**社交网络在线社区关键词频分析**

——基于图卷积网络社区发现算法

1. **引言**

## 研究背景

21世纪以来，特别是手机迅速发展以来，社交网络软件都是许多人生活的必备品。截止2021年6月30日，微信月活跃用户已经达到12.51亿人。不论是更封闭的微信还是更加开放的微博，社交网络都很大程度上地提高了信息传播的效率，在整个过程中，信息不断的被过滤然后再被传播。社交媒体是web2.0时代的产物，无论是对消费者、企业还是销售商，他都是一个非常热门的话题，也在逐渐成为时代的一个巨大特点。

首先，社交网络主要分为五大类:自我主义、基于社团、机会主义、兴趣驱动以及媒体共享。 自我主义主要就类似于 Facebook、微博等注重于个人平台，充当交友的社交软件。基于社团的社交软件会更集中于有共同特征的一部分人，例如黑人或者同性恋等。而机会主义的社交软件最闻名的就是国内外都具有影响力的“领英”:不同行业的人加入领英然后去建立更多的商业联系或者求职。兴趣驱动的社交网络会针对不同兴趣爱好的人群开展，更像是不同主题的社区平台。最后就是媒体共享的社交网络，这也是这两年在中国市场占有率突增的一个板块，例如抖音和YouTube。

而“社区”这个词，尽管社会学家对此有多种不同的定义，但许多学者认为，社区是需要建立在一定的地域基础之上的。但是在新时代中，社交媒体的出现已经导致人与人之间的联系不再依赖在地域上的局限性。这样开放的社交网络改变了传统的社会结构，在现在的社交网络系统中，每个人都能更加自由的发言，对于更开放的社交网络，我们能对所有用户发送消息也能获取到几乎所有用户愿意公开的基本信息。通过社交网络，人们有意或者无意的获取到了别人的信息，认识到了很多在以前他们不可能认识的人，对于个体来说，这样虽然加大了人们垃圾社交的冗余度，但同时也提高了满足社交需求的便捷度。但是，不论社交媒体怎样打破了从前对社区群体的定义，一个社区能够形成始终是归因于里面节点具有相似性。根据社交媒体中各个节点之间的关系能够构成一个网络世界的人际关系图，在这样的基础上，能够挖掘在新时代社交特征下多个具有不同属性的社区。

## 研究意义

从分析的角度来看社交媒体主要分为两种类型：一类为以Facebook和微信为代表的人与人之间联系的强关系网络，而另一类就是以微博、Twitter等为代表的基于关注与被关注关系的弱关系网络。

不管对于强关系网络还是弱关系的社交网络，其社区的存在都是依据一定的共同特征。在脱离了地域上的局限，所以相对于传统社区划分，社交媒体的社区划分能更大程度的体现个体之间更内在的联系，由此，对于销售商来说，能够更准确、更有效的获取具有内在共同点的群体，精准进行营销。

