基于复杂网络的机器学习方法

右武卫大将军

2019 年 8 月 12 日

1 复杂网络

1.1 摘要

复杂网络是一个新兴的跨学科研究领域，他引起了物理学，数学，生物学，工程，计算机等学科的专家以及其他许多人的广泛关注。复杂网络可以描述诸如因特网，万维网，生物化学反应，金融网络，社交网络，神经网络和通信网络等各种各样高度技术化和智能化的系统。为了了解这些相互交织系统的复杂内在结构，人们越来越关注复杂网络这个工具。为了更好地使用复杂网络，我们对复杂网络中数据结构的复杂性和节点及其相关连接的多样性进行了同一。当利用复杂网络进行系统动态分析时，会有许多问题需要研究，比如设计多大的网络，选择什么样的拓扑结构可以表示一个相互作用的复杂动力系统。网络的拓扑结构选择至关重要，它决定了系统的功能。例如，社交网络的结构可能影响信息和灾害传播速度，金融网络的拓扑结构可能会对冲击造成不同程度的放大，电力网络的不同配置可能影响电力传输的鲁棒性和稳定性。由于相关理论和技术的快速发展，对复杂网络的全面综述已经十分困难。这一章中，我们重点介绍复杂网络的基本概念和基本思路，这也是复杂网络在机器学习中应用的基础。2.1 节介绍网络的主要概念。由于复杂网络和图论有相同的定义，我们将介绍图论的基本符号。2.2 节对复杂网络研究的进展和主要研究成果进行总结。2.3 节对捕捉网络系统结构特征的网络测量方法进行了全面探讨。2.4 节总结了在复杂网络框架下定义的一些著名的动态过程。

1.2 图论简介

1.2.1 图的定义

本节介绍图论或者网络理论的主要技术术语。本书中，图和网络表达相同类型的信息，二者可以互换。构成图的数据相互之间的关系也可以成为网络结构和拓扑结构。

下面我们给出图的一般定义。

定义 1.1 图： 有序二元组 (V,E) 称为图，记为 $G,G=(V,E)$，其中非空有限集合 V 是节点集合，E 是 V 上的一个二元关系，即节点和节点之间的关系，称为边集合，即 $E\subseteq\{(u,v)|u,v\in V\}$。

两种特殊图的定义如下：

- **无环边图**：对于图 $G=(V,E)$，集合 E 不是自反关系，即 $\forall v \in V,(v,v) \notin E$，称图 G 为无环边图。也就是说，无环边图中的节点经过一个转换过程之后不可能回到原来的节点。
- **有环边图**：对于图 $G=(V,E)$，集合 E 满足条件：$\exists v \in V,(v,v) \in E$，则称图 G 为有环边图。也就是说，有环边图中存在节点，经过一个转换过程，即经过某条边后，可以回到原来的节点。

此外，$\mathbb{V}=|V|$ 称为节点的个数，$\mathbb{E}=|E|$ 称为节点的条数。下面介绍一些著名的图形拓扑结构。

定义 1.2 完全图：对于图 $G=(V,E)$，任意两个节点相邻，则称图 G 为完全图。包含 V 个节点的完全图记为 H_V

完全图也可以根据有无环边来进行进一步的分类。

定义 1.3 零图：对于图 $G=(V,E)$，G 只有节点没有边的图，即 $E=\emptyset$，那么我们称 G 为零图。

但需要注意的是，尽管零图中边的集合可以为空，但节点的集合不能为空。

图也可以根据边的类型进行分类。下面介绍根据图中边的类型分类的图形拓扑。

定义 1.4 无向图：对于图 $G=(V,E)$，若节点之间的相互关系没有方向，即 $\forall (u,v) \in E$，则一定有 $(v,u) \in E$，则称 G 为无向图。换句话说，当存在一条边将节点 u 连接到节点 v，那么从节点 v 到节点 u 的通路也存在。

一般情况下，如果图中的边没有箭头，那么该图就是无向图。如果图中的边是有箭头的，那么该图就是有向图，定义如下：

定义 1.5 有向图：对于图 $G=(V,E)$，当边集合 E 满足：$\exists (u,v) \in E | (v,u) \notin E$，那么称图 G 为有向图。也就是说，有向图中至少有一条边是具有方向的。

定义 1.6 带权图 如果 G 是一个三元组 (V,E,W)，其中新增的 W 是一个 $V \times V$ 的矩阵，表示每一条边的权重，那么称 G 为带权图。如果 $(u,v) \in E$，且这条边的权重为 $w>0$，那么 $W_{uv}=w$，如果 $(u,v) \notin E$，则 $W_{uv}=0$。

