

第1章

绪 论

1.1 研究背景及意义

在现实世界中,许多个人或组织之间的复杂关系都以复杂网络的形式存在,如社交网络、引文网络和通信网络等。其中,网络中的节点代表个体(实体),边代表个体之间的关系。复杂网络科学作为一门跨越社会和计算机科学的现代学科,逐渐成为科学研究的引擎和跨学科的活跃话题。在复杂网络中,普遍存在三个重要的统计特性,分别为小世界特性^[1]、无标度特性^[2]和强社团结构特性^[3]。小世界特性反映了复杂网络短平均路径和高聚类系数的特征;无标度特性反映了网络中节点服从幂律分布的特征;强社团结构特性反映了网络呈现明显的社团结构(community structure)特征,即社团内节点比社团之间节点交互更为紧密。复杂网络分析任务为探索复杂网络的这些潜在特性提供了有效的工具。其中,挖掘复杂网络中的社团结构是网络分析的重要研究内容之一,对揭示网络内部结构、理解复杂网络潜在特性、发现网络隐藏规律等具有重要意义。因此,数十年来复杂网络受到了持续的关注和广泛的研究。

尽管目前尚没有关于社团结构统一的定义,但从网络链接关系的角度来看,学术界普遍认为社团是一组内部链接紧密而与外部链接稀疏的节点的集合,这样的社团结构普遍存在于现实世界的社会网络、生物网络、信息网络等复杂网络中。图 1-1 所示为一个科学家合作网络,图中节点代表科学家,节点之间的边表示科学家之间

作为共同作者形成的合作关系。可以发现,同一研究领域的科学家相互之间存在密切的合作关系,表现出很强的社团结构。在蛋白质交互网络中,社团结构代表同一细胞内具有相似功能的蛋白质群,社团内部的蛋白质之间交互非常密切,表示同一蛋白质复合体内部的蛋白质相互作用,共同实现特定的生物过程。在社交网络中,社团结构表示组成网络的个体通常具有相同的兴趣、职业或者具有相似的教育背景等;在万维网(world wide web, WWW)构成的网络中,同一社团内的网页通常具有相同或相关的主题;在新陈代谢网络或神经网络中,社团反映不同的功能单元;在食物链网络中,生态系统的子系统对应着复杂网络中的社团结构。这样的社团结构在其他很多现实网络中比比皆是,此处不再一一赘述。由此可见,社团结构广泛地存在于现实网络中,是了解网络结构和功能的重要手段。

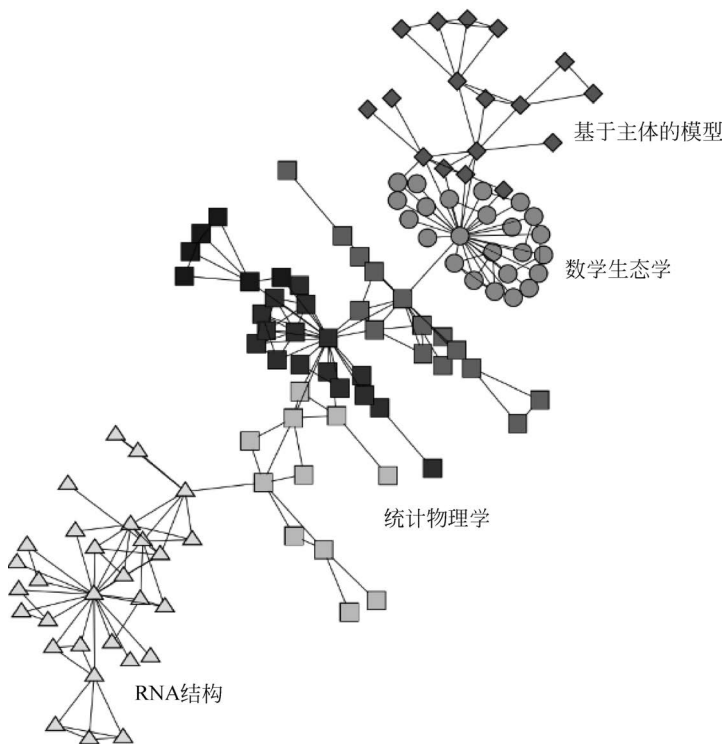


图 1-1 现实网络中的社团结构示意图(见文后彩图)