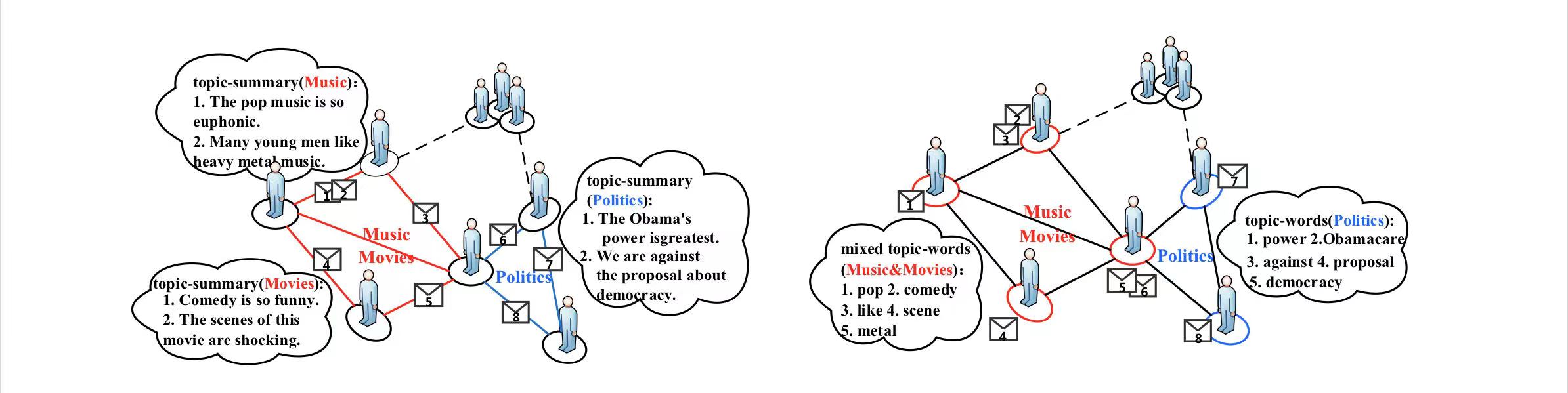
对于强关系网络的关系，其公司对其进行数据分析能够有效的划分真实生活中的社交圈，并对其进行相似性聚类和好友推荐等。但对于微博等弱连接网络，其大部分个人用户数据公开透明，采用社区的形式主要能够集中体现群体意识。但是对于使用常使用微博的普通人而言，基于关注与被关注的关系也能够非常有效生动的刻画出一个人的真实社交圈。另外，除现实生活中的交流者以外，在微博等能够产生关系的连接通常是基于有相似的兴趣爱好，于是，对弱关系网络的社区进行划分并对每个社区进行分析是有效且有意义的。

完成社区划分后，根据其公开的发布内容，基于自然语言处理，能够非常有效的分析出一个群体的兴趣点，并能够基于此兴趣点能够完成针对性的推荐等其他数据分析需求。除此以外，分析大量的社交网络数据有助于揭示潜在的社会结构，辨别组织行为和预测未来的趋势。由此可见，此选题有一定的现实意义。

## 研究创新点

现存的社区发现算法主要是基于统计概率的原理以及网络图论实现，但是在考虑在线社交媒体时，除了简单的相互关系以外，还有非常丰富的节点特征。对于微博用户而言，可以挖取到经过用户同意后公开的个人信息、发布的内容文本、以及其他例如注册日期等与账号紧密相关的特征。除了内容本身以外，每条内容背后的转发数量、支持数量以及评论数量也能够给分析者带来非常丰富的数据信息。另外，当对内容的分析从局限于用户一个人到一个社区内容的分析，能够非常大程度的提高数据量，使之泛化能力更强。

除此之外，考虑在线社交平台是一个动态的平台，且总体而言，对于一个稳健的社交平台而言，注销用户是一个小概率事件，而新用户的加入却是一个长期稳定的事件。在考虑社区发现时，可以更多的考虑新用户加入带来的社区结构改变。



## 研究内容及论文框架

1. **理论基础**

## 基于关系及文本的动态社区发现

* + 1. **社区发现**

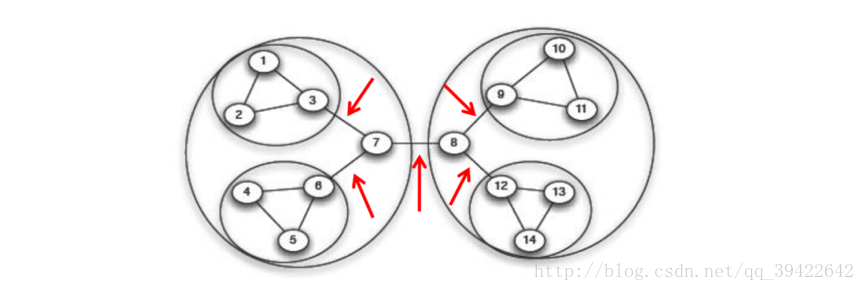
复杂网络是复杂系统的抽象，现实生活中许多复杂系统都能够通过复杂网络的相关特性进行描述以及分析。在复杂网络系统中，社区内部连接的密度高于社区之间的连接密度，结构是复杂网络的一个重要特征。