定义 1.7 二分图 当图中的节点集合可以分为两个不相交的非空集合 V_1 和 V_2 时，对于 $\forall (u,v) \in E, u \in V_1, v \in V_2$，我们称图 G 为二分图。此时，V_1 和 V_2 集合内的节点不相邻，即没有同一个集合内的边。当 $|V_1|=M, |V_2|=N$，如果对于 $\forall u \in V_1, \forall v \in V_2$，都满足 $(u,v) \in G$，则该图是一个完全二分图 $H_{M,N}$。

显然，如果图 G 是二分图，则，图 G 一定没有环边。

1.2.2 图的连通性

在这一节，我们将介绍与图连通相关的常用术语，它们贯穿于本书中。

定义 1.8 节点相邻 在图中，如果 $e=(u,v) \in E$，则称节点 u 和节点 v 是相邻的。

不过在有向图和无向图有所区别。在无向图中，如果 u 相邻于 v，那么 v 就相邻于 u。而在有向图中，节点 u 相邻于节点 v，节点 v 未必相邻于节点 u。

定义 1.9 节点的邻域 在图中，节点 $v \in V$，图中与 v 相邻的点的集合称为节点 v 的邻域，记为 $N(v) = \{u : (v, u) \in E\}$

邻域又可以区分为开邻域和闭邻域。区别在于节点 v 是否包含在邻域 $N(v)$ 中，如果包含的话，就是闭邻域，如果不包含的话，就是开邻域。在机器学习的场景下，大部分情形下说的都是开邻域，除非节点 v 上有环边。

定义 1.10 节点的度 在无向图中，与节点 v 关联的边的数目称为节点的度，记为 k_v。任意节点 v 的度与节点 v 邻域集合中元素的个数相等，记为 $k_v = |N(v)|$。即：

$$k_v = |N(v)| = |\{u : (v, u) \in E\}| = \sum_{u \in V} \mathbb{I}[(v, u) \in E] \tag{1}$$

其中，$\mathbb{I}[K]$ 为克罗内克函数或者指示函数，当逻辑表达式 K 为真时，返回值为 *1*，否则，其返回值为 *0*。

当节点 v 的度为 0 的时候，节点 v 称为孤立节点。当节点 v 的度比其他节点的度大时，则称该节点为关键节点。

定义 1.11 入度和出度 在有向图中，以节点 v 为终点的边的数量称为节点 v 的入度，记为 $k_v^{(in)}$；以节点 v 为起点的边的数量称为节点 v 的出度，记为 $k_v^{(out)}$。该节点的度是出度和入度的和。

定义 1.12 平均度 所有节点的度的平均值称为图的平均度。**平均出度** 所有节点的平均出度为图的平均出度。**平均入度** 所有节点的平均入度为图的平均入度。

定义 1.13 强度 在无向图的节点 $v \in V$，我们把节点 v 邻域内全部权重值的和称为节点 v 的强度，记为 s_v。即

$$s_v = \sum_{u \in V} W_{vu} \tag{2}$$

同样在有向图中，定义有**入强度**和**出强度**。

在介绍了有关图的连通性基本概念的基础上，我们介绍一个著名的拓扑图。

定义 1.14 正则图 在图中，所有节点都具有相同数量的邻点，即每个节点具有相同的度 k，则称图 G 为正则图，图 G 也可称为 $k-$ 正则图。

当然，根据这个定义，如果图是完全图，那么该图是 $(V-1)-$ 正则图。

1.2.3 路径和环路

定义 1.15 路径/游走 令 $v_1, \cdots, v_K \in V, K \leq 2$，若 $\forall k \in (2, \cdots, K) : (v_{k-1}, v_k) \in E$，则我们把从 v_1 到 v_K 的有序边序列 $\{(v_1, v_2), (v_2, v_3), \cdots, (v_{K-1}, v_K)\}$ 称为路径，记做 W。v_1 为路径的起点，v_K 为路径的终点。注意，节点可以在路径中重复出现。

如果起点和终点重合时，则称路径为闭合路径，否则是开发路径；仅包含一个节点的路径称为环路径。

定义 1.16 行迹 如果路径中没有边是重复的，那么该路径就是行迹。当行迹的起点和终点重合时，称为闭合行迹（**回路**），否则称为开放行迹。

定义 1.17 路径长度 路径中涉及到的边的数量。

定义 1.18 轨迹 当路径不是环路径，且除了起始点和结束点之外，其他节点均不重复时，这样的路径称为轨迹。如果起点和终点重合的轨迹称为环路。

定义 1.19 路径/轨迹距离 节点 u 和节点 v 的路径/轨迹距离是该路径上所有边的权重和，记为 $d(W)$。

定义 1.20 节点的最短路径距离 图中节点 u 和节点 v 的最短路径距离记为 $d_{uv} = \min\limits_{W_{u\to v}} d(W_{u\to v})$。节点间的最短路径距离也称为节点的距离。