作为重要的网络分析任务,社团发现(community discovery)又称为社团检测(community detection),是指通过分析网络中节点之间的相互作用和潜在信息,从介观角度挖掘网络隐藏的社团结构的过程^[4]。通过挖掘社团结构,人们能够认清网络中隐含的内部结构,深入理解个体的行为特点和网络演化趋势,揭示网络中存在的普遍性特征和规律。精准挖掘社团结构不仅对揭示复杂网络的结构和网络功能具有重要的理论意义,而且对信息扩散、舆情传播、病毒感染等不同领域管理水

平与治理能力的提高具有重要的现实指导意义。例如,在微博等社交网络中,对拥有共同好友的用户推荐好友,有助于提高用户的忠诚度;对拥有相同兴趣爱好的用户推送感兴趣的内容,会提高用户的依赖度;在引文网络中,对引文网络进行社团挖掘,可以发现相同研究领域的关键性文章,以及历年来研究热点和学科建设的演变过程,有助于预测未来研究方向和前沿学科;在舆论网络中,对网络热点话题进行挖掘和分析有助于舆情控制和舆论导向,从而起到净化网络环境的作用;在犯罪网络中,对嫌疑人进行社团发现,可以侦查潜在的犯罪团伙或恐怖组织,对国家安全具有重要意义;在疾病蛋白质网络中,挖掘社团结构有助于找到药物靶标,对疾病的治疗起到关键的作用;在云计算中,分析服务间的流量挖掘虚拟网元之间的社团结构,可以为数据中心的微服务部署和调度提供指导意见,进一步优化运营商的效率,减少运营代价;在IP网络中,对IP网络的社团结构进行挖掘和分析有利于理解网络流量,从而为网络优化和安全管理提供有用的信息和决策支持;在通信网络中,对用户类别进行识别,有助于分析用户通信模式和特征,从而提供定向的推荐服务和安全布控;在卫星通信网络中,地球站通信单元可以建模成节点,通信单元之间的通信流量可以建模成边,通信单元之间的互通需求关系构成了一个复杂的通信网络,然后对卫星通信地球站中的通信单元进行社团发现,将通信频繁的通信单元组到一个子网中,通过一颗卫星转发与其他通信单元实现通信,有助于提高组网规划的准确率和工作效率,最大限度地利用卫星资源,实现高效智能的卫星通信组网规划任务。

1.2 基本概念

1.2.1 社团结构定义

社团结构是真实世界中许多复杂网络所具有的一种普遍性质。迄今为止,还没有关于社团的统一定义,但是很多研究尝试将社团的定义形式化。Fortunato^[5]将社团从三个层次进行定义:局部定义、全局定义和基于节点相似性的定义。之后,Fortunato又进一步将社团定义为可能共享相同属性或在网络中扮演类似角色的节点集合,并将其称为“群”。Porter等^[6]通过回顾社会学和人类学领域中关于社团研究的起源,将社团定义为紧密相连的节点组,组内节点间的连接比不同组节点间的连接更紧密。Yang等^[7]也提出了被广泛认可的社团定义,即社团是一组网络的节点,社团内节点之间连接紧密,不同社团节点之间连接稀疏。不仅如此,他们还从基于图划分、基于模块度和基于节点相似性等角度对社团结构进行了定义。Rheingold^[8]运用图论来描述社团结构,他将社团定义为网络中具有高度内聚性的节点集合,并提出了基于子图的社团概念,将社团视为复杂网络中若干节点子集,每个子集内的节点之间联系紧密,而不同子集之间的节点联系则相对稀疏。因此,