2002年以来，Newman提出“社区”的概念后，社区发现（community detection）就经历了非常高速的发展，直到现在，已经有非常多的基于图论以及其他统计概率模型的方式进行社区的划分与分析。社区发现的基本在于一个社区内部节点联系是较为紧密的，而社区之间的联系是较为稀疏的。在国内外的统计研究中，主要有使用基于层次聚类的社区发现算法、马尔可夫动态决策以及基于贝叶斯模型进行。但是常用的社区发现算法常常局限于网络拓扑结构图本身，而忽视了基于用户内容进行文本分析去社区划分与界定，所以可能会导致划分的精度不够大，最终推荐结果的针对性不够强的后果。

在传统的机器学习领域，社区发现的工作常常被看作为图聚类问题，吴恩达团队等人用特征向量实现了将节点划分到社区中的谱聚类方法，但是这种方法的性能较差，计算速度也较慢。近年来，随着深度学习算法的发展，研究者们逐渐从谱聚类等传统的统计学分析中解放出来，采用了效率更高的深度学习算法，2020年Philip S yu.团队根据深度神经网络、图神经网络等分析在此领域进行了更深一步的发展。显然当前的深度学习的能力无法完全替代统计分析处理复杂的网络结构，此团队也提出了这方面的机遇及挑战。[1]

图卷积网络是基于CNN研发的，面对图输入的数据，GCN展现出了更好的特性。Jin等人通过马尔可夫随机场解决了包含语义信息的带属性网络中的监督社区发现问题等。

国内的相关研究也是基于这几个大类算法进行，总体上也是对各个算法做出了改进。朱玲等人发现，直接使用节点系数很难刻画社区关系，需要定义一些基于节点聚类系数的社会度量去识别社区中的网络[2]。另外，赵宇红等人发现使用超图以及K-means算法能够有效的解决异质社区的划分等问题[3]。



* + 1. **动态社区发现**

随着时间的发展，社区内部的各个节点也会有很大的变化，动态社区发现是社区发现最具有挑战的研究方向之一。动态社区发现需要构建在动态网络上，初次之外，各个节点具有生存时间，节点之间的联系也有其生存时间范围。现在的动态社区发现主要有即时最佳社团发现（instant optimal CD）、时间权衡社团发现(temporal Trade-off CD)以及跨时间社团发现（Cross-Time CD）这三种算法。[4]

即时最佳社团发现的基础是当前的社区仅与当前的网络结构相关，本算法主要分为两步，首先需要将各个时间区间上的社区，然后将不同时间区间的社区联系，最终形成动态的社区结构。这种算法能够将不同时间上的社区发现任务并行处理，在之后的阶段再进行合并的操作。

时间权衡社区发现的方法为在最早的时间节点做第一次社区发现的工作，动态社区会在上一次的社区发现的基础上进行更新，并将新的社区发现结果输入，作为动态的社区结构。在考虑了时间因素进行平滑处理后，时间权衡的社区发现能够很有效的考虑网络进行巨变的情况。

跨时间的社区发现算法主要是把动态的网络压缩成为一个网络，并建立压缩网络和动态网络的节点对应关系，接着在压缩的网络上做静态的社区发现，再通过对应关系构建一系列的动态社区结构。

* + 1. **图卷积神经网络进行社区发现**

现在的研究中，利用图神经网络进行社区发现主要分为两大类：重叠的社区发现和非重叠的社区发现。这两类主要是在基本假设上有所不同。非重叠的社区发现假定识别出来的社区互不相交，各个节点只能属于一个社团。而重叠的社区发现的基本假定为在真实的社交媒体中，社区的情况常常都是有重叠的，一个节点会隶属于多个社区。重叠的社区更符合真实世界的规律，这也是现阶段社区发现的热点内容。[5]