当节点之间没有路径相连的时候，它们之间的距离是无穷大。节点到自身的距离是 0。

1.2.4 子图

定义 1.21 可达性 在图中两个节点之间的距离为有限实数的时候，称为节点 u 可达节点 v，即两点之间存在路径。

定义 1.22 可连通性 即图中任何两个节点之间都是可达的，那么该图就是连通图。在有向图中，随机选两个节点 u, v，只要满足 u 可达 v，或者 v 可达 u，都称为连通图。

定义 1.23 强可连通性 在有向图中，随机选两个节点 u, v，同时有 u 可达 v，v 可达 u，那么称该图为强可连通图。

基于上面几个定义可知，强可连通图一定是可连通图；对于无向图中，强可连通图和可连通图是等价的。

定义 1.24 连通子图 当图 G 的子图 G_C 满足条件：G_C 为连通图且 G_C 的任意子图均为非连通图，称 G_C 是图 G 的一个连通子图。

当然连通图只有一个连通子图，即其自身。

定义 1.25 派系 在无向图中，对于节点的某一个子集，满足子集中任意两个节点均有一条边相连，那么称该子集为图的派系。因此派系一定是子图，并且为完全图。

1.3 树和森林

定义 1.26 树 不含环路的连通图称为树。在树中，度为 *1* 的节点称为树叶。其他节点的度至少为 *2*。

定义 1.27 森林 当无向图的所有连通子图均为树时，这种图称为森林。

从该定义来看，森林是由不相交的树组成的图。所有的树都是森林，但反之不成立。

定义 1.28 生成树 图 G 为连通图，当包含图中所有节点的子图为树时，我们称该子图为图 G 的生成树。

1.3.1 图的矩阵表示

从数学上看，无权重图或者带权重图经常用节点和边集构建的邻接矩阵来表示。邻接矩阵的定义如下：

定义 1.29 邻接矩阵 带权重图 $G=(V,E,W)$，图 G 的邻接矩阵 A 定义为：

- 矩阵 A 是 V 阶方阵，维数由节点数来确定。
- 矩阵 A 中各元素的初始值为边的权重值，$A_{ij}=W_{ij}$。

邻接矩阵的一般形式为：

$$A=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,V} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,V} \\ \vdots & \vdots & & \vdots \\ a_{V,1} & a_{V,2} & \cdots & a_{V,V} \end{bmatrix} \tag{3}$$

如果该图是无向图，那么它对应的邻接矩阵就是对称矩阵；但如果是有向图，那么就可能不是对称矩阵。

1.4 网络演化模型

为了研究真实网络的拓扑结构，学者提出了许多网络模型。其中有一些网络模型由于其广泛的应用价值而得到了学者十分深入的研究。这些重要的网络模型包括：随机网络，小世界网络，随机聚类网络，无标度网络和核心 -边缘网络。下面分别对这些复杂网络模型进行介绍。

1.4.1 随机网络

1959 年，有两位学者首次提出了一种由 V 个节点和 E 条边组成的随机网络模型。一种描述方式是，给定网络中节点的个数为 V，逐步增加 L 条边，其中不存在自循环边。另一种描述方式是，给定网络中节点的个数 V，网络中任意两个点以概率 $p(p>0)$ 相连接。后者的模式是学者们公认的随机网络模型。在随机网络中，忽略节点之间的空间关系。在这种网络形成的过程中，边的创建以均匀概率的方式进行，不考虑节点之间的相似性。

因此，在随机网络中的任意节点 $i\in V$，它都有 $V-1$ 种方案与其他节点相连接，其连接方案的数量和样本空间的基数大小相关，样本空间的大小决定了网络中边的最大理论值，记为 $|\Sigma|$，其中：

$$|\Sigma|=\frac{V(V-1)}{2} \tag{4}$$

上式中的分母 2 表示我们默认随机网络为无向图。两点之间连接的概率为 p，不连接的概率为 $1-p$，对于某一个特定节点，在 $V-1$ 中连接方案中选择 k 个节点相连接的方案有 C_{V-1}^{k} 种，

该特点节点与此 k 个节点都有连接的联合概率为 p^k，因此其期望值为 $C_{V-1}^k p^k$，而与其他节点都没有连接，其概率为 $(1-p)^{(V-1)-k}$。即此时节点的度服从二项分布，即：

$$P(k) = C_{V-1}^k p^k (1-p)^{(V-1)-k} \tag{5}$$

当 $V \to \infty, p \ll 1$ 时，二项分布会转变为泊松分布，参数 $\lambda = (V-1)p$。

此外，随机网络中节点的平均最短轨迹距离 $< d >$ 很小，其大小和网络的大小呈对数关系，即 $< d > \sim \frac{\ln V}{\ln(<k>)}$，其中 $< k >$ 为平均度。

