齐 齐 哈 尔 大 学

外文翻译译文初稿

学 院： 计算机与控制工程学院

专业班级： 软件工程131

学生姓名： 许建华

指导教师： 王一萍

重叠社区发现方法：调查

Alessia Amelio和Clara Pizzuti

意大利国家研究委员会（CNR）

高性能计算与网络研究所（ICAR），

通过Pietro Bucci，41c87036仁德（CS），意大利

电子邮件：{amelio,pizzuti}@icar.cnr.it

摘要

由于个性化的态度，真实世界网络工作的探索，同时参与多个组等问题在近些年兴趣比之前高昂，重叠社区的检测因此成为一个具有挑战性的问题。本文描述了在该领域的主要对策。除了静态网络设计的方法，一些处理重叠社区的新方法按时改变，是可描述的。方法分类方面的基本原则引导他们获得一个网络划分组分享他们的一部分节点。对于他们中的每一个人，我们也报告，当需要用的时候，计算复杂度和网站地址就可以引用下载软件实现的方法。

1引言

复杂的网络构成代表关系的有效形式主义—船舶，构成许多现实世界系统的对象合作网络，互联网，万维网，生物网络，通讯与运输网络，社交网络只是一些例子。网络建模被作为图，其中节点代表对象并且边代表在这个对象的交互。复杂网络研究的主要问题之一是社区结构的检测，进入节点组（簇或模块）分为密集的内部连接，和稀疏的连接。在过去的几年中，许多不同的方法已被提出揭示网络中的社区结构。

许多社区明确了目标，朝着有效方法的努力，已检测发现不相交的社区导向。然而，社区可能重叠，即一些节点可能属于多个组。一个节点具有多组数据在现实世界中的网络非常普遍。例如，在社交网络中，一个人可以参与若干利益集团，在与一个网络进行合作的研究人员可能会与许多团体合作，在生物网络中，论文可能涉及多个主题，蛋白质在细胞中扮演不同的角色去参与几个过程。

在这篇论文中描述的重叠社区的最新建议—社区检测是给定的，和在不同类别的分类设置也被提供。算法已经被列在考虑的基本原则引导网络共享的方法来实现网络共享节点。

最近的许多描述社区检测算法的评论已经出版[ 12，20，11，7，38 ]。然而，我们的审查不同于那些[ 12，20，11，7 ]因为我们只专注于重叠的方法，并从[ 38 ]，也动态方法描述。

本文的结构如下。下一节给出了一些初步对方法的描述必要的定义。3节介绍了本文提出的方法分类。第3.1节描述节点种子和局部扩展方法。3.2节考虑集团展开方法。第3.3节描述链路聚类算法。第3.4节介绍标签

传播途径。这一方法的定义分类不可能在第3.5节报告类。第3.6节认为更多

动态网络的近期建议。第4节给出了一些可以用来测试算法性能的信息。最后，第5节提出和思考的方法和结论。

2初步了解

本节中的一些基本定义，一个清晰的认识的必要在调查中描述的概念将会给出。

网络N可以被建模为图G =（V，E），其中v是一组n = | V |对象，称为顶点，E是一组m = | E |链接，称为连接两个元素的边。网络中的社区是一组顶点 ，具有高密度的边缘，其中，和下群间边缘密度。在[ 11 ]，它是观察到一个正式的定义社会不存在因为这个定义往往取决于应用域。然而，令人印象深刻的方法已经被复杂网络中的检测社团提出。开始之前描述重叠的方法，由于它的普及和研究人员之间的大量使用，有必要引入模块化的概念，。

模块化的概念已被最初定义的Girvan和Newman【26】作为质量函数来评价划分的良性。然而,随着岁月，它已被公认为分区的最有意义的措施之一一个网络，更密切地同意与直观的社会概念广泛的现实世界网络。模块化的思想是随机图没有聚类结构，因此集群的边缘密度应高于预期一个图的节点,连接在随机密度。这一预期的边缘密度取决于所选择的空模型。模块化可以写在下面的方式：