* + 1. **文本分析社区发现**

在以前的社区发现算法中，也常常会考虑到文本分析的内容。但是文本分析常常会局限于单个节点的文本分析中，而很少考虑节点与节点之间的联系。另外，尽管一些文献考虑了文本与文本的之间的联系，更多的是假定了一个群体需要有共同的文本内容。但是事实上，不论是推特数据还是微博数据都显示了一个群体中，各个节点的信息可能会有不同，所以其实节点的文本信息并不能完美的和网络结构进行对应，所以算法的设计过程中需要分开对文本信息和网络信息进行建模。除此以外，常见对于文本信息的处理会集中于一个话题，但是在现实生活中，一个群体常常会有多个共同话题。最后，对于文本内容分析，不能采用简单的分词工具，获取到单独的一个常讨论的对象，更需要的应该是更完整的语言，需要更精准的表达出一个群体的观点和情感色彩。[6]

## 自然语言处理

1. **研究对象与数据**

## 研究对象介绍

在国内，各种社交媒体百花齐放，社交网络的各种形式都蓬勃发展。但是考虑到数据的可获得性、数据的丰富性以及数据的可挖掘空间，本文的研究对象选用了微博上的用户数据。截止2020年9月，微博的月活跃用户达到了5.11亿，最高日活跃用户达到了2.24亿。值得注意的是，其用户群体也体现了非常大的年轻化趋势，其中90后和00后占比高达80%，女性用户占比也远高于男性用户。在电子商务平台的报告中，年轻的女性也是占据了消费者的绝大部分。

根据2019年的统计，微博在全球的社交软件中排名第十，但是考虑到全球使用者最高的Facebook以及whatsapp无法在国内使用，微信、qq等软件的社交网络较为封闭，且用户在更公共空间上的发言较少，研究其内容的可行性和有效性较低，抖音的传播文件主要是以视频的形式。综合以上的所有因素，从数据可得性、丰富性以及可研究性，微博上的数据是更合适的。

## 网页数据爬取处理

如前文所述，本文将利用微博用户的关注者、粉丝以及其发布的内容构建网络系统，在信息的收集过程中，微博并未设计接口供研究者直接获取到数据，需要自行编写网络爬虫进行抓取数据，因而此部分将对后文获取数据所使用到的网络爬虫技术细节进行简要的介绍。

* + 1. **网络爬虫**

网络爬虫是是一段可以自动提取网页源代码、复制归档和保存其所访问过的网站页面上的信息至本地（localhost）的计算机程序。其基本工作流程为：首先，选取一部分种子网络地址（如，目标网站的主页）并将这些地址放入待抓取队列；其次，按照“先进先出”原则逐个打开队列中的域名地址并解析其主机IP地址，通过向主机地址发送请求以获取所需信息源文件，从源文件中提取目的信息，并将之存储于本地的信息数据库中；同时，将正在解析处理的源文件中包含但尚未访问的网络地址放入待抓取队列，从而进入下一个循环；反复前述步骤直至待爬取队列再次为空时停止爬虫程序的运行。

* + 1. **Ajax**

ajax即Asynchronous Javascript And XML，使用Ajax技术网页应用能够快速地将增量更新呈现在用户界面上，而不需要重新刷新整个页面，这使得程序能够更快地回应用户的操作。ajax其实有其特殊的请求类型，叫做xhr，其内容能够很容易的在浏览器中查阅到，也能够直接的获取到其request headers、request URL和response Headers。由于最原始的链接常常会采用javascript渲染，所以我们在每个页面能看到的真实数据并不是最原始的页面返回的，而是后来执行了javascript后再次向后台发送了Ajax请求，进一步渲染出来的。于是，采用Ajax爬虫获取数据能够观察到实际上我们真正需要的返回数据。

## 网络数据获取与存储

由于同一个用户具有多个信息，且信息是需要多层级的。本研究获取到的用户数据用json进行存储。另外，本研究的数据量较大，json作为轻量级的数据交换格式，采用完全独立于编程语言的文本格式来存储和表示数据，其层次结构简洁清晰，不仅可理解性强，同时也非常利于机器的解析和生成，能够有效的完成网络的传输。在数据获取的阶段，由于url接口获取到的源代码也是读取的json文件，所以在数据进行存储前，即可对有效信息进行第一次的提取，并将同一个用户的不同信息进行整合存储。另外，json数据能够非常清晰的表达出用户之间的双向关联关系，能够非常简便的绘出网络结构。