学者研究发现，对于连接概率 p 大于临界概率 p_c 的随机网络会是一个连通图，反之，随机网络不再是一个连通图，而变成了几个不连通的子图构成。

1.4.2 小世界网络

一些现实世界的网络表现出网络的小世界性，即大多数节点可以经过少量的几步到达另外一个节点。例如社交网络便具有明显的小世界性，世界上的每一个人都可能通过一个很短的关联与其他任意一个人产生联系。

为了构建一个具有小世界性的网络，我们可以采用如下的方法

- 第一步，形成包括 V 个节点的规则网络，每个节点与其相邻的 k 个节点相连接，累计有 $2k$ 个连接。
- 第二步，每一条边被随即重新连接，即对于网络中的任何一个节点 i，我们随机选择一个连接来替换它原来的连接。被选定的边，即连接节点 i 和节点 j 的边，以概率 p 任意重新连接到另一个节点 $u(u \in V, j \neq u)$，这样原来的边 (i,j) 便转换成了边 (i,u)。

当 $p=0$ 时，网络没有重新连接，此时网络仍然保持为规则网络。当 $p \neq 0$ 时，所有的边将在概率 p 下进行重新连接。随着 p 值得增加，网络的小世界性越来越显著。当 $p=1$ 时，网络将转换为随机网络，此时网络中节点度数分布的峰值接近 $2k$。网络的小世界特性最直接的理解是在这种网络上信息传播的速度非常迅速。

1.4.3 无标度网络

1998 年学者研究发现，某些网络中的极小数节点拥有很高的度数，而大部分节点只有很小的度数，基于这一发现，提出一种新的网络模型，叫作无标度网络，网络中节点的度服从幂律分布，即：

$$P(k) \sim k^{-\gamma} \tag{6}$$

其中 γ 是幂指数。在无标度网络中，给定标度指数 γ，随着度数 k 的增加，度数为 k 的节点个数急剧减少。网络的无标度性与网络受到随机攻击的鲁棒性紧密相关。在无标度网络拓扑结构中，主要关键节点周围有大量度数相对较小的次要节点。而次要节点周围则连接着度数更小的节点，如此发展直至网络的边缘节点。这种层次结构具有一定的容错能力。如果网络受到随机攻击，由于网络中绝大多数节点度数都较小，那么关键节点受到攻击的可能性是极低的。但另一方面，如果我们单独把几个关键节点从网络中提取出来，原始网络就转换为一组相当简单的图，任意一个关键节点的缺失都会造成巨大的影响。因此无标度网络对随机攻击具有较强的

鲁棒性，而对于蓄意攻击则很脆弱。相关学者使用渗流理论对无标度网络的这一属性进行了详细研究。

无标度网络的形成是根据节点的优先连接法则而产生的。这种方式可以从网络增长的角度来理解。这种背景下网络的增长过程即网络中节点数量随时间的推移而增加的过程。优先连接意味着某个节点的度越大，新加入网络的节点与其连接的可能性就越大。因为度数更大的节点具有更强的网络连接能力。比如说，在互联网上出现一个新的站点，这个站点的超链接更有可能指向那些知名的大站，而不是小站点。这就是优先连接法则。优先连接法则是正反馈循环的一种，在该法则下，随机变化会逐渐增强，因此个体间的差异会逐渐放大。

学者提出了利用这种优先连接法则生成无标度网络的算法。初始阶段，网络中有 V_0 个节点相连接组成。每一步在网络中添加一个新节点。每个新节点与现有网络中的 $V(V \leq V_0)$ 个节点相连接，其连接概率与当前网络中节点的度数成比例。新节点与网络中节点 i 相连接的概率 p_i 为：

$$p_i = \frac{k_i}{\sum_{j \in V} k_j} \tag{7}$$

因此网络中的关键节点或者度数更大的节点更容易与新节点相连，从而形成更多的连接；而度数较小的节点则很难与新节点相连接。

1.4.4 随机聚类网络

一些现实世界的网络，如社交网络和生物网络，呈现出模块化的结构特性，我们将其称为社团。这些社团的节点集合满足一个简单的条件：属于同一社团的节点有许多相互连接的边，而不同的社团由相对较小的边相连接。在随机聚类网络的形成中，两个节点如果属于同一个社团，那它们连接的概率为 p_{in}，而不同社团内的节点相连接的概率为 p_{out}。p_{in} 和 p_{out} 可以为任意值，它们主要用来控制网络平均度 $< k >$ 下社团内的联系 Z_{in} 和社团间的联系 Z_{out}。

