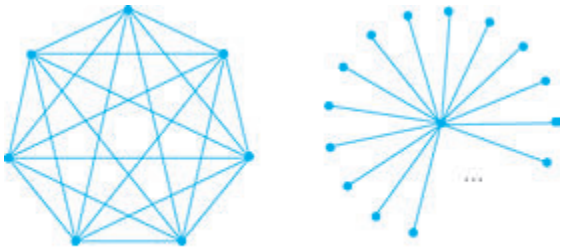
3.2.2　深入三角形计数算法

给定一个无向网络 $G=(V,E)$，三角形计数算法旨在找到网络中所有的节点三元组 $\{i,j,k\}$，其中 $i\in V, j\in V, k\in V$，满足 $(i,j)\in E, (j,k)\in E, (i,k)\in E$。用 N 表示网络中节点的数量 $N=|V|$，L 表示边的数量 $L=|E|$。

假设判断网络中一对节点间是否存在边的时间复杂度为 $O(1)$，即可以在常数时间内回答网络中的“边查询”操作。那么一种显然的计数算法是暴力搜索方案，即枚举网络中所有可能的节点三元组，并记录其中有多少个是三角形。网络中共有 C_n^3 组节点三元组，因此这种暴力搜索方案的时间复杂度为 $O(N^3)$。对于顶点数只有几百个的网络来

说，这种算法显然是可以接受的；但对于更大的网络，需要更好的算法。

是否可以在$O(N^3)$的运行时间内取得更好的效果？这取决于网络的性质。如图 3.15 所示，所有节点间都有边相连的完全图结构，具有$O(N^3)$个三角形。任何按顺序逐个计算三角形的算法，不论对暴力搜索算法进行何种改进，都注定在这样的网络上运行$O(N^3)$的时间。然而，上述暴力搜索算法在每个网络上都运行$O(N^3)$的时间，即使对于根本没有三角形的网络也是如此。因此，可以在许多网络上做得比暴力搜索算法更好。



(a) 完全图结构　　(b) 星形图结构

图 3.15　特殊的网络结构

一种更简洁的方法是枚举网络中所有长度为 2 的路径，而不是枚举节点三元组。也就是说，只检查已经知道包含两条边的节点三元组。具体来说，对于网络中每个节点$v \in V$，记v的所有邻居节点为$N(v)$，对于每对不同的邻居节点u、$w \in N(v)$，判断u、w间是否有边相连，若有，则$\{u,v,w\}$形成一个三角形。由于边查询的时间复杂度为$O(1)$，此算法在节点v处运行的时间复杂度为$O(k_v^2)$，其中k_v为节点v的度。在整个网络上运行的时间复杂度为$O\left(\sum_{v \in V} k_v^2\right)$。这个界限在某些情况下可能是$O(N^3)$，例如，如果每个顶点的度数都与$N$呈线性关系。然而，在每个顶点的度数都是常数的图中，算法具有线性的$O(N)$运行时间，这一复杂度相比$O(N^3)$有巨大提升。

从边的角度出发，还可以用类似的方式查找网络中的三角形。具体地说，对于网络中每条边$(u,v) \in E$，遍历网络中所有节点$w \in V$，判断u、w之间及v、w之间是否有边相连，若都有，则$\{u,v,w\}$形成一个三角形。由于边查询的时间复杂度为$O(1)$，此算法在每条边(u,v)的时间复杂度为$O(N)$。在整个网络上运行的时间复杂度为$O(N \cdot L)$。

在社交网络中希望有更好的三角形计数算法。万维网等网络的度分布服从幂律分布，即使网络的平均度可能是常数，但存在度非常高的“重尾”节点。作为一个极端的例子，考虑如图 3.15 所示的具有星形图结构的网络。网络中只有$N-1$条边，没有三角形，但上述两种算法的运行时间是$O(N^2)$，因为在度很高的中心节点处执行了大量工作。

以上两种方法在某些场景下的时间复杂度是相似的，并且对于网络中的所有三角形它们都将重复查找到该三角形三次。例如，从节点开始查找三元组的方法会在三个节点处都查找到三角形$\{u,v,w\}$，从边开始查找的方法也都会在三条边各查找到一次。因此，一种优化方案是对网络中的每个三角形只会遍历一次，这看起来似乎只能节省一个因子 3，但如果我们对选择用于计数的节点采取高效的策略，算法的执行时间可能被大大

缩短。关键是选取三角形中最有代表性的节点负责计数。

对于一个节点数为 N、边数目为 L 的网络来说，节点度的总和为 $2L$。若将节点数不低于 $\sqrt{L}$ 的节点定义为“重节点”，则“重节点”的数目不会多于 $2L/\sqrt{L}=2\sqrt{L}$ 个。首先对所有“重节点”的集合，在其上运行暴力搜索算法枚举所有“重节点”三元组并判断是否形成三角形，如此可以枚举得到所有三个节点均为“重节点”的三角形。由于“重节点”最多有 $2\sqrt{L}$ 个，暴力搜索算法的时间复查度不超过 $O(L^{1.5})$。对于网络中其他并非所有节点都是“重节点”的三角形，用如下方式对其进行搜索：对于网络中每条边，若其两个节点均为“重节点”，跳过；若其中至少有一个非“重节点”v，则对 v 的所有不超过 $\sqrt{L}$ 个邻居节点 $u\in N(v)$，判断 u 是否和这条边中不为 v 的另一个节点相连。如此，可以遍历得到网络中所有三角形。由于有 L 条边，这部分搜索的时间复杂度仍然为 $O(L^{1.5})$。因此，这种方法可以在 $O(L^{1.5})$ 的时间复杂度内得到网络中的所有三角形。是否还有更优的计数方法呢？对于任意的 N 和 L，都可以构造出一个有 N 个节点、L 条边，并且有 $O(L^{1.5})$ 个三角形的网络。在这种情况下具有 $O(L^{1.5})$ 的时间复杂度的算法是最优的。更快的方法需要对常数项进行优化，如进一步减少同一三角形被重复查找的次数，本案例中每个所有节点均为非“重节点”的三角形仍将被重复查找三次。

重节点算法流程如表 3.1 所示。

表 3.1　重节点算法流程

输入：节点数 N，边数 L
输出：网络中的三角形数量 Δ
$\Delta\leftarrow 0$
重节点←度超过 $\sqrt{L}$ 的节点
对于每个“重节点”v_1，循环遍历：
　对于每个“重节点”v_2，循环遍历：
　　对于每个“重节点”v_3，循环遍历：
　　如果 $v_1<v_2<v_3$ 并且 v_1、v_2、v_3 之间都有边相连：
　　$\Delta\leftarrow\Delta+1$
对于网络中的每条边(v,w)，循环遍历：
　如果 v 和 w 都是重节点，那么跳过此边；
　否则，对于 v 的每个度不超过 $\sqrt{L}$ 的邻居节点 $u\in N(v)$，循环遍历，即
　　如果 u 和 w 之间有边相连：
　　$\Delta\leftarrow\Delta+1$
返回 Δ

对于拥有数亿条边的非常大的网络，如社交媒体上的好友关系形成的网络，$O(L^{1.5})$ 的时间复杂度仍可能认为是昂贵的。因此，近年来，近似三角形计数算法流行起来。近似三角形计数方法并不枚举出网络中所有的三角形，而是提供三角形的近似计数，有时还带有近似保证。此外，它们的执行时间更短，通常能够在若干数量级差距下得到令人满意的结果。对于许多应用程序来说，以较长的运行时间换取良好的近似是足够的。

本节简要介绍基于网络稀疏化的三角形计数方法。这种方法的核心思想是以一定概率删除网络中的一部分边(图 3.16),然后从稀疏网络中使用先前介绍的方法精确计数三角形来推断原始网络中的三角形数目。由于稀疏网络比原始网络小得多,因此在稀疏网络中的三角形计数通常可以在更短的时间内完成,使得近似方法比精确计数方法快得多。这种方法也可以视为基于均匀三角形采样的方法,因为稀疏网络中的三角形是根据原始网络中所有三角形的均匀概率采样的。

图 3.16　网络的稀疏化:原始网络和稀疏网络[8]

Tsourakakis 等提出了一种最早的基于网络稀疏化的近似三角形计数方法 DOULION[9],其算法流程如表 3.2 所示。给定一个网络 $G(V,E)$,DOULION 以概率 p 保留 G 的每条边,并以概率 $1-p$ 删除边,以生成稀疏网络 G_s。然后,在 G_s 上运行精确的三角形计数方法,得到 G_s 的精确三角形计数 $t(G_s)$。只要其三条边都保留,原始网络中的每个三角形都会在稀疏网络中保留,此概率为 p^3。从 G 中采样一个三角形在 G_s 中的概率为 p^3,因此原始网络中的三角形的期望计数为

$$\hat{t}(G)=\left(\frac{1}{p^3}\right)t(G_s)$$

对于具有数百万个顶点的大型网络,p 低至 0.01 都可以提供非常好的近似三角形计数,而这将使运行时间几乎提高 100 倍。

表 3.2　DOULION 算法流程

输入:无权图 $G(V,E)$,稀疏化参数 p
输出:G 的全局三角形估计值 $\Delta'(G)$
对于 G 中的每条边 e_j 执行以下操作:
a. 抛掷一个偏置的硬币,成功概率为 p
b. 如果成功,则设置 $w(e_j)\leftarrow 1/p$
c. 否则,设置 $w(e_j)\leftarrow 0$
使用三角形计数算法计算 $\Delta'(G)$
$\Delta'(G)\leftarrow\Delta'(G)/p^3$
返回 $\Delta'(G)$