他认为社团发现的过程实际上是将网络节点根据它们的连接紧密程度划分为多个子图的过程。Girvan 和 Newman^[3] 提出了基于模块度的社团结构评价指标用于网络中社团内部连接的比例与在随机网络中相同规模节点集合的连接比例之间的差异。模块度的值越大,表明社团结构越明显。

社团根据网络中节点之间的相互作用对社团结构进行识别,从而达到挖掘网络中连接紧密、具有相同性质节点集合的目的。一个理想的社团结构应该是社团内部的节点之间连接紧密并且共享相同的属性或扮演类似的角色,而不同社团的节点之间连接稀疏。

1.2.2 社团发现评价指标

为评估不同社团发现算法的性能,研究者们提出了多种评价指标,这些指标可分为两大类:已知真实社团划分结果的指标和未知真实社团结构的指标。前者主要包括标准化互信息(normalized mutual information, NMI)、准确率(accuracy, ACC)、纯度(purity)、Jaccard 相似度(Jaccard similarity)等,后者主要包括著名的模块度(Q)指标。下面分别对这些指标进行具体介绍。

1. NMI

NMI 是众多社团发现算法评价指标,来源于信息理论,其利用信息熵衡量算法划分的社团与真实社团划分之间的差别。NMI 的定义为

$$NMI(C, G) = \frac{-2 \sum_{i=1}^k \sum_{j=1}^l F_{ij} \log_2 \left(\frac{F_{ij} n}{F_{i \cdot} F_{\cdot j}} \right)}{\sum_{i=1}^k F_{i \cdot} \log_2 \left(\frac{F_{i \cdot}}{n} \right) + \sum_{j=1}^l F_{\cdot j} \log_2 \left(\frac{F_{\cdot j}}{n} \right)} \quad (1-1)$$

式中, $C = \{C_1, C_2, \dots, C_k\}$, $G = \{G_1, G_2, \dots, G_l\}$, 分别是算法得到的社团划分和真实的社团划分; F 是混合矩阵,其元素 F_{ij} 表示既属于社团 C_i 又属于社团 G_j 的节点的数量, $F_{i \cdot}$ 和 $F_{\cdot j}$ 分别是矩阵 F 的第 i 行元素之和、第 j 列元素之和,代表社团 C_i 和 G_j 中节点的数量。NMI 的取值介于 $0 \sim 1$, 值越大说明算法的划分结果与真实社团划分越接近。若 NMI 等于 1, 说明两种划分完全一致。

上述 NMI 指标在检验非重叠社团发现算法性能中得到了广泛应用,但其并不能对重叠社团划分的优劣进行有效和准确的评估。因此,本章对 NMI 进行改进,提出一个能够用于评价重叠社团的标准化互信息指标——重叠归一化互信息(overlapping normalized mutual information, ONMI)。ONMI 的定义为

$$I(C : G) = \frac{1}{2} (H(C) - H(C | G) + H(G) - H(G | C))$$

$$ONMI = \frac{I(C : G)}{\max\{H(C), H(G)\}} \quad (1-2)$$

式中, $I(C : G)$ 为两种社团划分 C 和 G 的互信息,其示意图如图 1-2 所示, C 和 G 与式(1-1)一致,分别表示算法检测的社团划分和真实社团划分。ONMI 在 NMI

的基础上给出了更加直观的社团划分质量评价指标,可用于对重叠社团发现算法的性能进行有效评估。

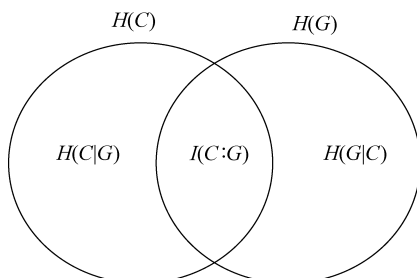


图 1-2 互信息示意图

2. ACC

ACC 是数据挖掘和机器学习领域中常用的检验聚类算法性能的指标,在社团发现问题中,任意两个相似的节点应被认作同一社团的成员。但在实际的社团划分中,会发生以下四种情况:①TP(true positive)表示相似的成员被划分到同一个社团中;②TN(true negative)表示不相似的成员被划分到不同的社团中;③FN(false negative)表示相似成员被划分到不同的社团中;④FP(false positive)表示不相似的成员被划分到相同的社团中。其中,①和②是正确的划分,③和④是错误的划分。ACC 定义为那些正确分配到同一社团的节点对的比例,其表达式为