典型的随机聚类网络中，p_{in} 值较大，p_{out} 值较小，即社团之内的节点连接紧密，而社团之间的节点连接稀疏。反之，当 p_{in} 值较小而 p_{out} 较大时，网络中节点的聚类效果不显著。基于以上参数，我们定义 $Z_{in}/ < k >$ 表示社团中节点的联系程度，而 $Z_{out}/ < k >$ 表示不同社区间的联系程度。评价社团检测技术优劣程度的 $Girvan - Newman$ 算法就是基于这里的两个定义。

根据以上讨论，典型的随机聚类网络必须满足 $p_{out} << p_{in}$。

1.4.5 核心 -边缘网络

网络可以采用局部网络，全局网络或者中等尺度网络的方法来描述。从这个角度而言，网络理论的一个主要目标是识别大型网络统计学意义上的主要结构，以便于分析和比较复杂网络的框架。在这个目标下，中等尺度网络结构的识别算法使得我们能够发现节点和边在局部网络以及全局网络中不明显的特征。特别是针对一种特定类型的中等尺度结构 –即社团结构 –的算法识别已经进行了许多尝试。其中称为社团的部分由密集相连的节点组成，而不同社团中节点之间的连接相对稀疏。

虽然对社团结构的研究已经非常完善，但对其他类型的中等尺度结构的识别，通常以不同形式的“块模型”来进行，却很少有学者进行研究。本节我们主要探索被称为核心 -边缘结构

的中等尺度结构。在社交网络中这种结构十分常见，在社会学，国际关系学和经济学等学科的研究中很早便已提出。识别核心 -边缘结构最常用的定量方法是在 1999 年提出。

这种方法通过计算确定核心 -边缘网络的结构，划分哪些节点属于密集连接的核心节点，哪些节点属于外围稀疏连接的边缘节点。其中核心节点应合理地与边缘节点相连接，而边缘节点不一定与核心节点或者相互之间相连接。因此，某节点当且仅当它与其他核心节点以及边缘节点“连接良好”时，该节点才属于“核心节点”。网络中可以有嵌套的核心 -边缘结构网络，以及核心 -边缘结构和社团结构相结合的网络。因此，开发一种能够同时检验两种中等尺度结构的算法是很有必要的。

关键节点是一种具有很大度数的节点，这种节点在现实世界中普遍存在。关键节点的存在常常会造成社团检测的失误，因为它们和网络中的许多节点相连接，它们与网络中几个不同的社团有很强的联系。当我们对某一个网络进行聚类分析时，对于其中的关键节点，采用不同的社团检测算法可以使它们分配到不同的社团中。因此，如何判断它们与不同社团之间关系的强弱至关重要，即使用允许社团之间有重叠的算法。这种情况下，一般意义上的社团概念可能并不能很好地实现对中等尺度网络实际结构的理解，而若采用核心 -边缘网络结构，将关键节点划分为核心结构来考虑可能更为合理。比如，我们可以把单个社团看做网络核心结构的一部分，整个核心结构由多个社团组成，社团之间可以存在重叠。

1.5 复杂网络的统计描述

1.5.1 度和度相关性

定义 1.30 密度 网络的密度主要用来衡量网络中各个节点间的连接强度。密度以分数的形式表示，它以实际的连接数为分子，所有可能的连结方案数为分母。

对于有向网络，密度 D 为：

$$D = \frac{E}{2C_V^2} = \frac{E}{V(V-1)} \tag{8}$$

具体来说，就是现有连接数和任选两个节点建立连接的可能连接数的比率。

对于无向网络，密度 D 为：

$$D = \frac{E}{C_V^2} = \frac{2E}{V(V-1)} \tag{9}$$

即不考虑方向的问题。

网络密度的取值区间为 $[0,1]$。当密度为 $D=0$ 时，此时的网络是一个零图，当密度为 1 的时候，此时的网络是一个完全图。

定义 1.31 网络同配性 网络同配性主要根据网络中节点的度，从网络结构的角度考虑网络中节点相连的可能性。同配性通常作为网络中节点相关性的判断因素，同配系数 r 是一种基于“度”的皮尔森相关系数。r 为正值时，表示度大的节点倾向于连接度数大的节点；r 为负值时，表示度大的节点倾向于连接度小的节点。通常来说，r 的值在 *-1* 到 *+1* 之间。$r=+1$ 时网络具有很好的同配性；$r=-1$ 时表示网络具有很好的异配性。在非零图 $G=(V,E)$ 中，u_e 和 v_e 表示某一条边 $e \in E$ 两个节点的度，则网络的同配系数 r 为：

$$r = \frac{E^{-1}\sum_{e\in E} u_e v_e - [0.5E^{-1}\sum_{e\in E}(u_e+v_e)]^2}{0.5E^{-1}\sum_{e\in E}(u_e^2+v_e^2) - [0.5E^{-1}\sum_{e\in E}(u_e+v_e)]^2} \tag{10}$$