Etemadi 等对 DOULION 算法进行改进[10],其算法流程如表 3.3 所示。与 DOULION

类似，该方法还使用均匀概率 p 对给定网络 G 的边进行采样，以获得稀疏网络 G_s。然而，除了在 G_s 中计算三角形之外，它还检查 G_s 中每个开放三元组（仅存在两条边的三元组）的缺失边是否存在于原始网络 G 中，若该边存在，则将这一三角形计数到 $t(G_s)$。在 G 中的三角形以 p^2 的概率在 $t(G_s)$ 中计数，因为只要 G 中的三角形的两条边保留在 G_s 中，它们就会在 $t(G_s)$ 中被计数。因此，使用这种方法计算原始网络中的三角形的期望计数为

$$\hat{t}(G)=\left(\frac{1}{p^2}\right)t(G_s)$$

可以证明，这种方法相比 DOULION 原始方法只需更少的样本就能达到相同的准确性水平。但需注意的是，本方法的时间成本更高，因为除了在稀疏网络 G_s 中计数三角形之外，它还需要检查原始网络 G 的结构，以确定 G_s 中开放三元组的缺失边是否存在。

表 3.3　改进 DOULION 算法流程

输入：无权图 $G(V,E)$，稀疏化参数 p
输出：G 的全局三角形估计值 $\Delta'(G)$
对于 G 中的每条边 e_j 执行以下操作：
a. 抛掷一个偏置的硬币，成功概率为 p
b. 如果成功，则设置 $w(e_j)\leftarrow 1/p$
c. 否则，设置 $w(e_j)\leftarrow 0$
使用三角形计数算法计算子图中的三角形数量 $\Delta'(G)$
对于子图 g 中的每个开放三元组，检查其在原图 G 中的闭合情况，若闭合，$\Delta'(G)\leftarrow\Delta'(G)+1$
$\Delta'(G)\leftarrow\Delta'(G)/p^2$
返回 $\Delta'(G)$

3.3 社群发现算法

3.3.1 社群发现与模块度

网络科学研究中通常会在真实世界的网络中发现一些共同的特征，如小世界属性、无标度律以及节点集聚等，其中一个重要的特征是网络的社群结构，它指的是网络中存在一些节点的组，这些组内部的连接比与网络的其他部分连接更紧密，如图 3.17 所示。这种连接的不均匀性表明网络内部存在某些自然的分割。社群通常是根据节点集的划分来定义的，本节讨论每个节点都被放入一个且仅一个社群的情况，如图 3.17 所有划分方式一样。这是一种有用的简化，大多数社群发现方法都会找到这种类型的社群结构。

在网络中找到潜在的社群结构具有许多重要意义。社群可以用于创建网络的大规模地图，因为个体社群在网络中起到了类似元节点的作用，从而使研究变得更容易。个体社群还可以揭示由网络表示的系统的功能，因为社群通常对应于系统的功能单元。在代谢网络中，这样的功能群对应于循环或路径，而在蛋白质相互作用网络中，社群对应于在生物细胞内具有类似功能的蛋白质。同样，引文网络按研究主题形成社群，能够识别

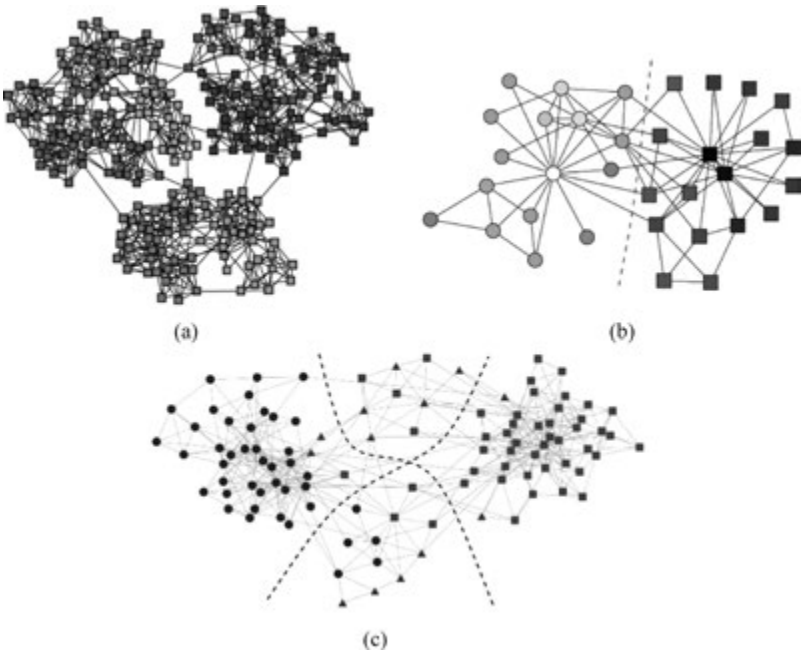


图 3.17　网络的社群结构及社群发现示例

网络内的这些子结构可以洞察网络功能和拓扑如何相互影响。另外，社群通常具有与网络的平均性质非常不同的性质。因此，仅关注平均性质通常会忽略网络内许多重要和有趣的特征。社群的存在通常也会影响网络上发生的各种过程，如谣言传播或流行病传播。因此，正确理解这些过程，重要的是检测社群，以及研究它们如何在不同情境中影响传播过程。

用来衡量网络社群结构强度的一个重要指标是模块度 Q，其由 Mark Newman 提出。具有高模块度的网络在社群内的节点之间具有密集的连接，而在不同社群之间的节点之间具有稀疏的连接。在经典的定义中，模块度是指落入给定社群的边的比例与在边随机分布下的期望比例间的差异。模块度的定义如下：

给定网络 $G=(V,E)$ 和对 V 的一种划分方式 $S=\{s_1,s_2,\cdots,s_m\}$，其中 s_i 为 V 的一个非空子集，对任意 $1\leqslant i<j\leqslant m$，都有 $s_i\cap s_j=\varnothing$，且 $s_1\cup s_2\cup\cdots\cup s_m=V$。在此划分下，模块度

$$Q(G,S)=\frac{1}{2m}\sum_{s\in S}\sum_{i\in s}\sum_{j\in s}\left(A_{ij}-\frac{k_ik_j}{2m}\right)$$

式中：A_{ij} 表示节点 i 和节点 j 间是否有边相连；k_i 为节点 i 的度，m 为网络中边的总数。

在随机情况下，具有度 k_i 和 k_j 的节点间有边相连的概率为 $\frac{k_ik_j}{2m}$，因此上述定义可以体现社群中节点相比于随机分布时的连接强度差异。对于无权网络和无向网络，其模块度的取值为 $-1\sim1$。如果社群内边的数量超过了基于随机分布预期的数量，那么模块度

将为正值。一般认为模块度达到 0.3～0.7 代表已经有显著的社群结构。

3.3.2 Louvain 算法

模块度定义了网络社群结构的强度，为了得到具有更高模块度的社群分割方式，科学家提出了许多算法，Louvain 大学的 Vincent Blondel 等于 2008 年提出的 Louvain 算法是网络社群结构发现的经典算法之一[11]，该方法是一种贪婪优化方法，在节点数为 N 的网络上具有 $O(N\log N)$的时间复杂度。

Louvain 算法分为两个迭代阶段：第一阶段通过仅对节点在社群间进行本地移动来优化模块度；第二阶段将已识别的社群进行聚合。迭代重复进行，直到模块度不再增加为止。算法的整体流程如图 3.18 所示。

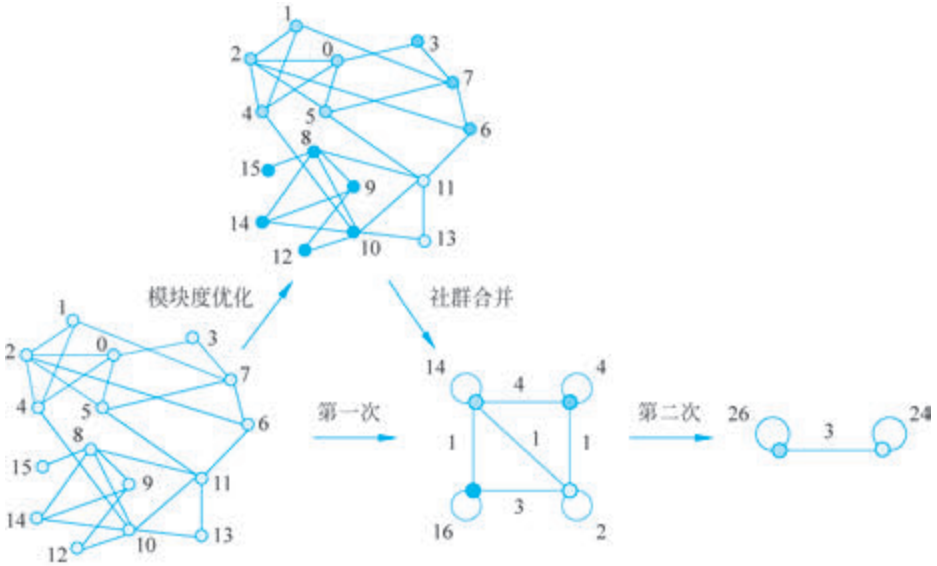


图 3.18 Louvain 算法的流程示意图[11]