A是图G的邻接矩阵，m是G的边数，Pi j是空模型中节点i和j之间的期望数。δ是Kronecker函数并产生一个如果我和J在同一个社区,否则为零。当假定随机图具有相同程度时K原始图的分布，PI = 2mj，Ki和K J的程度

节点i和j分别。

因为只有属于同一簇的顶点对的总和，模块化可以改写为;

其中k是在网络中找到的模块数，LS是在模块S中加入顶点的顶点，DS是加入节点是的位的和。因此每个节点和的第一项是边缘内的分数社区，第二个是预期值的分数的边缘。在网络中，如果边缘随机下降，不考虑社会结构。接近1的值表明强烈的群落结构。

3方法

在本节中，重叠社区的检测方法是描述。他们被分在六个不同类别的基

识别社区的方法学。类别如下：

--节点种子与局部展开

--集团的扩张

--链接聚类

--标签传播

--其他的方法

--动态网络

对于每个类别，简短描述了主要的共同特征，提供该类的算法。

3.1节点种子与局部展开

这些方法的基本思想是从一个节点或一个小节点，可以通过增加邻居节点，提高了一个社区质量功能。质量函数表征所得到的结构聚类。

博梅等人。[ 4 ]介绍了密度函数的概念与定义的社区作为子图是局部最优就这个密度函数。

其中E（C）是C和Eout内边数（C）是从c节点的节点的边缘，不属于C三重量或公制

这些指标测量集群内的通信强度，可以有效地更新，当新节点加入或从群集中删除。最后，测量两个簇之间的差异，汉明，或编辑距离，和非重叠的百分比定义。

为了对两个重叠的群落进行分类，提出了两种方法。第一个算法，迭代扫描（IS），从随机选择的边缘，称为种子，每次添加或移除一个顶点，直到选择的密度满足度量的提高。当没有更多的改进可以得到，该算法停止它重新启动一个新的种子。重叠是可能的，因为重新启动过程可以重新分配现有集群中的节点到另一个新的成形社区。

第二种算法，Rank Removal（RaRe）。假设有一些高级节点，当从图中删除，断开图成更小的连接组件，称为核心。然后删除已删除的节点对一个或多个核心。这意味着两个簇之间的重叠是可能只有通过这些顶点。

为了验证方法，每个群集之间的Hamming距离真正的聚类和所得到的聚类计算，然后算出所有这些距离的平均值。

一个随机边缘的选择能负面影响IS结果。因此，同一作者的修改他们的迭代扫描的方法，并提出了一个更有效的找到重叠社区[ 3 ]，命名为IS2，结合IS和RaRe。IS2也依赖于一个新的策略，用于初始化种子集群，能够一次计算每个节点的排名。该战略是基于观察，只有节点能够提高集群密度，要么是集群本身的成员，要么是直接邻居集群的成员，其中邻近簇是那些包含节点内的节点集群.因此，而不是访问每个节点都要迭代，节点不属于

这两个组中的一个，可以跳过。

该方法开始随机选择一个节点，并考虑它作为一个社区。然后对所有不包括S的邻居节点执行循环。为了要选择添加到S的邻居节点，选择是由计算每个节点的节点性，并增强与节点对性的最高价值。在这一点上每个节点的网格重新计算，如果一个节点出现负性价值是从S.进程停止时，所有尚未包含相邻节点的节点有一个负性。一旦获得一个社区，一个新的节点被选中和进程重新启动，直到所有的节点都被分配给至少一个组。作者发现，分辨率参数α= 1获得的部有关。然而，他们引入了一个标准，选择一个聚类的基础上稳定性概念。一个聚类被认为是稳定的，如果它被交付α值范围。这个范围的长度决定了更稳定的划分，这被认为是最好的结果。

另一种方法是lancichinetti等人应用。在[ 23 ]中选择一个节点添加到集群美国事实上，一种称为oslom使用统计检验成长的社区评估的统计意义一个节点。直觉的方法是，如果一个顶点V股份更多，与空节点的S在空模型中比预期的强大，那么关系V和S之间是出乎意料，因此V可以包括在S.