学者们已经对此展开了很多研究，并且在现实世界网络的分析中进行了应用。例如，社交网络就表现出很明显的同配性。而科技网络，生物网络以及金融网络则表现出很强的异配性。

定义 1.32 局部同配性 局部同配性主要用于分析局部范围内的同配性或者异配性。局部同配性为每个节点在整个网络同配性中所占的比例，记做 r_{local}：

$$r_{local}(u) = \frac{(j+1)(j\bar{k} - \mu_q^2)}{2E\sigma_q^2} \tag{11}$$

其中，$\bar{k}$ 为节点 u 领域内的点的平均度数，E 为网络中边的数量，μ_q 和 σ_q 分别为该网络中除节点 u 以外的度数分布的均值和标准差。节点同配性系数 r_{local} 和网络同配性系数 r 的关系为：

$$r = \sum_{u \in V} r_{local}(u) \tag{12}$$

定义 1.33 非归一化的富人俱乐部系数 富人俱乐部系数最初被提出时是作为节点度数的不成比例度量参数。现在该系数主要以参数化的形式表示节点的度数 k，限定节点度数的取值范围。富人俱乐部系数衡量复杂网络的“富人俱乐部”现象的结构属性。这种性质是指节点度数较大的节点之间具有紧密相连的趋势，从而形成团状结构。

对于给定的阈值 k，$N_{>k}$ 表示度数大于 k 的节点的数量，$E_{>k}$ 代表与这些节点想连接的边的数量，那么该网络的富人俱乐部系数的形式为：

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)} \tag{13}$$

虽然网络同配性系数同样用于衡量度数类似的节点连接的可能性，但富人俱乐部系数可以看作是一个更为具体的系数，只关注那些超过一定度数的节点的相连的可能性。

定义 1.34 归一化富人俱乐部系数 对于非归一化富人俱乐部系数最大的争议是其没有很好地反映富人俱乐部效应，即便是对于随机网络而言，该系数也是单调增加的。因为度数更大的节点具有更大的概率相互之间建立连接，这未必反映特定的规律。为了真正体现这一规律，必须将其归一化，方法是：根据该网络的度分布，构建一个随机网络，然后计算随机网络的富人俱乐部系数 $\phi_{rand}(k)$，那么该网络归一化的富人俱乐部系数为：

$$\phi_{norm}(k) = \frac{\phi(k)}{\phi_{rand}(k)} \tag{14}$$

如果在某一特定阈值 k，归一化的富人俱乐部系数超过了 1，那么就说明该网络中存在富人俱乐部效应。

1.5.2 距离和路径

定义 1.35 直径 在网络 $G = (V, E)$ 中，节点间的最大路径长度称为网络 G 的直径，记为 T。

对于无向网络中，直径 T 的最大可能值为 $V - 1$。

定义 1.36 节点偏心率 在网络中，节点 u 的偏心率表示网络中其他节点与其距离最长的路径，记为 e_u。

定义 1.37 半径 在网络中，各个节点的最小节点偏心率称为网络的半径。

定义 1.38 维纳指数 在网络中，所有节点的距离之和称为维纳指数 λ。

但当网络中有大量分割子图的时候，维纳指数就会出现严重的偏差，因为不同的分割子图中的节点距离无穷大。即使不计算无穷大的情况，维纳指数也无法反映间断的局面。为了解决这一问题，引入如下两个概念：

定义 1.39 网络全局效率 在网络中，网络的信息传播效率记为 GE，它和网络中节点间的距离成反比，即：

$$GE = \frac{1}{V(V-1)} \sum_{u,v \in V, u \neq v} \frac{1}{d_{uv}}$$

定义 1.40 网络平均一致估计 网络的全局效率的倒数称为网络的平均一致估计，记为 $h = \frac{1}{GE}$。

网络平均一致估计很好地避免了维纳指数在不连通网络中的偏差问题，适合应用中不连通网络。

1.5.3 网络结构

定义 1.41 局部聚类系数 聚类系数是度量网络中节点聚集程度的系数。局部聚集系数量化了局部集聚的能力。从数学上讲，节点 i 的局部聚类系数为：

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

上式中 $|e_i|$ 表示节点 i 的邻域内节点间的连接边数，k_i 是节点 i 的度数。

定义 1.42 网络聚类系数 在局部聚类系数类似，网络的聚类系数用于度量网络的集聚情况，其数学表达式为：

$$CC = \frac{1}{V} \sum_{i \in V} CC_i$$

定义 1.43 循环系数 循环系数描述了复杂网络的流通度，它考虑了从 3 到无穷大的所有的圈。节点 i 的循环系数 θ_i 是连接节点和它两个相邻节点的最小圈数倒数的平均值，数学表达式为：

$$\theta_i = \frac{2}{k_i(k_i-1)} \sum_{j,k \in N(i)} \frac{1}{S_{jk}^i}$$

上式中 S_{jk}^i 是连接节点 i 及其相邻节点 j 和 k 的最小圈数。注意节点 i 的所有邻域节点对 (j,k) 都包含在内。如果节点 j 和 k 直接相连，那么这三个节点形成一个三角形圈，此时的圈数就是 3,；如果连接三个节点的路径只有一条，且不能形成圈，那么这条路径就是树，圈数是无穷大。