假设从一个具有 N 个节点的加权网络开始。首先为网络中的每个节点分配一个不同的社群，该社群内有且仅有此一个节点。因此，在这个初始分区中，社群的数量与节点的数量相同。然后，对于每个节点 i，考虑 i 的邻居 j，评估将 i 从其社群中移除后并入 j 所处社群获得的模块度增益。再把节点 i 放置在模块度增益最大的社群中（如果有多个最大增益的社群，那么可以使用一个规则来选择），但只有在正增益的情况下才会这样做。如果不存在正增益，节点 i 将保持在其原始社群中。此过程对所有节点重复且顺序执行，直到无法进一步提高模块度为止，第一阶段完成。需要强调的是，一个节点可能被多次考虑。当没有个体移动可以改善模块度时，认为社群分割情况已达到模块度的局部极大值。如图 3.18 所示，在第一阶段算法将 0、1、2、4、5 号节点合并为一个社群，3、6、7 号节点合并为第二个社群，8、9、10、12、14、15 号节点合并为第三个社群，11、13 号节点合并为第四个社群。

上述算法需要不断地计算将节点 i 移入某社群 C 所获得的模块度增益：

$$\Delta Q=\left[\frac{\Sigma_{\text{in}}+k_{i,\text{in}}}{2m}-\left(\frac{\Sigma_{\text{tot}}+k_i}{2m}\right)^2\right]-\left[\frac{\Sigma_{\text{in}}}{2m}-\left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right]$$

式中，Σ_{in} 为社群 C 内部边的权重之和；Σ_{tot} 为与社群 C 中的节点相连的边的权重之和；k_i 为与节点 i 相连的边的权重之和；$k_{i,\text{in}}$ 为从 i 到社群 C 中节点的边的权重之和；m 为网络中所有边的权重之和。

类似的表达式也可以用于评估当 i 从其社群中移除时模块度的变化。在实践中，通过将 i 移入相邻的社群中来评估模块度的变化。

算法的第二阶段包括构建一个新网络，该网络的节点现在是在第一阶段找到的社群。为此，新节点之间边的权重由对应的两个社群中的节点之间的链接的权重之和给出。同一社群节点之间的边导致新网络中该社群对应节点存在自环(与自身相连的边)。如图 3.18 所示，在第一阶段将网络分成四个社群后，可以聚合得到一个有四个节点的网络，各边的权重均已标出。构建新网络后，可以重新将算法的第一阶段应用于所得到的加权网络并进行迭代。若用“轮”来表示这两个阶段的组合，则每轮后中社群的数量都会减少，因此大多数的计算时间都用在第一个迭代上。这些轮将会被反复迭代，直到模块度不再发生变化，达到局部最大值。Louvain 算法类似于网络的自相似性，并在过程中自然地融入了层次结构的概念，并且构建的层次结构的层数由轮的数量决定，并且通常是一个较小的数量。

Louvain 算法具有的优点：首先，它的步骤直观且易于实施，而且结果是无监督的。其次，Louvain 算法非常快，即在大型特定的稀疏数据上进行的计算机模拟表明，它的复杂性在典型和稀疏数据上是线性的。这是因为可能的模块度增益很容易使用公式计算，并且在只经过几轮之后，社群的数量会急剧减少，大部分运行时间集中在最初的迭代上。模块度的分辨率限制问题也因算法的固有多层性质而被规避。事实上，一般模块度优化无法识别小于某个尺度的社群，从而在纯模块度优化方法检测到的社群上引入了分辨率限制。在 Louvain 算法中此情况不一定会出现，因为算法的第一阶段涉及将单个节点从一个社群移到另一个社群。因此，通过逐个移动节点来合并两个不同的社群的可能性非常低。这些社群可能在以后的轮中被合并。Louvain 算法还提供了将网络分解为不同层次组织的社群的解决方案，这表明算法找到的中间解也可能是有意义的，而且揭示了发现的层次结构可能允许最终用户放大网络并以所需的分辨率观察其结构。

Louvain 算法也存在相应的缺点：首先，它是一种贪心算法，通过不断地优化每个节点的社群分配来寻找局部最优解。这种局部性质导致其不能保证找到全局最优的社群结构。其次，Louvain 算法对初始条件敏感，它在迭代过程中依赖初始的社群分配，不同的初始条件可能导致不同的最终结果。虽然 Louvain 算法通常对大规模网络具有较高的效率，但在某些情况下它可能需要大量的内存来存储网络的表示，限制了其在资源有限的环境中的使用。但从整体来看，Louvain 算法仍然是网络科学中社群发现的一个重要手段。

3.3.3 基于 PageRank 的社群发现算法

尽管 Louvain 算法能够在时间复杂度 $O(N\log N)$ 内发现网络中的社群结构，但在 N

很大的超大规模网络中使用 Louvain 算法仍然十分吃力。因此，对超大规模网络进行社区发现时可以从局部网络入手。本节将介绍如何从某个节点出发，利用 PageRank 算法的思想在局部网络中找到该节点所在的社区。可以证明这种算法的时间复杂度与网络中社群的数量呈线性关系，这在超大规模网络中将大大提升算法速度[12]。

PageRank 算法最初由 Larry Page 和 Sergey Brin 开发，是一种用于网页排名的算法。其核心思想是通过分析网页之间的链接结构确定网页的重要性和排名。PageRank 算法认为，一个网页的重要性可以通过其被其他网页链接的数量和质量来衡量，如图 3.19(a)所示。具体来说，如果一个网页被许多其他重要的网页链接，那么它本身也应该被视为重要的。此外，从链接到一个网页的其他网页的重要性也会影响该网页的重要性。PageRank 使用迭代计算的方式，不断更新每个网页的重要性分数，直到达到稳定状态，具有稳定 PageRank 分数的网络如图 3.19(b)所示。

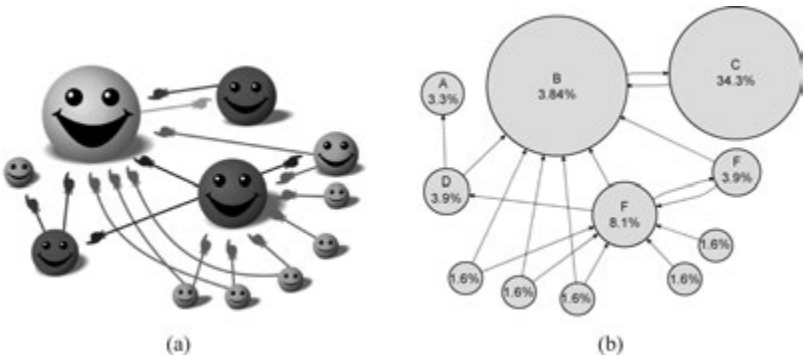


图 3.19　PageRank 思想示意图及一个示例网络的 PageRank 分数

在对网络进行社群分割时，可以利用 PageRank 分数辅助发现能够形成紧密社群的节点集合并且找到合适的分割不同社群的边。整体思路：对于一个网络，其 PageRank 向量是从指定的初始分布开始进行随机游走序列得到的概率分布的加权和。通过对 PageRank 向量进行扫描操作可以找到这个切割，这一步骤涉及按照 PageRank 向量顺序检查网络节点，并计算由此顺序产生的每个集合的传导率，随后选取传导率最低的集合进行社群的分割。

具体来说，对于一个无向无权网络 $G=(V,E)$，用 $d(v)$ 表示节点 v 的度。并令 $\boldsymbol{D}$ 为度矩阵，其中 $D_{i,i}=d(v_i)$。$\boldsymbol{A}$ 为邻接矩阵，当且仅当 v_i 和 v_j 之间存在边时，$A_{i,j}=1$。当考虑节点集 V 上的向量时，将它们写成行向量的形式，因此向量 $\boldsymbol{p}$ 与矩阵 $\boldsymbol{A}$ 的乘积将写作 $\boldsymbol{pA}$。PageRank 向量 $\mathrm{pr}_\alpha(\boldsymbol{s})$ 定义为如下线性系统的唯一解：

$$\mathrm{pr}_\alpha(\boldsymbol{s})=\alpha\boldsymbol{s}+(1-\alpha)\mathrm{pr}_\alpha(\boldsymbol{s})\boldsymbol{W}$$

式中：α 为传送概率 $\alpha\in(0,1]$；$\boldsymbol{s}$ 为起始向量；$\boldsymbol{W}$ 为惰性随机行走转移矩阵，$\boldsymbol{W}=\frac{1}{2}(\boldsymbol{I}+\boldsymbol{D}^{-1}\boldsymbol{A})$。

在初始的 PageRank 定义中，起始向量为均匀的，$\boldsymbol{s}=\boldsymbol{1}/n$。由不均匀起始向量得到的 PageRank 向量称为个性化 PageRank 向量，并用于提供个性化搜索排名等领域。为了减

少计算复杂度，在局部网络上进行社群分割可以采用仅有单一节点对应的值不为 0 的起始向量，这等价于从该向量进行随机游走。已有一些算法提出在大规模网络中使用近似方法快速计算。