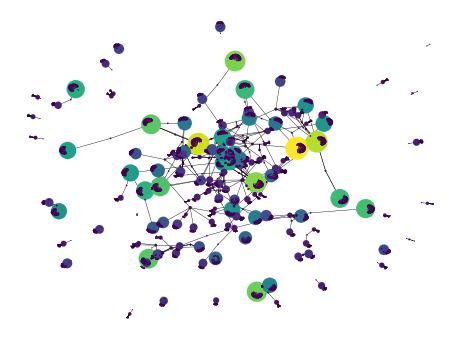
根据微博爬虫内容，本研究主要获取到的属性有，用户ID，所有粉丝用户ID，所有关注者用户ID，文本内容。在进行数据爬取前，由于每个用户的粉丝数量以及关注者数量不定，微博文本数量也不定，无法用传统的表格结构进行数据存储，而json能够非常自由的存储数据，以满足本研究的实验内容。

经过试验后，一位用户的数据存储空间大概需要0.2MB左右，对于一共4000位用户的所有数据，一共需要占800MB左右，不到1个G。

当读取783个用户的关联关系绘画其关系网络图时，如下图所示。虽然数据量还未达到一定的情况，但实际上已经能看到很明显的呈现出了社区内部聚集情况。

针对不同算法，除了Json外，也有多种不同的存储方式。基于上面内容获得的json文件，需要获得其特征邻接矩阵，网络结构等内容。

* + 1. **GML文件**
    2. **邻接矩阵存储**
    3. **Csv文件**



1. **模型建立**

## 基于图神经网络的结构社区分类

* + 1. **Bernoulli–Poisson 模型**

BP模型是一种允许重叠社区出现的模型的图生成模型根据关系矩阵, 可以生成对应的邻接矩阵，表示为：

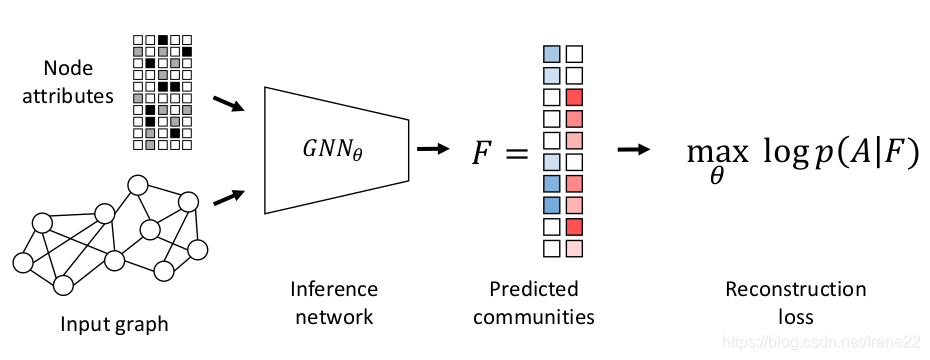
是隶属关系矩阵F中节点u的行向量，简单来说，u和v的共同社区越多，即的点积越大，他们的边就越有可能是相连的。