$$ACC = \frac{TP}{TP + FP} \quad (1-3)$$

可见,ACC 介于 0~1,其值越大代表算法得到的社团划分与真实社团结构越接近。

3. 纯度(purity)

与 ACC 类似,purity 也是一个常用的聚类算法评价指标。针对社团发现问题,给定真实社团划分 $G = \{G_1, G_2, \dots, G_l\}$ 和算法得到的社团划分 $C = \{C_1, C_2, \dots, C_k\}$,其具体步骤如下:首先,为每个真实社团 G_j 分配一个标签,同时将其标签分配给社团内每个节点;其次,将算法得到的每个社团 C_i 中节点数最多的标签作为该社团的标签;最后,计算具有相同标签的 C_i 和 G_j 中相同节点的数量占网络中所有节点的比例,其数学表达式为

$$purity = \frac{1}{N} \sum_{i=1}^k \max_j |C_i \cap G_j| \quad (1-4)$$

式中, N 为节点个数。纯度的取值也在 0~1,其值越大代表算法的社团划分性能越好。

4. Jaccard 相似度

Jaccard 相似度又称 Jaccard 系数,用于比较有限样本集之间的相似性与差异性。在给定检测出的社团 C 和真实社团 G 的情况下,利用每个检测出的社团去匹配最相似的真实社团。Jaccard 相似度定义为

$$\text{Jaccard similarity} = \frac{1}{2|G|} \sum_{G_i \in G} \max_{C_j \in C} \delta(G_i, C_j) + \frac{1}{2|C|} \sum_{C_j \in C} \max_{G_i \in G} \delta(G_i, C_j) \quad (1-5)$$

式中, $\delta(G_i, C_j) = \frac{|G_i \cap C_j|}{|G_i \cup C_j|}$ 。该指标可用于对重叠社团发现算法的性能进行评估, 其值越大对应的社团划分结构越接近真实社团。

5. Q

在真实社团结构未知的情况下, 通常采用模块度(Q)来量化评价检测到的社团结构, 通过对比随机网络中连接密度和实际情况下社团中连接密度之间的差异来评价社团划分的质量。Q的形式化定义为

$$Q = \frac{1}{2m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \quad (1-6)$$

式中, a_{ij} 为矩阵 **A** 中的元素, **A** 为网络的邻接矩阵; m 为网络中的总边数; k_i 为节点 v_i 的度, 即 $k_i = \sum a_{ij}$; v_i 为网络中第 i 个节点; $\frac{k_i k_j}{2m}$ 为随机图中节点 v_i 和 v_j 之间有边的概率; C_i 和 C_j 分别为节点 v_i 和 v_j 所属的社团。式(1-6)适用于对无符号网络中社团发现算法的性能进行量化评估, 模块度的值越大代表社团划分的结果越好。

更进一步, 对于符号网络、重叠社团和多层网络也相继提出了对应的模块度函数用于评估社团发现算法性能。为了评价符号网络中的社团发现的质量, 提出可用于符号网络的模块度指标:

$$Q_s = \frac{1}{2w^+ + 2w^-} \sum_{i,j \in V} \left(w_{ij} + \frac{w_i^- w_j^-}{2w^-} - \frac{w_i^+ w_j^+}{2w^+} \right) \delta(C_i, C_j) \quad (1-7)$$

式中, w^+ 和 w^- 分别表示网络中正边和负边的数量; w_i^+ 和 w_i^- 分别表示与节点 v_i 相连的正边和负边的数量; V 为节点集合。

为了评价重叠社团发现的质量对传统的模块度进行了扩展, 定义了重叠模块度函数 EQ, 即

$$\text{EQ} = \frac{1}{2m} \sum_i \sum_j \frac{1}{O_i O_j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \quad (1-8)$$