定义 1.44 网络全局循环系数 网络中所有节点的循环系数的平均值称为网络的全局循环系数 θ。

全局循环系数的取值区间为 $[0, \frac{1}{3}]$，当 $\theta = 0$ 时网络为树状结构，网络中不存在圈；当 $\theta = \frac{1}{3}$ 时，网络的局部聚类系数为 1，网络中所有的节点对均相连。

定义 1.45 模块化系数 模块化系数用于度量网络中某一特定聚类的可能性，即度量网络中聚类的强度。一般来说，模块化系数的取值区间为 *[0,1]*，当模块系数为 *0* 时，网络中不存在社团结构，即网络中的节点是随意相连的；随着模块化系数的增加，社团的结构越来越清晰，社团内的边在总边中的比例越来越大。无权重网络的模块化系数为：

$$Q = \frac{1}{2E} \sum_{i,j \in V} (A_i j - \frac{k_i k_j}{2E}) \mathbb{I}[c_i = c_j]$$

其中 E 表示网络中边的数量，k_i 表示节点 i 的度数，c_i 为节点 i 所属的社团；A_{ij} 表示两个节点之间的距离。

由于在计算中使用了指示函数，所以需要提前知道每一个节点所属的社团，同时不同社团内节点对模块化系数的计算没有影响。求和项中有两项，第一项表示社团内所节点的跨度，第二项表示需要将随机连接带来的影响进行扣除。非零的模块化系数表示忘了偏离随机性的程度，当数值大于 0.3 时，网络就具有较好的分离性了。

定义 1.46 拓扑重叠指数 拓扑重叠指数度量了网络中大致处于同一个社团中的两个节点的连接程度。本质上说，拓扑重叠评价两个直接或者间接相邻的节点的相似性，即比较这两个节点在邻域节点组上的重叠程度，领域也可以是不同深度的，比如 *1* 阶邻域代表直接的连接点，*2* 阶邻域需要额外增加和直接连接点直接连接的节点。

1.5.4 网络中心性

网络中心性度量网络中节点或者边的中心性或者重要性。首先想到的网络中心性度量的指标是节点的度数。在这个指标下，我们很自然地会假设度数比较大的节点是网络的中心。此外还有一些其他定义。

基于距离的网络中心性定义：第一是极小极大准则：即节点的中心性度量是其偏心率（距离其他节点的最大距离）的倒数，偏心率最小的节点具有最大的中心度；第二是极小求和准则：即节点的中心度取决于和其他所有节点的距离之和，距离之和越小，中心度越高。

基于路径的网络中心性定义：基于路径的网络中心性，计算时不考虑节点到节点之间的距离，这种方法重点考虑通过节点的路径数量。通过某节点的最短路径越多，该节点的中心度越高。

定义 1.47 介数中心性 介数中心性主要度量网络中每对节点位于最短路径上的程度。假设有一个节点相互交换信息的网络。首先简单地假设网络中的每一对节点以相等的概率交换信息，而信息总是利用网络的最短路径传播。那么某一个节点的介数中心性定义为

$$B_v = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \frac{\eta_{st}^v}{\eta_{st}}$$

η_{st}^v 表示节点 s, t 之间经过节点 v 的最短路径数，而 η_{st} 代表节点 s, t 的最短路径数量。

定义 1.48 连通度 在有一些情况下，难以保证网络上的信息都沿着最短路径运行，所以有额外的连接路径也是有意义的，所以将最短路径的数量和普通路径的数量相加就得到了该节点对的连通度 $G_{pq}(M)$:

$$G_{pq}(M) = \frac{1}{s!}P_{pq} + \sum_{k>s}\frac{1}{k!}(A^k)_{pq}$$

其中 P_{pq} 表示从 p 到 q 最短长度的路径数量；s 表示最短路径长度；A 表示网络的二进制邻接矩阵。G_{pq} 越大，说明 p 到 q 之间的路径越多。

