数据科学导论第 13 讲—— 社交网络分析

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 06 月 19 日

目录

社交网络概述

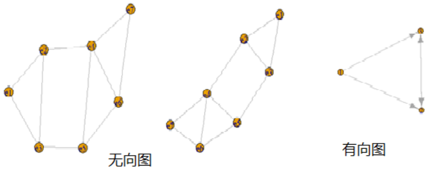
社交网络

- 社交网络即社交网络服务，源自英文 SNS（Social Network Service）的翻译，中文直译为社交网络服务，意译为社交网络服务。
- 社交网络分析 (Social Network Analysis) 是指基于信息学、数学、社会学、管理学、心理学等多学科的融合理论和方法，为理解人类各种社交关系的形成、行为特点分析以及信息传播的规律提供的一种可计算的分析方法。
- 社交网络分析最早是由英国著名人类学家 Radcliffe-Brown(拉德克利夫-布朗) 在对社会结构的分析关注中提出的，他呼吁开展社会网络的系统研究分析。
- 随着社会学家、人类学家、物理学家、数学家，特别是图论、统计学家对社会网络分析的日益深入，社交网络分析中形成的理论、方法和技术已经成为一种重要的社会结构研究范式。
- 由于在线社交网络具有的规模庞大、动态性、匿名性、内容与数据丰富等特性，近年来以社交网站、博客、微博等为研究对象的新兴在线社交网络分析研究得到了蓬勃发展，在社会结构研究中具有举足轻重的地位。

网络的基本概念

网络的基本概念

- 网络（图，Graph): 有序三元组 (V, E, φ), 其中 V 为顶点集，非空，E 为边集，φ 是有序或无序对簇 $V \times V$ 的函数，也称关联函数
- V 中的元素叫顶点（Vertex),E 中元素叫边 (Edge),φ 描述了 V 中元素的关系
- 若 $V \times V$ 中的元素都为无序对，(V, E, φ) 称为无向图（Undirected Graph），记作 $G = (V(G), E(G), \varphi_G)$
- 假定 $e \in E(G)$，则存在 $x, y \in V(D)$ 及无序对 $(x, y) \in V \times V$, 使得 $\varphi_D(e) = (x, y)$
- 对于无序对来说，$\{x, y\}$ 和 $\{y, x\}$ 代表同一元素，因此 $\varphi_D(e) = xy$ 或 yx，e 称作连接 x 和 y 的边。
- 对于有序对来说，(V, E, φ) 为有向图，记 $D = (V(D), E(D), \varphi_D)$
- 假定 $a \in E(D)$, 存在 $x, y \in V(D)$ 和有序对 $(x, y) \in V \times V$ 使得 $\varphi_D(a) = (x, y)$。
- a 是从 x 到 y 的有向边，x 称为 a 的起点，y 称为 a 的终点，起点和终点统称为端点。



* 可以用平面上的一个点来表示网络中的某一个节点，用点与点之间的线段代表节点之间的边，边可以使用直线来表示，也可以使用曲线来表示，这样的表示方法叫做网络的图示。

- 网络的邻接矩阵表示如下：

$$A(G) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1v} \\ a_{21} & \cdots & \cdots & a_{2v} \\ \cdots & \cdots & \cdots & \cdots \\ a_{v1} & a_{v2} & \cdots & a_{vv} \end{bmatrix}$$

- 网络的关联矩阵表示如下：

$$M(G) = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1\tau} \\ m_{21} & \dots & \dots & m_{2\tau} \\ \dots & \dots & \dots & \dots \\ v_{v1} & m_{v2} & \dots & m_{v\tau} \end{bmatrix}$$

网络特征的描述性分析

节点度

- 网络 G 中节点 v 的度，简称点度。
- $d_G(v)$ 是该网络里同 v 关联的边的数量。
- 有向图：根据方向分为出度和入度。
- 以节点 v 作为终点的边的次数之和是 v 的入度，以节点 v 作为起点的边的次数之和是 v 的出度。

节点中心性

- 关于网络中节点的很多问题，本质上是在试图理解它在网络中的“重要性”。
- 对于生物体，移除相应基因调控网络中的哪个基因可能是致命的？这个网络中哪个行动者最有权势？
- 互联网上的某个路由器对于信息流动有多重要？万维网上某个网络的权威性应当如何评判？
- 度量“中心性”（centrality）可以量化“重要性”，从而协助解决这些问题。