式中, O_i 和 O_j 分别是节点 v_i 和 v_j 所属社团的总数量。

1.3 社团发现研究现状

在实际的复杂网络中, 社团结构通常隐藏于复杂的网络拓扑结构和(或)节点属性特征中, 难以轻松获取。社团发现的目的是通过有效的算法从复杂网络数据中定量挖掘出这种中尺度特征, 将连接紧密或具有相似特征的节点进行合理的划分, 以便揭示网络中节点之间、节点与社团之间以及社团与社团之间的相互关系。

社团发现作为重要的网络分析任务之一,一直是社会计算领域的研究热点,近年来陆续有大量的社团发现算法被提出。本书后续章节中会针对符号网络、动态网络、大规模网络等社团发现的相关工作进行总结,因此本节从方法论的层面总结社团发现研究进展中的一些经典方法。

1.3.1 基于层次聚类的社团发现方法

基于层次聚类的社团发现方法是根据网络的层次特征,基于节点相似性进行迭代聚类,从而识别出网络中的多级社团结构。层次聚类方法主要分为凝聚式层次聚类方法和分裂式层次聚类方法。

(1) 自底向上的凝聚式层次聚类方法是将网络中的每个节点都视为一个社团,然后通过对相似度高的社团进行迭代组合获得最终的社团结构。Newman 提出的 FN(fast Newman)算法^[9]首先计算任意两个社团之间合并后的模块度增量,然后将模块度增量最大的两个社团进行合并,直至获得网络的社团结构。标记传播算法(label propagation algorithm, LPA)^[10]在初始时为每个节点分配一个唯一标识,通过不断将多数邻居的标识更新为自己的标识,直至节点标识分布稳定,具有统一标识的节点构成一个社团,如图 1-3 所示。Zhang 等^[11]提出了一种基于真实连接概念的凝聚式方法来发现网络中的重叠社团,该方法通过对原始网络进行预处理得到“真实连接”图,然后通过不停地重复迭代,将相似度高的社团进行合并来获得最终的网络社团结构。Bahulkar 等^[12]提出了一种凝聚式层次聚类方法用于发现犯罪网络中的社团结构,该方法首先捕捉网络中隐藏的边并在划分社团前将其添加到网络中;然后采用自底向上的搜索方法,通过优化社团的局部模块度来发现社团结构。Riedy 等^[13]首先将网络中的每个节点看作一个社团,然后根据两个相邻社团合并后优化指标的增量对其进行评价,最后通过迭代将社团进行合并直到模块度目标被最大化。Blondel 等^[14]提出的凝聚式层次聚类算法首先采用局部搜索来选择小社团,然后对社团进行不断的聚和直到模块度停止增加。Shang 等^[15]提出了一种基于增量模块度的动态网络社团发现算法,算法首先采用静态社团发现方法初始化得到初始社团,然后执行增量更新策略来发现动态的社团结构。

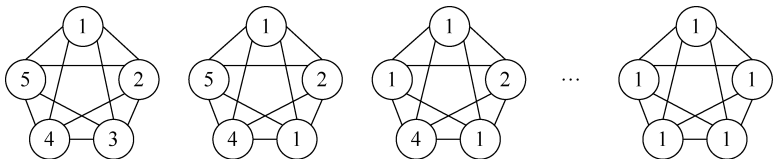


图 1-3 LPA 算法示意图

(2) 自顶向下的分裂式层次聚类方法是将整个网络看作一个社团,然后计算网络中节点之间的相似度,通过不断迭代删除网络中相似度最“弱”的连接,进而得到最终的社团结构。分裂式层次聚类方法的典型代表——GN(Girvan-Newman)