接下来定义网络中节点集合 S 的一些属性，注意社群也是网络中节点的集合。首先，S 的体积 $\mathrm{vol}(S)=\sum_{x\in S}d(x)$，那么整个网络的体积为 $2m$，m 为边的总数。S 的边界 $\partial(S)=\{\{x,y\}\in E\mid x\in S,y\notin S\}$。$S$ 的传导率为

$$\Phi(S)=\frac{|\partial(S)|}{\min(\mathrm{vol}(S),2m-\mathrm{vol}(S))}$$

形成社群结构的集合有着更少的外部连接和更多的内部连接，对应更低的传导率。如图 3.20(a)所示，左侧网络比右侧网络的传导率更高。扫描是一种从向量生成切割的技术，给定一个在有 N_p 个节点的局部网络随机游走得到的个性化 PageRank 向量 $\boldsymbol{p}=\mathrm{pr}_\alpha(\boldsymbol{s})$，令 $v_1,\cdots,v_{N_p}$ 为节点个性化 PageRank 向量值从高到低的顺序，即 $p(v_i)\geqslant p(v_{i+1})$。对于$\{1,\cdots,N_p\}$的每个整数 j 生成集合 $S_j^p=\{v_1,\cdots,v_j\}$，称为扫描集。令 $\Phi(\boldsymbol{p})$为这些扫描集中传导率最小的值，即

$$\Phi(\boldsymbol{p})=\min_{j\in[1,N_p]}\Phi(S_j^p)$$

如图 3.20 所示，当节点数量为 18 时达到最低的传导率。由于社群对应最低传导率，找到 $\Phi(p)$并进行相应的切割(保留 $\Phi(p)$对应的 S_j^p 中的节点集合)即可得到一个社群，这可以在时间复杂度 $O(\mathrm{vol}(\{v_1,\cdots,v_{N_p}\})+N_p\log N_p)$内完成。相比于 Louvain 算法，基于 PageRank 的社群发现方法能够更好地处理大规模网络。在面对不同的网络拓扑及规模时，可以根据具体情况选取合适的社群发现方法。

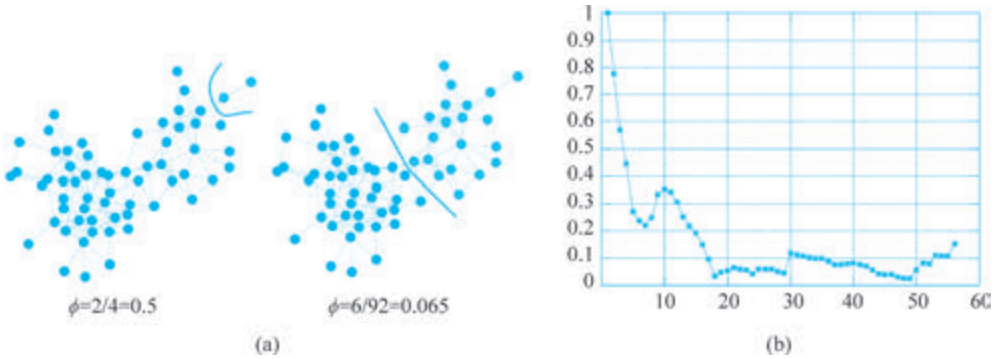


图 3.20　传导率的定义及在一次扫描过程中不同集合大小对应的传导率

3.4　案例一：移动行为模式分析

3.4.1　问题描述

现代社会和环境都是由人们不同尺度上的移动模式塑造的。长时间和长距离的出行通常包括国际飞行或城市之间的移动等罕见和不经常发生的事件。短时间出行主要

包括城内出行，如通勤上班或购物。这些出行表现出高度的规律性，通常遵循每日的生物钟节律。出于对了解全球流行病传播的动机，对大尺度上的人类移动性的研究已经揭示了底层移动模式的有趣特性。如今，大尺度的人类移动模式通常由出行距离分布 $p(r)$、旋转半径 $r_g(t)$ 以及随时间变化的访问地点数 $S(t)$ 三个广泛接受的指标来描述。

整个人口的出行距离分布一般遵循幂律分布 $p(r)\sim r^{-\beta}$，其中 $\beta\approx 1.59$。从移动电话数据中提取个体轨迹使得研究个体访问的区域成为可能，其特征由回转半径 $r_g(t)$ 来描述。个体的 r_g 可以理解为个体在给定时间段 t 内旅行的特征距离。回转半径分布显示了人口的异质性；大多数个体在短半径范围内旅行，但其他人则常在长距离上旅行。因此，每个个体在他的特征距离 $r_g(t)$ 内遵循 $p(r)$。在人口内部，r_g 的分布 $p(r_g)$ 导致了在聚合的出行距离分布 $p(r)$ 中观察到的幂律。对访问的地点的频繁返回由随时间变化的访问地点数 $S(t)$ 捕获。这个值呈亚线性增长，$S(t)\sim t^{-\mu}$，其中 $\mu=0.6$，捕获了个体重返地点的倾向。

这三个指标包含了描述个体轨迹的基本要素，人们经常在有限数量的地点之间出行，较少地在个体半径之外的新地点出行。这种长时间尺度上的行为可以通过具有位移分布作为输入的探索和有选择性返回模型来复现，该模型可用于模拟航空网络上的流行病传播。然而，当前的模型旨在捕捉长期的移动行为。例如，对于 $t<24$h，访问地点数 $S(t)$ 并不显示出稳定的缩放指数 μ。此外，旋转半径只在观察几个月后才稳定下来。因此，在建模城际和城内移动时存在不同的基本机制。

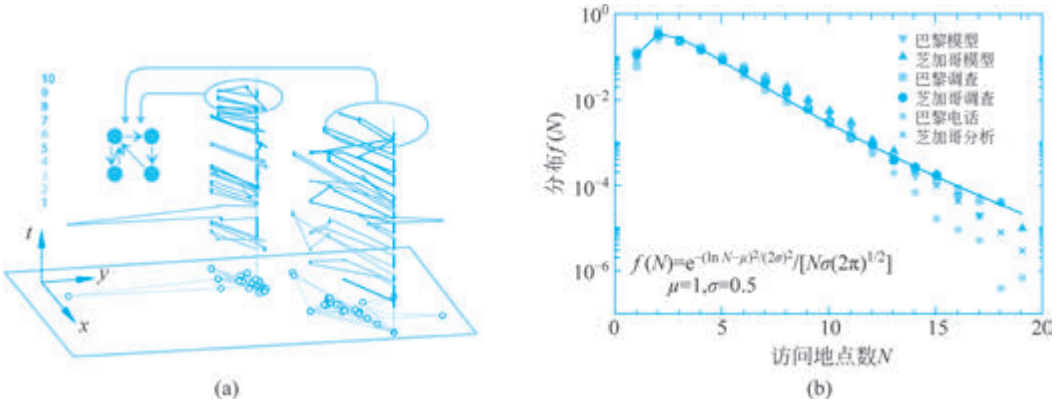
本应用案例源通过结合不同大规模数据来源构建人类移动网络，研究日常人类移动模式的共同基本机制。在各尺度、不同城市的数据集中，采用统计方法记录网络中不同主题结构的出现概率分布，观察到移动网络中普遍存在的日常移动模式，并进一步研究能否设计基于规则的模型来复现从移动网络中观察到的统计分布。

3.4.2 结果分析

本案例使用了三种不同的数据集：来自巴黎的移动调查和手机话费数据，以及来自芝加哥的移动调查数据。在调查中选择了芝加哥和巴黎整个人口代表性数据。在芝加哥的调查中，每个参与者回答了一两个整天的活动信息调查表，包括工作日、持续时间、地点、原因和出行方式。基于此信息可以重现匿名个体的整个日常活动模式。巴黎的调查具有相同的信息，但提供的是行程长度而非地理位置。从数百万手机用户的电话费数据中，根据基站位置、事件时间和用户识别号等信息，可以重建用户在 6 个月内的每日移动网络。

人类的移动特征是一系列访问的地点以及它们之间的出行。图 3.21 展示了两个用户个体的聚合移动轨迹以及在为期 10 天的观察期内对应的每日轨迹。不同日期的轨迹从底部到顶部以褐色(第 1 天)到红色(第 10 天)进行着色。xy 平面上的黑色圆圈和灰色线是每日轨迹的投影。每日和聚合的轨迹都可以描述为有向网络，其中节点代表访问的地点，有向边表示它们之间的出行。为了在每日基础上对这些网络进行分类，进一步舍弃了关于活动目的、出行时间、活动持续时间以及访问地点之间的距离等额外信息，因

此节点和边都没有加权。仅通过边的方向来体现出行的方向，如图 3.21 中构建的最后一天的移动网络(红色)。



彩图

图 3.21　两个用户 10 天的移动轨迹及移动网络示意图及个体每日访问地点数 N 的分布[13]

在此移动网络上首先研究不同访问地点数量的分布，即每日网络整体规模的分布。如图 3.21(b)所示，网络的规模分布 $f(N)$对于所有数据集都相似。观察到的分布 $f(N)$的形状可以近似为对数正态分布：

$$f(N) \sim \exp\left(-(\ln N-\mu)^2/(2\sigma^2)\right)/(\sqrt{2\pi}N\sigma)$$

式中：$\mu=1\pm0.1$；$\sigma=0.5\pm0.1$；平均访问地点数$\langle N\rangle$约为 3。

因此，大多数人每天只访问几个地点。事实上，90%的人口每天只访问不到 7 个地点。在不同城市，无论数据来源是出行调查还是手机数据，$f(N)$在所有数据集都遵循相同的分布。