定义 1.49 活力指数 假设有某一个函数映射 $f : G \leftarrow \mathbb{R}$，现有网络 G 在映射 f 下的值和移除节点 u 之后的网络在映射 f 下的值的差称为活力指数

$$V(G,u) = f(G) - f(G/\{u\})$$

。

定义 1.50 流介数中心性活力指数 节点 $u \in V$ 的最大网络流介数中心性活力指数

$$BV(u) = \sum_{s,t\in V, s,t\neq u}\frac{f_{st}(u)}{f_{st}}$$

上式中 $f_{st}(u)$ 是必须通过节点 u 的网络总流量，而 f_{st} 指的是不通过节点 u 的网络总流量

定义 1.51 紧密中心性活力指数 假设从节点 s 到节点 t 的距离表示信息在两者之间传递的费用。那么从网络中删除节点 u 后网络随机通信将会增加的费用，称为节点 u 的紧密中心性活力指数，

$$\begin{aligned} CV(u) &= I(G) - I(G/u) \\ I(G) &= \sum_{v,w\in V} d_{vw} \end{aligned} \tag{15}$$

定义 1.52 动态活力指数 假设一个包含 V 个节点的有向图，满足 $A\mu = \lambda\mu$，$v^T A = \lambda v^T$，其中 A 是网络的邻接矩阵，λ 为矩阵 A 的最大特征值，u 和 v 分别是矩阵的左特征向量和右特征向量。边 (i,j) 的动态活力指数 DI_{ij} 的数学表达式为:

$$DI_{ij} = -\frac{\Delta\lambda_{ij}}{\lambda}$$

其中 $-\Delta\lambda_{ij}$ 指的是移除边 (i,j) 之后的特征值变化量。同样节点 k 的动态活力指数为移除节点 k 后的特征值的变化值 $-\Delta\lambda_k$ 和原始特征值的比值，即:

$$DI_k = -\frac{\Delta\lambda_k}{\lambda}$$

1.5.5 复杂网络度量方法的分类

这里涉及到的各种度量复杂网络的方法可以按照使用的信息的不同进行分类

- 严格局部计算方法：这类方法仅仅根据节点自身的信息进行计算。
- 混合计算法：这类方法除了根据节点信息之外，还利用直接或者间接邻域内的信息进行计算。
- 全局计算方法：这类方法根据整个网络结构进行计算。

1.6 复杂网络上的动力学过程

图论和复杂网络的根本区别在于后者不仅研究网络的静态结构，还要关注网络的动力学特性。因此，本节将讨论网络的五种动态过程：随机游走，惰性随机游走，自避行走，游客漫步和流行病传播。

1.6.1 随机游走

给定一个网络和一个起始节点，随机选择一个邻居节点，将焦点移动至该邻居节点，然后再次选择邻居节点，然后移动，以此类推。如果是网络是加权网络，那么选择邻居节点的概率可以依赖于权重。这样的过程就是网络上的随机游走过程。本质上，网络上的随机游走过程就是有限的离散马尔科夫链过程，未来的某一个状态的概率只与当前的所处的节点相关，而与之前的轨迹无关。

定义 1.53 离散时间马尔科夫链 离散时间马尔科夫链是一个随机过程 $\{X_t : t \in \mathbb{N}\}$，其中，假设随机变量 X 在任何给定的时间上都取可数集合 N 中的值。转移到状态 q 的概率为：

$$P[X_t = q|X_{t-1}, X_{t-2}, \cdots, X_0] = P[X_t = q|X_{t-1}]$$

即下一个输出的概率只取决于该过程的最后一个值。所以，与过去的轨迹并不相关。

相关的概念包括**转移概率，转移矩阵**等。如果转移概率或者转移矩阵不随着时间发生变化，那么这个马尔科夫过程就是时间齐次的。

在随机游走过程中，采用传代时间来计算给定节点被访问的次数。

定义 1.54 传代时间 指的是一个函数 $pt : V \leftarrow \mathbb{N}$，$pt(q)$ 为马尔科夫过程访问状态 q 的次数。

$$pt(q) = |\{t \in \mathbb{N}|X_t = q\}| = \sum_{t=0}^{\infty} \mathbb{I}[X_t(w) = q]$$

基于传代时间，可以构造马尔科夫过程的势能矩阵。

定义 1.55 势能矩阵 $\boldsymbol{R}$ 指的是当我们从任何给定的其他节点开始，每一个节点被访问的逾期次数，即

$$R_{ij} = \mathbb{E}[pt(j)|X(0) = i]$$

这可以看做是从节点 i 开始的游走到达 j 的平均传代时间。

定义 1.56 循环状态 满足以下条件时，状态 j 具有周期性：

$$P(T < \infty|X(0) = j) = 1$$

定义 1.57 过渡状态 满足以下条件时，状态 j 具有短暂性：

$$P(T = \infty|X(0) = j) > 0$$