3.2集团扩张

派系扩张的方法是类似于以前描述的方法部分。然而，他们认为作为种子核心高度连接的节点N组—废气通过拉帮结派，然后生成合并重叠社区拉帮结派，应用不同的标准。设置[ 1 ]是一个系统来识别和可视化的重叠、密集的收集在无向图中节点组。它也允许浏览原始图和发现的社区。搜索算法采用集团设置渗漉法[ 29 ]和渗流集群主体。一个主体是一个完全图K节点构成。

两派都说相邻的，如果他们分享完全k 1节点。一个主体的渗流集群定义为所有的K联盟集团可以从彼此通过一系列相邻的k-cliques达到。参数k必须在输入中提供。K值越高，越小高度稠密组的大小。作者建议，4到6之间的值给出了最富有的群体结构。

3.3链接聚类

链路聚类方法提出的检测重叠社区划分集的链接，不是一组节点。为此，使用线图。无向图G的行图L（G）是另一个图L（G），如1）L（G）的每个顶点代表G的边，2个顶点）L（G）是依据—当且仅当它们相应的边在G中共享一个共同端点时一个直线图表示G的边之间的邻接应用聚类分析的线图是它产生的重叠图分—原有的交互图等，从而使节点存在多社区。

Pereira等人。[ 30 ]一直在用线图和重叠的RST蛋白质相互作用网络模块。为此他们应用了一口井已知的方法（MCL [ 8 ]），对蛋白质相互作用网络的线图。 Evans和Lambiotte [ 10 ]认为，任何算法分割网络可用于发现重叠社区结构的分区链接。他们首先回顾了定义模块使用的统计性质动态过程发生在一个图的边缘，然后提出链接应用新模块化概念划分。特别是，传统模块化Q的定义是在一个随机行走运动的联系网络。因此，这样的沃克将位于链接，而不是节点在每个时间t，其运动是在相邻的边缘，即链接有一个共同点。三个分区的质量函数网络G已被提出。每个不同的动力学过程的形式化并以不同的方式探讨了原始图G的结构。

在第一个动态过程，链接随机游走，沃克跳具有相等概率的相邻边。在第二过程中，链接—节点的链路的随机游走，沃克移动第一个相邻节点等概率，然后跳转到一个新的链接，选择平等的概率从那些节点处的新边。在最后一个过程中，动态是由原始随机游走，但预计在网络的链接的稳定性。

3.4品牌传播

在标签传播方法社区被认为是一组节点由同一属性的传播组合在一起，行动或信息—网络化。格雷戈瑞在[ 19 ]提出的算法椰干（社区重叠传播—

传算法），作为拉加标签传播技术的延伸等人。[ 32 ]。主要的改进在于分配多个社区认同—每个顶点的层次。因此，该方法与每个顶点x关联一组夫妇（C，B），其中C是一个社区的标识符和B是属于系数表示x的力量为社区成员开始给C.椰子每个顶点属于系数设置为1单标签。

然后，反复，每一个顶点x更新总结和规范属于系数标签—其相邻节点系数。新的一组x的标签是由邻居标签的联合。然而，为了限制通信数—一个顶点可以参与，参数v必须在输入。特别地,标签的归属系数小于1 / V被删除。椰子有一个不确定的行为，当所有属于系数对应与顶点相关联的标签是相同的，但低于阈值。在这样的一个随机选择的标签的情况下保持，而其余的被丢弃。最后，社区完全包含在其他被删除，并断开可以生成的社区在连接的。

3.5其他方法

在本节中，不能在上述类中的一个中进行分类的方法据报道。章等。【41】提出了一种检测重叠社团的算法。结合模块化的概念结构，光谱松弛和模糊C均值聚类。特别是，一个新的模块化功能扩展纽曼的模块化CON组—概念介绍，考虑到软任务节点的社区。最大化的模块函数的问题转化为特征—向量问题。 固定一个上限的k个数的社区，顶部计算了广义特征系统的K 1特征向量，并给出了一个映射网络中的节点到一个d维欧氏空间中进行的，在那里D≤K 1。之后，模糊均值聚类的最大应用组节点—模块化功能化改性。