为进一步研究观察到的每日移动模式，案例中调查了每日网络的数量。这些网络揭示了人们是喜欢在返回起始地点之前单一循环旅行中访问不同地点，还是更喜欢在返回起始地点之后访问其他地点。实际上，对于给定的网络规模 N，存在 $N_p(N)=2^{N^2-N}$ 种边的组合。为了描绘人类每日出行的网络，合理网络的数量可以通过睡眠的需要和行程的一致性两个主要约束条件大大减少。睡眠的需要要求出行在同一地点开始和结束，最有可能是在家里。一致性确保每个 N 个地点至少被访问一次。这两个条件意味着：对于 $N>1$，所有节点至少有一个入边和一个出边。通过计算满足这两个约束条件的可行每日网络数量 N_f 随着地点数量迅速增加($N_f(1)=1$，$N_f(2)=1$，$N_f(3)=5$，$N_f(4)=83$，$N_f(5)=5048$，$N_f(6)=1047008$)。在实际移动网络中，高达 90%的每日出行可用 17 种网络结构来描述，本应用案例将 17 种每日网络称为“主题”，利用这些主题构建整个人口的聚合移动网络。应用三角形发现算法可以快速找出网络中具有三角形拓扑结构的主题，并设计类似的算法发现其他多边形拓扑结构的主题。

图 3.22 中比较了在芝加哥和巴黎的调查、巴黎的手机数据及本案例所提模型获得的主题分布，这些主题按大小和出现频率排序。尽管数据源涵盖了来自不同国家的不同城市，但观察特定主题的频率行为相似。因此，可以假设主题是常见且通用的每日移动

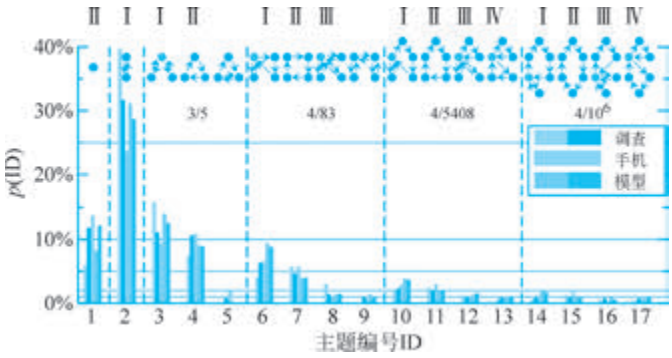


图 3.22　移动网络中 17 种主题出现频率的分布及相应的规则[13]

特征，进一步用于建模和模拟城市活动。常见的主题（ID 2）由两个访问地点和两个地点之间的出行组成，其次是仅有一个地点的主题（ID 1）。接下来的主题包括具有四次出行的三个地点，这些出行都在同一地点开始和结束（ID 3），或者具有一次往返出行（ID 4）。有趣的是，在所有数据集中没有观察到具有大小 N 和多于 $N+2$ 次出行的主题。

除了 ID 5 和 ID 9 外的主题最多只有一个具有两个以上有向边的节点的中心地点，这个中心节点是遍历路径 $T(x)$的起点，即在返回起始地点之前访问 x 个其他地点的出行，其中 $x<N$。存在唯一的中心节点确保了主题的边属于仅有一个遍历路径。因此，多次沿着相同有向边出行被抑制。整个主题由一个欧拉环组成，即可以准确地访问所有边一次，该路径以起始节点结束。基于上述定义，主题可以根据四个规则进行分类：规则Ⅰ，$T(1)$和 $T(N-2)$；规则Ⅱ，$T(N-1)$；规则Ⅲ，$T(1)$，$T(1)$和 $T(N-3)$；规则Ⅳ，$T(2)$和 $T(N-3)$。每个主题对应的规则写在图 3.22 的顶部。对于给定地点的数量 N，观察到某个主题的可能性与对应的规则有关，最有可能出现的主题可以用规则Ⅰ来描述。对于 $N\leqslant 6$，遍历路径的上限是三次，因此主题越大，遍历路径次数就越多。为分析个体是否具有典型的每日移动主题，基于手机数据研究了个体主题之间的相关性：

$$C_{ij}=\begin{cases}\dfrac{N(i)N(j)}{N_r(i)N_r(j)}-1, & \dfrac{N(i)N(j)}{N_r(i)N_r(j)}>1\\[2ex] 1-\dfrac{N_r(i)N_r(j)}{N(i)N(j)}, & \dfrac{N(i)N(j)}{N_r(i)N_r(j)}\leqslant 1\end{cases}$$

式中：ID 为 i 的主题观测次数为 $N(i)$，平均值为 $N_r(i)$。

6 个月内移动网络主题间的相关性分布如图 3.23 所示。

每个主题的最高相关性是自相关 C_{ii}，通常比期望相关性高 10～30 倍。对于每个主题，发现一个具有类似访问地点数量的主题（变化在±2 个地点之内）的概率与平均值相似。但对于更大的差异，概率显著受到抑制。此外，$N>4$ 的活跃用户似乎在整个观察期内都保持活跃，因为他们访问 $N>4$ 的任何主题的概率都较高。有趣的是，在 N 为 4、5、6 的主题结构内，一些相关性被抑制或增强。如果两个图论结构遵循相同的规则但访问的位置数不同，如在一次旅行中访问所有位置，那么它们之间的相关性会增强。如果主题结构不太相似，如旅行次数相差超过一个单位，那么相关性会被抑制。这实际上对应

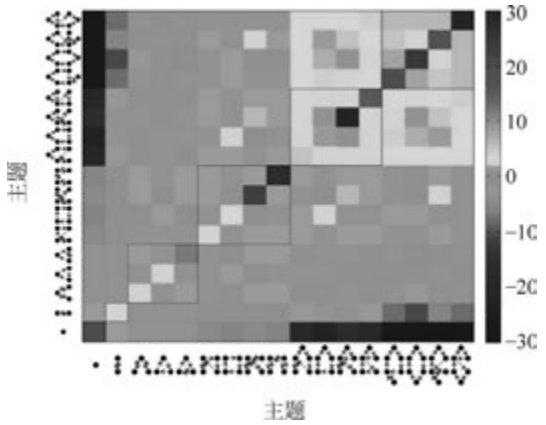


图 3.23　6 个月内移动网络主题间的相关性分布[13]

到根据规则Ⅱ和规则Ⅲ创建的主题。总的来说，主题不是唯一的，因为一个人可能在一天内重复遍历多次。然而，重复遍历是不常见的，因此一条边对应于一次出行。这些对移动网络上的主题分析表明，尽管访问的地点可能会改变，但是每人每天只有一个特征主题。因此，用户每天都有一个固定数量的首选地点，最有可能以特定顺序访问，这由其特征主题确定。

几乎整个人口的移动行为都可以用几个独特的主题来描述。为了理解这一观察结果，进一步研究在特定地点停留的时间，以及在一项活动的开始时间和同一类型下一项活动之间的时间。从两次调查数据中提取了在家庭、工作和其他三组活动中停留特定时间段的频率，如图 3.24 所示。工作和在家中的时间相对平均分布，工作时间大约为 3.5h 在家中的时间约为 8.6h，其他地点的活动的概率随其持续时间而减小。这种停留时间分布没有特定的持续时间，表明地点的变化不是均匀分布在时间上，而是在休息期间穿插在一起。为了支持这一观察，图 3.24 展示了两个相似活动之间的时间。虽然基于家庭和工作的时间受日常常规的影响，但其他地点之间的时间遵循广泛的分布。

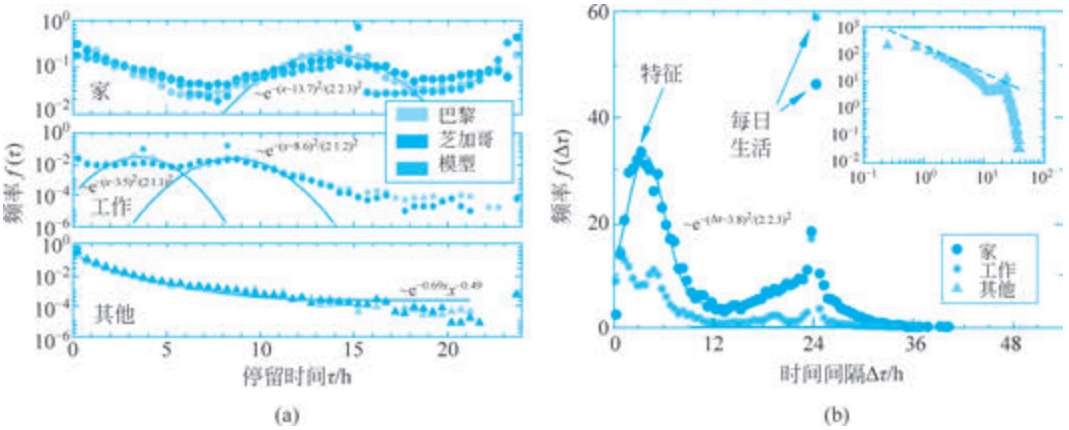


图 3.24　个体在家庭、工作地点及其他地点的停留时间分布及两次活动之间时间间隔分布[13]