- 接近中心性
- 主要思想：如果一个节点与许多其他节点都很“接近”（close），那么节点处于网络中心位置（central）。根据这一想法，我们可以用某节点到其他所有节点距离之和的倒数来表示接近中心性:

$$c_{CI}(v) = \frac{1}{\sum_{u \in V} Dist(v,u)}$$

- 其中 $Dist(u,v)$ 是节点 $u,v \in V$ 的捷径距离。
- 通常这一度量会乘以系数 $N_v - 1$ 归一化到 $[0,1]$ 区间，用于不同网络之间以及不同性度量之间的比较。

介数中心性

- 度量描述的是该节点在多大程度上“介于”（between）其他节点之间。
- 中心性基于这样一种观点：**节点的“重要性”与其在网络路径中的位置有关。**

* 如果我们将这些路径视作进行通信所需的渠道，那么处于多条路径上的节点就是通信过程中的关键环节。最常用的介数中心性的定义为：

$$c_B(v) = \frac{\sigma(s,t|v)}{\sum_{u\in V}\sigma(s,t)}$$

- 其中 $\sigma(s,t|v)$ 是 s 与 t 之间通过 v 的最短路径数，$\sigma(s,t)$ 是 s 与 t 之间 (无论是否通过 v) 的最短路径数。
- 当最短路径唯一时，$c_V(B)$ 仅计算通过 v 的最短路径数量。
- 这一中心性度量可以通过除以系数 $\frac{(N_v-1)(N_v-2)}{2}$ 归一化到单位区间。

节点中心性

- 如果一个节点的邻居中心性越高，节点本身的中心性也越高。这类度量本质上是隐式定义的，通常可以表达为某种恰当定义的线性系统方程的特征向量的形式。
- “特征向量中心性”（eigenvector centrality）的度量方法有很多，常用的为：

$$c_{E_i}(v) = \sum_{\{u,v\}\in E} c_{E_i}(u)$$

- 向量 $C_{E_i} = (C_{E_i}(1), \cdots, C_{E_i}(N_v))$ 是特征值问题，$Ac_{E_i} = \alpha^{-1} c_{E_i}$ 的解。其中，A 是 G 的邻接矩阵。

网络凝聚性

- 根据问题所属的领域，可以使用很多的方法定义网络的凝聚性。
- 定义有不同的尺度，既有局部的也有整体的；决定的明确程度也不同，有的很清晰（如团），有的相对比较模糊（如聚类或社团）。
- 定义网络凝聚性的一种方法是规定某种感兴趣的子图类型。
- 团是这类子图的典型例子，是一类完全子图，集合内的所有节点都由边相互连接，因而是完全凝聚的节点子集。
- 所有尺寸的团的普查（census）可以提供一个“快照”，让我们了解网络的结构是怎样的。
- 大尺寸的团包含了小尺寸的团。“极大团”（maximal clique）是不被任何更大的团包含的一类团。
- 由于现实生活里的网络大部分都是稀疏的，而团的存在要求网络本身相当稠密，所以实际上，大尺寸的团比较稀少。团的定义存在各种弱化了的版本。

网络凝聚性

- 网络 G 的 k 核 (k−core) 是一个网络 G 的子图，里面包含的所有节点的度最少是 k, 而且它是满足条件的最大的子图，即不被包含于满足条件的其他子图中。
- 核的概念在可视化中非常流行，因为它提供了一种将网络分解到类似洋葱的层的方法。
- 这种分解可以与辐射布局有效地结合起来（使用靶心图）。

- 其他如二元组和三元组。
- 二元组关注两个节点，他们在有向图中有三种可能的状态：**空（Null），非对称（Asymmetric, 存在一条有向边），双向（Mutual, 两条有向边）**
- 三元组是三个节点，对图中每个状态观察到的次数进行统计，得到的这两类子图可能状态的一个普查，它可以帮助理解途中连接的本质。