* + 1. **模型定义**

BP模型的最大负对数似然估计：

由于模型是非常稀疏的，所以第二项对损失的贡献会远大于第一项，通过不平衡分类可以平衡两项，使得损失函数更加合理。

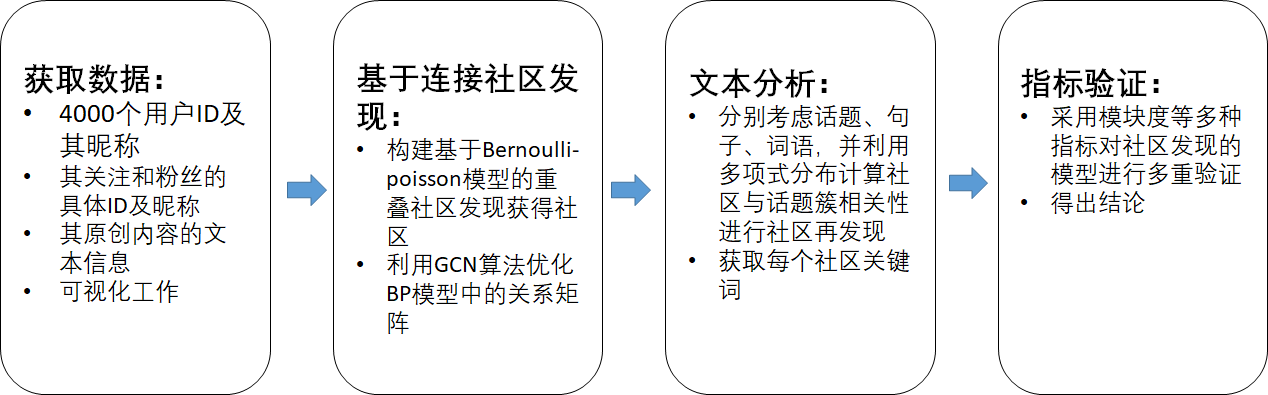
采用GNN可以直接寻找参数最小化平衡的负对数似然函数[7]：



## 基于文本内容的社区分类

1. **实验设计**

本文的实验流程如下图所示，可以看到，主要流程为：首先需要进行数据获取工作，并完成数据可视化展示，接着将各个节点用于构建网络模型，根据图卷积神经网络分析结合基础nlp进行社区发现，而后获取到各个群体的关键词，并将社区发现后的模块度等各种指标进行验证等，得出结论。



## 对比算法

* + 1. **louvain算法**

1. 模块度（modularity）：模块度是由Newman等人首先提出的，是目前使用最广泛的社区发现的优化目标，是一个用来衡量社区划分结果是否优良的评价指标。其根本的逻辑为，社区发现算法能够实现在社区内不得节点相似程度比较高，而在社区外部节点的相似程度比较低。总的来说，需要保证社区内部的边尽可能多，同时保证社区之间的连接尽可能的少。经过十余年的发展，模块度的定义也在不停地进行改变。

而louvain算法中，主要使用的模块度定义为以下公式：

即对于一个已经划分的社区，每个社区的内部边的权重之和减去所有与社区节点相连的边的权重之和。

另外，在代码实现阶段，模块度可以直接调用networkx中封装好的算法，直接进行计算。

* + 1. **FastNewman算法**

Fast Newman同样基于模块度，是一种贪心算法，其基本逻辑为：

* 将网络中的每个节点自定义为一个社区
* 计算出两两社区结合的模块度值，同时需要找到模块度增加程度最多或是减少程度最少的方式，进行社区的合并；
* 直到所有社区合并成为一个大社区时停止，并找出合并过程中最大的模块度结果，作为社区发现的结果。
  + 1. **LFM算法**

LFM算法是一个复杂网络中的重叠社区发现算法。LFM算法的基本思想为迭代地从不同的种子节点出发，以贪婪的方式检测社区。在每一次迭代过程中从不属于任何社区的节点中随机选择节点，并且为了确定某一节点是否属于当前社区，再计算社区中节点相对于社区的适应度之和：

内部度为社区内部节点的连接数量，外部度为与社区外部节点的连接数量。对于一个模块来说，他的内部度数等于模块内部连接数的两倍，外部度是将每个成员与图的其余部分连接起来的连接数。

和是社区G中节点的内部度和外部度。LFM算法是基于此度量方法，在检测过程中评估已经在社区中的节点的所有邻居节点，找到其中加入社区时能够增加社区适应度的节点，并选择其中的最佳贡献者加入到社区中。

一旦添加了新的节点到社区中，LFM算法会再次评估社区中现存节点对社区适应度的贡献，并移除贡献度为负的节点，即排除该节点能够增加社区适应度。当所有邻居的贡献度为负时，对一个社区的贪婪发现就此结束，当网络中的所有节点都已经分配社区时，整个过程结束。