受到上述观察的启发，本应用案例中进一步设计了一个基于扰动的模型，不仅可以复制观察到的每日主题，而且可以复制它们的发生频率。非工作代理(NW agent)模型的设计：考虑到家庭和其他地点之间的差异，该模型假定代理在家中有一个固定的活动，而在其他地方可以有任意数量的灵活活动(购物、娱乐等)。代理更倾向于待在家中，将其他活动视为一种干扰，因此在完成灵活活动后，如果他们没有其他灵活活动安排，将会返回家中。另外，当人们已经受到干扰时，更有可能在之后进行另一项灵活活动(如在城市中用餐后到访附近的酒吧)。在模型中，一天被分成 $K=48$ 个 30min 的时间段，在每个时间段中，代理根据相应的时间依赖概率 $p_{\mathrm{NW}}(t)$接收一个任务，并将其分配给下一个空闲时间段。最初，除了晚上的 9h 睡眠时间段，所有时间段都是空闲的。由于大多数任务发生在白天并在白天执行，可以假设接收任务的概率与昼夜节律有关。这个节律由整个人口的归一化手机活动 $p(t)$来近似表示。模拟观察到的调查和手机数据中的模式最重要的要素是假设在接收任务 $p_{\mathrm{NW}}(t)=p(t)$后，获得下一个任务的概率 $p_{\mathrm{NW}}(t+1)=\alpha p(t+1)$会显著增加，并定义 $\alpha=10$，如图 3.25(a)所示，这确保了灵活活动的事件间隔分布以短时间为主。依据上述模型生成代理每日的行程。图 3.25 展示了一个代理人建模的示例，$p_{\mathrm{NW}}(t)$峰值对应于家之外的灵活活动。

该模型可以通过将其映射到一个硬币翻转或独立非同分布的伯努利试验问题来进行分析，前提是只有两种概率 $p_1=\langle p(t)\rangle$和 $p_2=10p_1$，而非一个依赖时间的变量。一个拥有 K 个空闲时间段的人，可以投掷硬币以改变在下一个时间段中的位置。成功(记作 H)导致留在家里或返回家中，失败(记作 T)导致探索新的地点。硬币投掷的成功概率依赖于当前状态：

$$\mathrm{H}\xrightarrow{p_1}\mathrm{T},\mathrm{H}\xrightarrow{1-p_1}\mathrm{H},\mathrm{T}\xrightarrow{p_2}\mathrm{T},\mathrm{T}\xrightarrow{1-p_2}\mathrm{H}$$

通过对独立非同分布的伯努利试验使用修改的有限马尔可夫链嵌入技术，可以将一天内访问的地点数量(等价于试验成功的数量 N)的概率写为

$$P(N)=\boldsymbol{\xi}_0\left(\prod_{t=1}^{K}\boldsymbol{\Lambda}_t\right)\boldsymbol{U}'(C_N)$$

式中：$\boldsymbol{\xi}_0$ 为相应马尔可夫链状态空间中的初始条件向量；$\boldsymbol{\Lambda}_t$ 为转移概率矩阵；$\boldsymbol{U}'(C_N)$为与 N 次成功对应的子空间的转置向量。

如图 3.25(e)所示，这个简单的硬币投掷模型可以很好地复制观察到的主题分布。

为了确认 $\alpha\gg1$ 的假设是获得每日访问地点数量分布的关键，图 3.25 中展示了三种模型的分析结果：包含两种代理(工作和非工作代理)的模型；只包含非工作代理的模型；只有一个概率 $\alpha=1$ 的模型。尽管两种代理的存在对整体主题及其分布有微小影响，但消除扰动($p_2=p_1$)会将结果从近似对数正态分布变化为二项分布。此外，还出现了调查数据中不存在的星状主题结构。因此，扰动的行为 $p_2=10p_1$ 是复制每日移动的关键因素。

本应用案例通过对个体调查和匿名手机数据获得的每日移动网络分析发现，来自两个城市的旅行调查和手机轨迹都显示出相同的一组普遍网络，在此称之为主题。观察到的结论可以支持这些主题是一般人类移动的特征，并能够进一步用于建模和模拟城市活动。此外，具有高低活动周期交替出现的扰动状态是正确复制这些主题必不可少的因

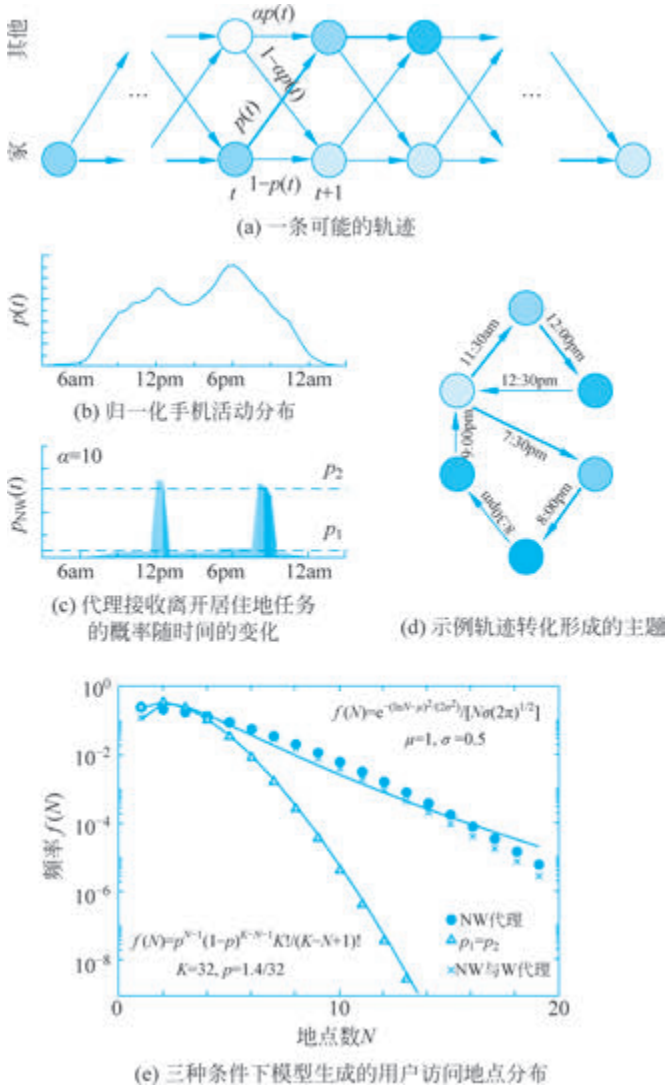


图 3.25　非工作代理模型示意图[13]

素。本案例可以为移动数据挖掘任务带来启示。通过减少基于代理建模中选择的维度，可以增强现有的城市移动模拟模型和流行病传播模型，并通过理解日常例行移动规律更好地评估城市规划和控制。

3.5　案例二：移动社交网络分析

3.5.1　问题描述

揭示人类社交网络的结构和功能一直受映射大量个体间真实交互关系的困难所制约。传统的基于问卷调查的方法通常只涵盖了几十至几百个个体，并依赖主观意见来表示交互关系的性质和强度。随着越来越多的人际交互都得到了记录（如电子邮件、电话

记录等),为揭示和探索通信和社交网络的大规模特征提供了前所未有的机会。其中,利用移动电话可以捕获数百万甚至上亿个个体的移动交互模式,从而构建一个全社会范围的通信社交交互网络。在此基础上能够探索网络拓扑与个体间联系强度的关系,这是传统方法在小范围网络上无法挖掘到的信息。本应用案例收集某国全国范围内约 20%的人口之间 126 天的所有手机通话记录并构建了一个移动通话网络,此网络能够为在整个社会尺度上理解社交网络的拓扑结构提供机遇。本应用案例,旨在观察网络度分布、联系强度等指标的分布,并在此基础上验证社交网络的组织结构[14],进一步分析社交网络中的链接关系对网络面对攻击的稳健性及信息在网络中的传播的影响。

3.5.2 结果分析

本应用案例中,根据某国全国范围内约 20%的人口之间 126 天的所有手机通话记录重建了该国通信网络的重要部分。构建方式:虽然在 126 天内两个用户之间的单次通话可能不包含太多信息,但两个用户之间的长时间通话则表明存在某种工作、家庭、休闲或服务关系的特征。因此,为了将电话日志数据转换为捕捉底层通信网络特征的网络表示,若两个用户间至少有一对电话通话是互相回拨的(A 给 B 打电话,B 给 A 打电话),则在网络中添加一条边连接这两个用户对应的节点,此边的权重 $w_{AB}=w_{BA}$ 为用户 A 和 B 之间通话的累计时长,定义为 A 与 B 的联系的强度。

由手机通话记录构建的移动通话网络包含 $N=4.6\times10^6$ 个节点和 $L=7.0\times10^6$ 条边,其中 84.1%的节点处于最大的连通分量中。移动通话网络的度分布呈偏斜分布,尾部较厚,如图 3.26(a)所示。这表明,尽管大多数用户只与少数人交流,但存在少量用户与数十人交流。如果尾部近似于幂律分布,得到的幂律指数 $k\approx8.4$,明显高于固定电话网络的幂律指数(约 2.1)。对于这种快速下降的度分布,中心枢纽较少,因此传统无标度网络的许多特性都不复存在,如异常扩散到误差容忍性。这种快速下降可能源于移动通话网络中机构电话号码的缺失,而在固定电话的情况下这些机构电话号码对应着绝大多数大型中心枢纽。