密度

- 密度是指实际出现的边与可能的边的频率之比。如，对于不存在多重边，而且没有自环的（无向）图 G, 子图 $H=(V_H,E_H)$ 的密度为：

$$den(H)=\frac{|E_H|}{|V_H|(|V_H|-1)/2}$$

- $den(H)$ 的值处于 0 到 1 之间，提供了一种 H 与团的接近程度的度量。G 为有向图时，上式中的分母将替换为 $|V_H|(|V_H|-1)$。
- 由于定义式子时子图 H 可以自由选择，这使简单的密度概念变得很有趣。
- 如令 $H=G$, 得到的是整个网络 G 的密度。而令 $H=H_v$ 为节点 $v\in V$ 的邻居集合以及节点间的边，度量的是 v 直接相邻邻居的密度。

聚类系数

- 相对频率也可以用于定义网络中的“聚集性”（clustering）概念。例如，术语“聚类系数”（clustering coefficient）的标准定义如下：

$$cl_{\tau}(G) = \frac{3\tau(G)}{\tau_3(G)}$$

- $\tau V(G)$ 指的是网络 G 的三角形个数，$\tau_3(G)$ 是联通三元组个数。其中，联通三元组指的是由两条边连接的三个节点，也称为 2-star。
- $cl_{\tau}(G)$ 的值也被称为网络的“传递性”，它是社会网络文献中的一个标准指标，表示传递性三元组的比例。
- $cl_{\tau}(G)$ 是对全局聚集性的度量，所概括的是联通三元组闭合形成三角形的相对频率。

分割

- 分割泛指将元素的集合划分到“自然的”子集之中的过程。即一个有限集 S 的分割 $l = \{C_1, \cdots, C_k\}$ 是将 S 分解成 K 个不相交的非空子集 C_k, 满足 $\bigcup_{k=1}^{K} C_k = S$。
- 在网络分析中，分割是一种无监督的方法，用于发现具有“凝聚性”的节点子集，揭示潜在的关系模式。
- 图分割 (Graph Partitioning) 问题在复杂网络方面的文献中也常被称为社团发现（Community Detection）问题。
- 描述这一问题可通过系统聚类来实现。

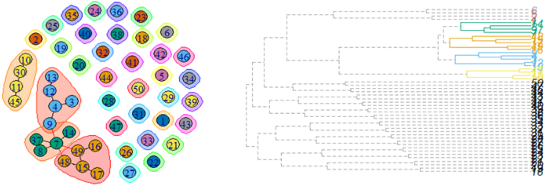
1. 聚类算法通过合并过程逐渐得到更粗粒度的分割
2. 分裂算法通过分裂过程逐步对分割进行优化。

- 在每一步中，当前候选的分割以最小化某种成本度量的方式进行修正。
- 凝聚算法选择最小化成本的方式，将两个之前存在的分割元素进行合并。
- 分裂算法选择最小化成本的方式，将一个分割的元素划分为两个。

分割

- 当进行网络图分割时，无论是采用凝聚还是分裂的方法，系统聚类实际上产生的是一个嵌套的图分割层级而非单个的分割。
- 这些分割包括了从最细的分割 $\{\{v_1\}, \cdots, \{v_{N_v}\}\}$ 到整个节点集 V 的情况。
- 凝聚方法从前面开始合并，而分裂方法从后面开始分解。层级结果通常使用树的形式进行表示，成为树状图（Dendrogram）。

分割-树状图



豆瓣朋友网络分割图

- 上左图为使用凝聚算法得到的分割，该网络可以分成四个社团和多个个体。
- 上右图是该分割对应的树状图，通过树状图可以得到整个凝聚过程，观察到每个节点是在什么阶段凝聚到一起的。

网络图的统计模型

经典随机图模型（Random Graph Model,RGM）

- 通常指一个给定了集合 l 及以上的均匀概率分布的模型。
- RGM 是发展最完备的一类网络图模型，其基础是一个认为所有给定阶数和规模的图具有相同概率的简单模型。
- RGM 给定了一个集合 ℓ_{N_v,N_e} 表示所有 $|V| = N_v |E| = N_e$ 的图 $G = (V, E)$, 并规定每个 $G \in \ell_{N_v,N_e}$ 的概率为 $P(G) = 1/C_N^{N_e}$，其中 $N = C_n^2$ 表示不同节点对的总数。