算法^[3],该算法通过迭代删除边界数最大的边获得社团划分的结果。Ni 等^[16]将网络的社团发现过程视为几何分解,并采用离散里奇流(Ricci flow)的基本原理,通过迭代地在里奇流过程中移动边来确定网络的社团结构。Görke 等^[17]提出了一个分裂式分层聚类算法用于发现时序网络的社团结构,该方法首先将网络中的所有节点看作社团,然后将合并后模块度增长最大的两个社团进行合并,重复此过程直到获得最终的社团结构。基于层次聚类社团发现算法的优点是能揭示网络的层次结构并直观地显示出来,可直观地对网络结构进行分析和研究。但是每个步骤完成都不能撤销,无法对次优或者错误的划分结果进行调整。

1.3.2 基于非负矩阵分解的社团发现方法

非负矩阵分解(non-negative matrix factorization, NMF)^[18]因其具有明确的物理意义和良好的可解释性被广泛应用于图像处理、语音识别和文本分析等领域。近年来,基于 NMF 的社团发现算法也受到了研究者的广泛关注。

NMF 可以看作一种数据的表示方法,即将原始数据矩阵中的数据用几个基向量的线性相加组合来表示。给定一个矩阵 $\mathbf{V} \in \mathbb{R}^{m \times n}$, NMF 旨在将矩阵 $\mathbf{V}$ 分解为两个非负矩阵 $\mathbf{W} \in \mathbb{R}^{m \times k}$ 和 $\mathbf{H} \in \mathbb{R}^{n \times k}$ 的乘积,使两个非负矩阵的乘积尽可能地逼近原来的数据矩阵 $\mathbf{V}$,即 $\mathbf{V} \approx \mathbf{WH}^T$ 。形式化地, NMF 模型可表示为如下的优化问题:

$$\min D(\mathbf{V}, \mathbf{WH}^T) \quad \text{s. t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (1-9)$$

式中, $\min D(\mathbf{V}, \mathbf{WH}^T)$ 表示矩阵 $\mathbf{V}$ 与 $\mathbf{WH}^T$ 之间的误差损失函数。常用的误差损失函数有以下两种:

1) 基于平方损失函数:

$$L_{\text{LSE}} = D(\mathbf{V}, \mathbf{WH}^T) = \|\mathbf{V} - \mathbf{WH}^T\|_{\text{F}}^2 \quad \text{s. t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (1-10)$$

式中, L_{LSE} 表示最小二乘误差损失; D 表示差异度量; $\|\cdot\|_{\text{F}}^2$ 表示 Frobenius 范数的平方, Frobenius 范数是矩阵元素的平方和的平方根,用于衡量两个矩阵之间的差异。

2) 基于广义 Kullback-Leibler 散度函数:

$$L_{\text{KL}} = D(\mathbf{V}, \mathbf{WH}^T) = \text{KL}(\mathbf{V} \parallel \mathbf{WH}^T) \quad \text{s. t. } \mathbf{W} \geq 0, \mathbf{H} \geq 0 \quad (1-11)$$

式中, L_{KL} 表示散度损失,用于衡量两个概率分布之间的差异; D 表示差异度量; $\text{KL}(\mathbf{V} \parallel \mathbf{WH}^T)$ 表示 $\mathbf{V}$ 和 $\mathbf{WH}^T$ 之间的 KL 散度。

在基于 NMF 的社团发现方法中,对邻接矩阵 $\mathbf{A}$ 进行非负矩阵分解,得到的非负矩阵 $\mathbf{W}$ 为基矩阵, $\mathbf{H}$ 为系数矩阵,矩阵 $\mathbf{H}$ 的每一列可以看作节点属于每个社团的隶属度^[19]。图 1-4 所示为基于 NMF 的社团发现算法的示意图,图左侧表示由 34 名成员构成的空手道俱乐部社交网络。在对其进行社团发现时,首先将网络表示为邻接矩阵的形式;然后对邻接矩阵进行非负矩阵分解,得到每个节点隶属于社团的隶属度;最后将节点分配到隶属度最高的社团,获得最终的社团结构。