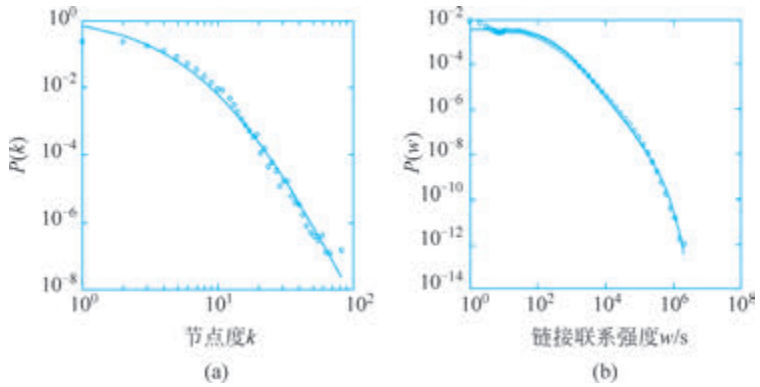


图 3.26 某国移动通话网络的节点度分布与链接联系强度分布示意图[14]

如图 3.26(b)所示,移动通话网络中联系强度 w_{AB} 的分布是宽泛的,1～10^6 s,衰减

指数 $\gamma_w \approx 1.9$。因此，尽管大多数联系对应着几分钟的通话时间，但有一小部分用户数小时与对方聊天。这一发现出人意料，因为重尾联系强度分布主要出现在全球运输过程的网络中，例如航空运输网络所承载的乘客数量、代谢网络中的反应通量或互联网上的数据包传输，在这些情况下个体通量由全局网络拓扑确定。这类全局流动过程的一个重要特征是局部守恒：到达机场的所有乘客都需要被送走，由反应产生的每个分子都需要被其他反应消耗，或者到达路由器的每个数据包都需要发送到其他路由器。尽管手机的主要目的是在两个个体之间传递信息，但缺少限制或驱动联系强度的局部守恒，使得移动通话网络的拓扑与局部联系强度之间的任何关系都不那么明显。

复杂网络通常根据全局效率原则自组织，这意味着联系强度被优化以最大化网络中的总体流动。在这种情况下，链接的权重应与其介数中心性相关，介数中心性与通过它的所有节点对之间的最短路径数量成正比，与集聚系数类似，用于衡量网络节点在网络拓扑结构中的重要性和连接性。另一种可能性是特定链接的强度仅取决于两个个体之间关系的性质，因此与围绕联系的网络无关。这种假设称为二元假设。广泛研究的弱链接强度假设认为，A 和 B 之间的链接强度随着他们友谊圈的重叠而增加，从而强调了弱链接在连接社区中的重要性。这一假设导致了弱链接的介数中心性较高，可以看作全局效率原则的镜像。

图 3.27(a)展示了在随机选择的个体附近的网络结构，其中链接的颜色对应于每个联系的强度。由图可以看出，网络由小的局部集群组成，通常围绕在一个高度连通的个体周围。与弱联系强度假设相一致，大多数强联系都在集群内部，表明用户将大部分通话时间用于与他们直接朋友的交流。相比之下，连接不同社区的大多数链接较弱，而不同社区内部的链接较强。作为比较，当随机排列连接用户对之间的链接强度时(对应二元假设)，由图 3.27(b)可以观察到在社区内部有更多的弱联系，而连接不同社区有更多的强联系。若根据全局效率原则和介数中心性的预测，即社区间的联系(“桥梁”)较强，社区内部的联系(“本地道路”)较弱时，将会呈现图 3.27(c)所示的结构，相比二元假设情况这与图 3.27(a)观察到的数据更不一致。

彩图

(a) 某个体附近的局部网络结构及链接强度　(b) 二元假设下随机排列得到的结构及链接强度　(c) 全局效率原则下随机排列得到的结构及链接强度

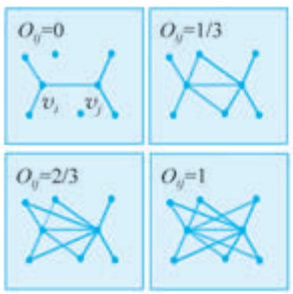
图 3.27　网络结构及链接强度[14]

为了量化图 3.27 中观察到的差异，定义网络中用户 v_i 和 v_j 附近的拓扑重叠的相对值：

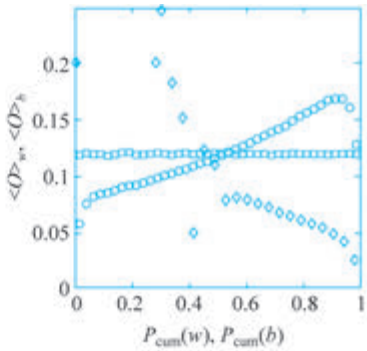
$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}}$$

式中：n_{ij} 为 v_i 和 v_j 共同邻居的数量；$k_i(k_j)$表示节点 $v_i(v_j)$的度；O_{ij} 为 v_i 和 v_j 共同朋友的比例。

与社群发现算法中的判断依据类似。如图 3.28 所示，i 和 j 之间的链接代表着连接两个不同社区的潜在“桥梁”，若 v_i 和 v_j 没有共同朋友，则 $O_{ij}=0$；若 i 和 j 属于同一个朋友社群，则 $O_{ij}=1$。二元假设暗示了局部网络拓扑与权重之间没有关系。事实上，当随机排列网络中的链接强度时，O_{ij} 与 w_{ij} 无关，如图 3.28 所示。然而，根据全局效率原则，在给定的 b_{ij} 的介数中心性下，$\langle O\rangle_b$ 随着 w_{ij} 的增加而减小。这表明在平均情况下具有最高介数中心性 b_{ij} 的链接具有最小的重叠。相比之下，图 3.28 中对于真实通信网络，$\langle O\rangle_w$ 随着权重小于 w 的链接百分比的增加而增加，表明用户之间的联系越强，他们的朋友重叠越多，这一相关性对 95%的链接都有效。这一结果与弱联系强度假设基本一致，在社会层面证实了弱联系强度假设。即网络节点间的联系强度在一定程度上受到了与联系的直接周围环境相关的网络结构的驱动。这一建议与纯粹的二元观点相矛盾，后者认为联系强度仅由它连接的个体的特性或全局观点确定。



(a) 网络拓扑重叠相对值O_{ij}的定义



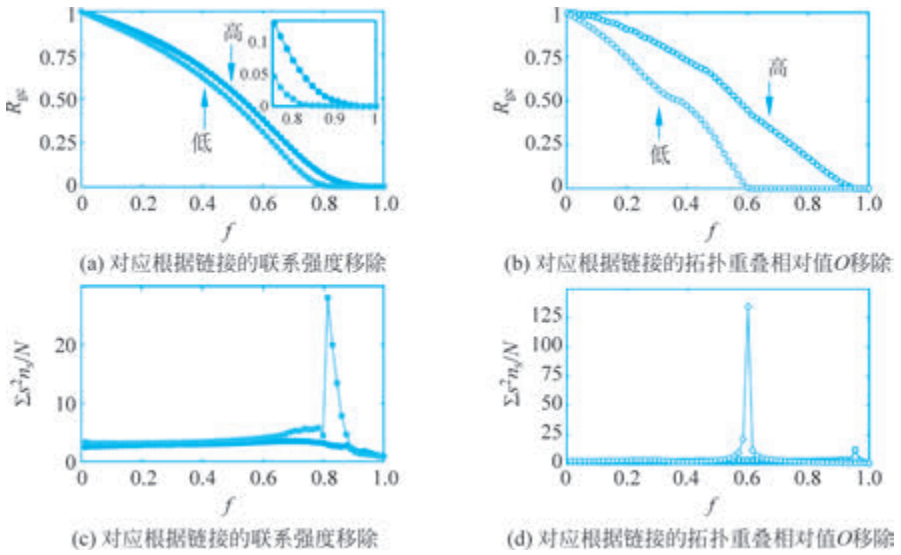
(b) 不同联系强度下O的均值(蓝色)、二元假设随机重连时不同联系强度下O的均值(红)及不同介数中心性下O的均值(黑色)

图 3.28 O_{ij} 的定义及 O 的均值[14]

为了理解联系强度与网络结构之间的这种局部关系对系统或全局的影响，进一步探讨网络抵抗删除强弱链接的能力。为了评估删除网络中链接的影响，可以测量巨大分量 $R_{gc}(f)$的相对大小，即在网络中删除链接的比例为 f 时，通过剩余的联通路径可以相互到达的节点的比例。如图 3.29 所示，按照从最弱的(或最小的 O_{ij})到最强的(或最大的 O_{ij})的顺序删除链接会导致网络在 $f^w=0.8$ 和 $f^O=0.6$ 时突然解体。相比之下，首先删除最强的链接会使网络缩小，但不会迅速解体。网络解体的确切点可以通过监测 $\widetilde{S}=\sum_{s<s_{max}} n_s s^2/N$ 来确定，其中 n_s 是包含 s 个节点的集群的数量。根据渗流理论，如果网络因为在 f_c 处的相变而解体，那么当 f 接近 f_c 时，$\widetilde{S}$ 的值会发散。如图 3.29 所示，如果从最弱的链接开始移除，$\widetilde{S}$ 会发展出一个峰值；如果从最强的链接开始移除，$\widetilde{S}$ 并不会出现峰值，表明网络不会发生相变。