所以LFM算法发现节点A的自然聚集的步骤为（若节点A在社区G中）：

* 在G的所有邻居节点中执行循环
* 将适应度最大的邻居节点加入G中，得到新的G’
* 重新计算G’中每个节点的适应度
* 若一个节点的适应度为负，则从G’中删除，生成一个新的 G”
* 若4发生，需要重复第三部，否则删除子图G”

当步骤1的所有都有负适应度时，进程结束。

## 模块工具

* + 1. **Networkx模块**

network是python的一个可调用的包，用于构建和操作复杂的图结构，提供分析图的算法。

另外，networkx工具可以以标准化和非标准化的数据格式存储网络、生成多种随机网络和经典网络、分析网络结构、建立网络模型、设计新的网络算法、进行网络绘制等。Networkx一共包含了四种图类型，包括无向图、有向图、两点间多边无向图以及两点间多边的有向图。

* + 1. **Community模块**

1. **实验结果与分析**
   1. **评价指标**

**模块度：**

模块度为Newman等人在2004年提出的，广泛应用于衡量社区划分的质量。一般适用于无向无权的同质网络。若社区内节点的边数相对越多，模块度值越大，相反，模块度越小。由此，若一个算法社区划分的效果越好，模块度越高。

从2004年到现在，模块度经过了非常多次的更新和变化。

其中常用的几个为：

模块度常用的一个原因为，其适用范围非常广泛，可以用于已知真实社区或未知真实社区划分的两种情况。

**ARI：**

**NMI（Normalized Mutual Information）：**

标准化互信息，也称为归一化互信息，也是目前使用非常广泛的一种社区划分评价指标，用于度量算法所得到的社区和真实社区划分之间的相似程度，也常常用于评价聚类结果的相似程度。

其中，矩阵N中的行表示真是的社区，而列表示不同算法得出的社区；而矩阵中第i行的元素表示为，第j列的元素表示为；

Nij​表示真实社区与算法所得到的社区相同的节点个数；S为节点个数；

CR​表示真实社区个数，CF​表示算法所得到的的社区个数。

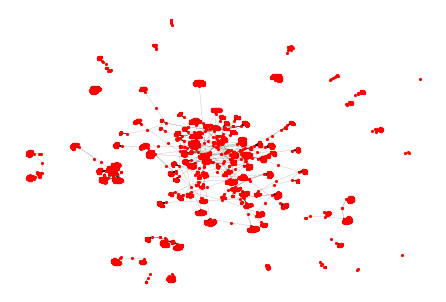
当算法得到的社区与真实社区完全一致时，NMI为1；而当算法得到的社区与真实社区完全独立时，NMI为0。

|  |  |  |  |
| --- | --- | --- | --- |
|  | NMI | 模块度 | ARI |
| LFM |  |  |  |
| Fast-Newman |  |  |  |
| NOCD |  |  |  |

当NOCD算法适用于不同数据集时：

|  |  |
| --- | --- |
| 数据集 | NMI |
| Facebook 348 | 34.7 |
| Facebook 414 | 56.3 |
| Facebook 686 | 20.6 |
| Facebook 1684 | 34.7 |
| Engineering | 18.4 |
| 微博 |  |

**LFM算法：**



1. **结论**

[1] Liu F , Xue S , Wu J , et al. Deep Learning for Community Detection: Progress, Challenges and Opportunities[C]// Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20. 2020.

[2]朱玲,余本国,冀庆斌.基于节点聚类系数的网络社区结构发现[J].中北大学学报(自然科学版),2021,42(05):435-440+459.

[3]赵宇红,张晓楠.基于超图和K-means改进的异质网络社区发现算法[J].计算机应用与软件,2021,38(10):290-296.

[4]端祥宇,袁冠,孟凡荣.动态社区发现方法研究综述[J].计算机科学与探索,2021,15(04):612-630.

[5]宁懿昕,谢辉,姜火文.图神经网络社区发现研究综述[J].计算机科学,2021,48(S2):11-16.

[6] Jin D , X Wang, He D , et al. Robust Detection of Link Communities With Summary Description in Social Networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, PP(99):1-1.

[7] Shchur O , S Günnemann. Overlapping Community Detection with Graph Neural Networks[J]. 2019.