针对社团发现的不同场景和需求, NMF 的一些变体也陆续应用于社团发现问

题,如对称非负矩阵分解(symmetric non-negative matrix factorization, SNMF)^[20]和非负矩阵三因子分解(non-negative matrix tri-factorization, NMTF)^[21]。SNMF 被提出并用于划分无向网络中的社团结构,将对称的邻接矩阵 $\mathbf{A}$ 分解为两个对称的低维矩阵乘积的形式 $\mathbf{H}\mathbf{H}^T$, 即 $\mathbf{A} = \mathbf{H}\mathbf{H}^T$ 。由于传统的二因子分解只能捕捉两种类型的关系, Ding 等^[22]将二因子的非负矩阵分解进行扩展,提出了可以捕捉更多类型关系的 NMTF, 即 $\mathbf{A} \approx \mathbf{W}\mathbf{X}\mathbf{H}^T$ 。其中, 矩阵 $\mathbf{X}$ 可以看作社团之间的关系矩阵, 用于描述社团之间的关系强度。更进一步, 对于无向网络的三因子分解, 可以得到 $\mathbf{A} \approx \mathbf{H}\mathbf{X}\mathbf{H}^T$ 。

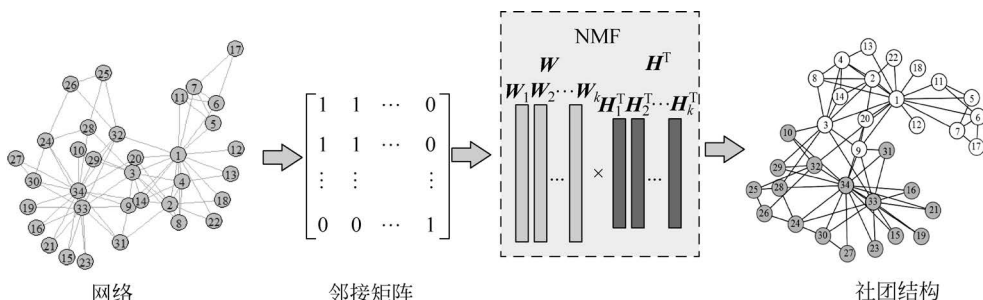


图 1-4 基于 NMF 社团发现算法示意图(见文后彩图)

近年来,因为 NMF 算法具有可解释性和模型简单等优点,越来越多基于 NMF 的社团发现算法被提出。Cao 等^[23]提出了一种基于 NMF 的社团发现方法——CLNCCD(combination of links and node contents for community discovery),算法同时考虑了网络拓扑和属性信息,并且认为在网络中拥有相同属性的节点有较大概率属于同一个社团。首先,基于 NMF 模型对邻接矩阵 $\mathbf{A}$ 进行分解,得到节点的社团指示矩阵 $\mathbf{H}$, 即

$$O^l(\mathbf{X}) = \|\mathbf{A} - \mathbf{H}\mathbf{H}^T\|_F^2 \quad (1-12)$$

式中,矩阵 $\mathbf{H}$ 的每一行为节点隶属于每个社团的概率。然后基于内容和节点所属社团的一致性来实现网络连接和节点内容的结合。算法通过引入图正则化项来惩罚拥有相同属性却被划分到不同社团的节点,则有

$$O^l(\mathbf{X}) = \|\mathbf{A} - \mathbf{H}\mathbf{H}^T\|_F^2 + \beta \text{tr}(\mathbf{H}^T \mathbf{L} \mathbf{H}) \quad (1-13)$$