总的来说，这些结果表明社交网络中强弱关系的全局作用存在根本差异：弱链接的

彩图



彩图

图 3.29　移动通话网络在不同链接移除比例 f 下的巨大分量 R_{gc} 与 $\widetilde{S}$

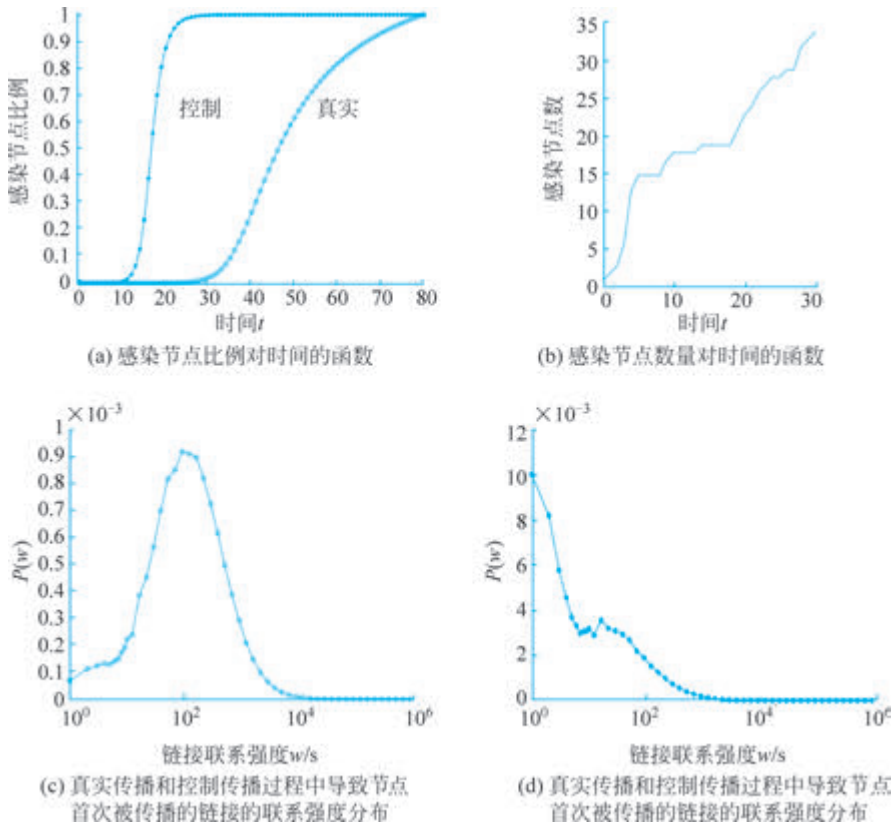
注：红色表示从低到高移除链接，黑色表示从高到低移除链接。

移除会导致整个网络发生突然的、受相变驱动的解体或崩溃；强链接的移除只会导致网络逐渐缩小而不会崩溃。这一发现出人意料，因为在大多数技术和生物网络中，人们认为强链接在结构上起到比弱链接更重要的作用，在这类系统中，移除强链接会导致网络崩溃。这一反直觉的发现揭示了弱链接和强链接在社交网络中发挥的不同作用：鉴于强链接主要存在于社区内部，它们的移除只会在局部破坏一个社区，但不会影响网络的整体完整性；移除弱链接将删除连接不同社区的桥梁，导致受相变驱动的网络崩溃。

手机在两个个体之间传递信息，鉴于这些个体嵌入在一个社交网络中，新闻和谣言能够传播到双方之外，有时会影响到大量个体，这是社会学和网络科学中广泛研究的传播问题。目前关于信息传播的大部分知识都是基于分析无权网络得出的，其中所有链接强度是相等的。为了查看观察到的网络拓扑和链接的联系强度之间的局部关系是否影响全局信息传播，可以进行如下模拟实验：在 0 时刻，将一位随机选择的个体感染上某些新信息，假设在每个时间步骤中，每个被感染的个体 v_i 都可以以有效概率 $P_{ij}=xw_{ij}$ 传递信息给其邻居 v_j，其中参数 x 控制整体传播速度。因此，两个个体在手机上花费的时间越多，传递此信息的机会就越高。这种传播机制类似于流行病学的易感-感染模型，其中不能康复的被感染个体将无限期地继续传播信息。作为对照，本应用案例中考虑了在相同网络上的传播，但用其平均值替换了所有连接强度，导致所有链接的传输概率恒定。

传播模拟过程中感染节点的比例与时间的关系如图 3.30 所示，蓝色曲线对应于在具有真实关系强度的网络上的真实传播过程，而黑色曲线表示控制模拟，在其中所有关系强度都视为相等。在所有权重相等的网络上信息传递速度更快，这种差异根植于信息在社区中的动态陷阱中。这种陷阱在监测扩散过程早期感染个体数量时清晰可见。图 3.30 展示了感染节点数量作为传播时间的函数。曲线的每个陡峭部分对应于入侵小社区，而较平的部分表示传播被困在社区内。可以观察到信息一开始在一个社区内迅速

传播，对应于感染用户数量的快速增加；然后出现平台期，对应于没有新节点被感染的时间间隔，直到信息传播到社区外。当用连接强度的平均值 w 替换了所有连接强度时，称为控制传播过程，如图 3.30 黑色点所示，社区之间的桥梁得到了加强，传播成为主要的全局过程，通过一系列枢纽迅速到达所有节点。



(a) 感染节点比例对时间的函数　(b) 感染节点数量对时间的函数

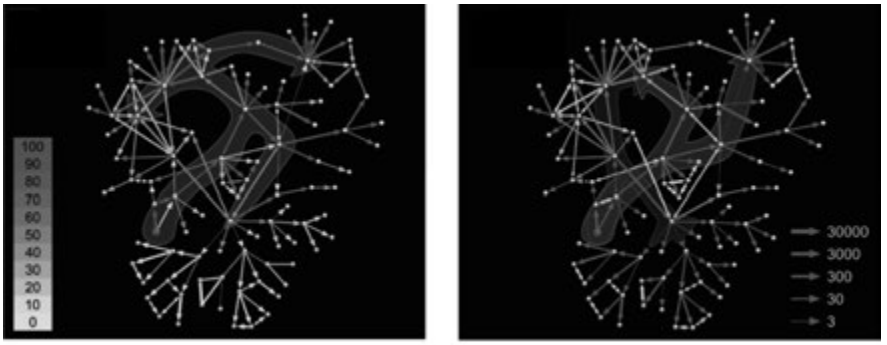
(c) 真实传播和控制传播过程中导致节点首次被传播的链接的联系强度分布　(d) 真实传播和控制传播过程中导致节点首次被传播的链接的联系强度分布

彩图

图 3.30　传播模拟过程示意图[14]

真实传播过程和控制传播过程之间的巨大差异引发了一个重要问题：个体从哪里获取信息？如图 3.30 所示，每个个体首次感染的联系强度分布在 $w \approx 10^2$ 处具有突出的峰值，这表明在绝大多数情况下个体是通过中等强度的联系了解新闻的。然而，在控制情况下，即在传播过程中将所有联系强度视为相等时，分布发生了显著变化。如图 3.30 所示，在这种情况下，大多数感染发生在较弱的联系上。因此，与弱联系在信息获取中的突出作用相反，弱联系和强联系都在信息传输中起到了相对不重要的作用，前者通话时间较短，信息传输机会较少，后者主要局限于社区内部，几乎无法获得新信息的访问。

为了说明真实模拟和控制模拟之间的差异，图 3.31 展示了信息在一个小邻域中的传播。首先，两种情况下的信息流动方向总体上是不同的，如大的阴影箭头所示。在控制运行中信息主要沿最短路径传播。然而，在考虑权重时(图 3.31(a))，信息沿着强链接骨干传播，与其余网络通过弱链接连接的大区域很少被感染。例如，在真实模拟中网络的下半部分很少被感染，但在控制运行中总是被感染。因此，当忽略导致图 3.30(a)两组



彩图

(a) 真实传播模拟　　(b) 控制传播模拟

图 3.31　信息在一个小邻域中的传播[14]

注：在真实模拟和控制模拟情况下，从红色节点释放信息的传输示意图。联系强度对应箭头的粗细。绿色等高线说明两个模拟中信息传递方向流的差异。

曲线差异的联系强度时，网络中的扩散机制发生了显著变化。

尽管社交网络的研究历史悠久，但研究社会范围网络中强度与拓扑之间的关系通常是不可能的。本应用案例从移动电话通话记录构建的移动通话网络中观察到网络中的关系强度与关系周围的局部网络结构相关，并且二元假设和全局效率原则都不能解释实际观察到的现象。长期以来人们已经知道许多网络在随机节点删除方面表现出弹性，但在删除中心节点的情况下是脆弱的。本应用案例的分析显示了与通信网络相反的效应：弱关系的删除导致了类似相变的网络崩溃，尽管强关系的移除对网络的整体完整性几乎没有影响。此外，观察到的网络结构与关系强度之间的耦合显著减缓了信息流动，将其困在社群中，这一现象也解释了为何在社交网络中成功的搜索主要通过中等到弱强度的关系而避免使用中心节点。因此，为了增强信息的传播，需要有意地通过弱关系传递信息，或者采用主动的信息搜索过程。

3.6　本章小结

本章介绍了用网络模型分析移动数据的方法与具体应用案例。移动数据特有的时空关系使得其可以天然地形成网络结构，用网络模型来抽象移动数据中的移动行为可以为分析移动数据带来诸多便利与应用。在介绍真实网络的重要性质及三角形和社群发现两个经典算法基础上，本章用两个案例分别阐释了网络拓扑结构和网络社群发现等思想的重要性。有兴趣的读者可以进一步探究网络模型相关的前沿研究。在使用网络模型分析移动数据时，可以使用 Python 的开源库 networkx 来实现网络模型的构建及多种知名算法[15]。

参考文献

参考文献