式中,拉普拉斯矩阵 $\mathbf{L} = \mathbf{D} - \mathbf{S}$, $\mathbf{S}$ 为通过余弦相似度计算得到的节点属性相似度矩阵, $\mathbf{D}$ 为矩阵 $\mathbf{S}$ 的对角矩阵; β 为正则化项的权重系数。更进一步,考虑到网络中节点度的不均衡性,为了减少节点度的非均匀分布对图正则化的负面影响,算法引入参数矩阵 $\mathbf{W}$ 对社团指示矩阵 $\mathbf{H}$ 进行归一化,得到最终的目标函数:

$$O(\mathbf{X}, \mathbf{W}) = \|\mathbf{A} - (\mathbf{W}\mathbf{H}) \cdot (\mathbf{W}\mathbf{H})^T\|_F^2 + \beta \text{tr}(\mathbf{H}^T \mathbf{L} \mathbf{H}) \quad \text{s. t. } \mathbf{H} \cdot \mathbf{1}_C^T = \mathbf{1}_N^T \quad (1-14)$$

式中,向量 $\mathbf{1}_C^T$ 是所有元素都为 1 的 C 维列向量。Wang 等^[24]提出了社团发现方法——SCI(semantic community identification),算法同时集成了表示网络拓扑的

社团成员矩阵以及表示语义信息的社团属性矩阵。SCI 不仅能有效地进行社团发现,还能利用语义信息对社团进行标注,使社团发现结果具有较强的解释性。针对多维网络的社团发现问题,Zhang 等^[25]提出一种基于 NMTF 的多维网络社团发现方法——JoNMTF(joint regularized non-negative matrix triple factorization),算法首先结合网络连接关系和属性内容构成统一的图表示,然后利用 NMTF 模型对统一的图表示进行分解,得到准确的社团结构。对于重叠社团结构的发掘,GNMTF(graph regularized non-negative matrix tri-factorization)^[26]是一种基于图正则化的 NMTF 模型,用于发现网络中的重叠社团结构,该算法不仅可以准确地获得每个节点隶属于社团的情况,还可以捕捉不同社团之间的相互作用。针对符号网络的社团发现,Li 等^[27]提出了一种基于 NMTF 的符号网络社团发现算法,算法通过引入正则化项和稀疏项获得准确的符号网络社团结构,同时捕获社团之间的“积极”和“消极”关系。针对多层网络的社团发现问题,Ma 等^[28]提出了一种基于联合非负矩阵分解的多层社团发现算法——MjNMF(multi-layer community detection by using joint non-negative matrix factorization),该算法通过用公共基矩阵同时分解所有层的邻接矩阵来提取每层的节点特征,并通过分解所有相关层的节点相似度矩阵来获得网络的社团结构。

NMF 算法在社团发现中得到了广泛的应用,主要源于以下原因:①NMF 模型可以发现隐含的模块结构;②因为分解矩阵的非负特性,分解的形式和分解结构都具有明确的物理含义,并且具有非常好的可解释性;③模型简单,是可加性模型,可扩展性强,适用于不同场景下的社团发现问题;④NMF 分解的结果具有天然的稀疏性。

1.3.3 基于深度学习的社团发现方法

近年来,随着深度学习在视觉、自然语言处理等领域的发展,基于深度学习的社团发现方法也逐渐成为研究的热点。现有基于深度学习的社团发现算法主要分为以下几种:基于自编码器(autoencoder)的社团发现方法、基于图卷积网络(graph convolutional network,GCN)的社团发现方法等。

(1) 基于自编码器的方法利用无监督自编码器,将网络编码成潜在空间中的低维表示,并在低维表示的基础上划分社团结构。在基于自编码器的方法中有两种获得社团结构的方式,一种是在低维表示向量上运行聚类策略获得社团结构;另一种是将社团信息整合到模型中用于直接发现社团结构。根据使用自编码器类型的不同,又可以将模型分为栈式、稀疏式、去噪式和变分式自编码器。Tian 等^[29]基于自编码器和谱聚类之间的相似性,提出了基于稀疏自编码器的图聚类方法——SAE(stacked autoencoders)。SAE 将归一化后的图相似度矩阵输入稀疏自编码器中,通过重构相似度矩阵并在目标函数上加入稀疏性约束,得到节点低维表示为

$$\text{Loss} = \sum_{i=1}^n \|y_i - x_i\|_2 + \beta \text{KL}(\rho \mid \hat{\rho}) \quad (1-15)$$