# 复杂网络课程论文

## 一、网络数据集选择和介绍



**1. 数据集概述**

Everglades 数据集包含了不同物种在生态网络中的相互作用关系，并用于描述这些物种如何通过食物链、捕食、共生等方式与其他物种互动。这个数据集是一个典型的 **生态网络**，可以用来分析生态系统中的物种间依赖关系以及整个生态系统的稳定性。

**2. 数据集的来源**

Everglades 是美国佛罗里达州的一片湿地，拥有丰富的生物多样性。生态学家和研究人员通过对物种互动的建模与分析，旨在更好地理解物种如何相互依赖，以及这些依赖关系如何影响整个生态系统的功能。

该数据集是由 **University of Miami** 和其他研究机构的科学家合作收集的，目的是帮助研究生态系统中的复杂互动关系，如捕食者-猎物关系、竞争和共生等。



**1. 数据集概述**

**ENZYMES-g532** 数据集包含了 532 个酶的化学分子结构，每个酶由一个化学分子图表示。这个数据集的重点在于将化学分子结构与酶的分类信息相结合，目标是通过网络或图数据分析来预测酶的类别。

* 图的节点：每个节点代表一个原子。
* 图的边：每条边代表两个原子之间的化学键。
* 酶的类别：每个酶被分配一个标签，表示它所属的类别。

**2. 数据集的来源**

这个数据集来自 KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库，它是酶分类和基因组研究中的一个重要资源。ENZYMES 数据集提供了基于酶的化学结构图和分类信息，这使得研究人员能够应用图神经网络（GNN）、图嵌入技术等方法进行酶的功能预测、药物筛选等工作。



**1. 数据集概述**

infect-hyper 数据集是一个高阶图数据集，通常用于接触网络或传染病传播模型中的研究。该数据集是基于人群接触信息、网络接触模式或社会行为模式等信息生成的，用来研究不同个体间的传播与接触情况。

* 图的节点：每个节点代表一个个体或实体。
* 图的边：每条边代表两个个体之间的接触或传播可能性。

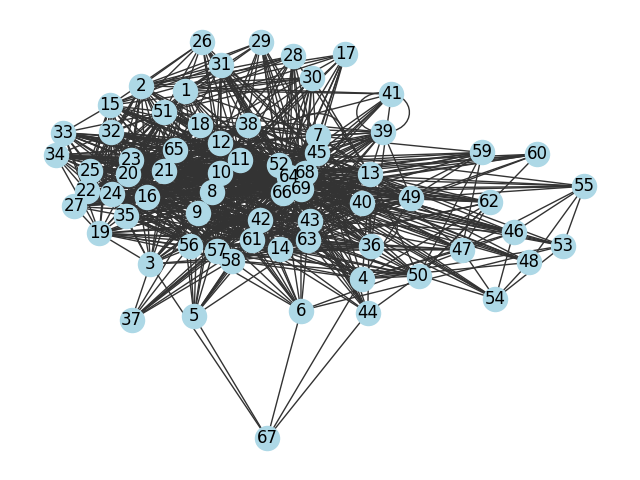
**2. 数据集的来源**

该数据集通常由 复杂网络研究或公共卫生领域中的研究人员收集，尤其是关于传染病传播的研究。这类数据集有助于研究个体间的接触模式如何影响疾病的传播，或如何通过网络结构分析来预测疾病的传播路径。

* 接触网络：网络中的节点代表个体，边则表示这些个体之间的物理接触、互动、或信息传播的潜力。
* 高阶网络：与传统的二阶图相比，高阶图能够表示多个节点之间的多重关系，即一条边可能连接多个节点，而不仅限于连接两个节点。

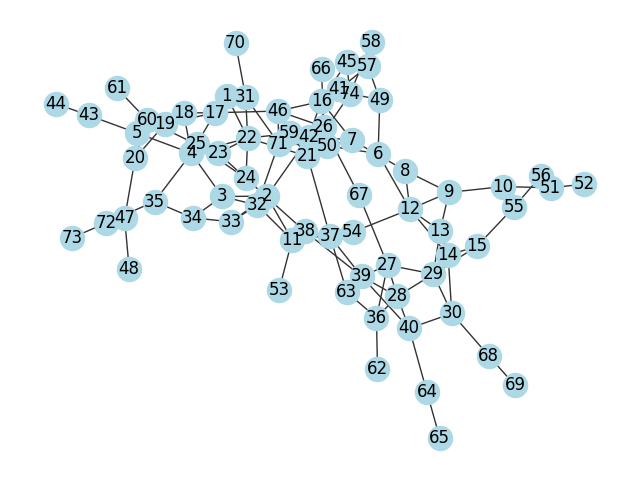
## 二、网络结构可视化





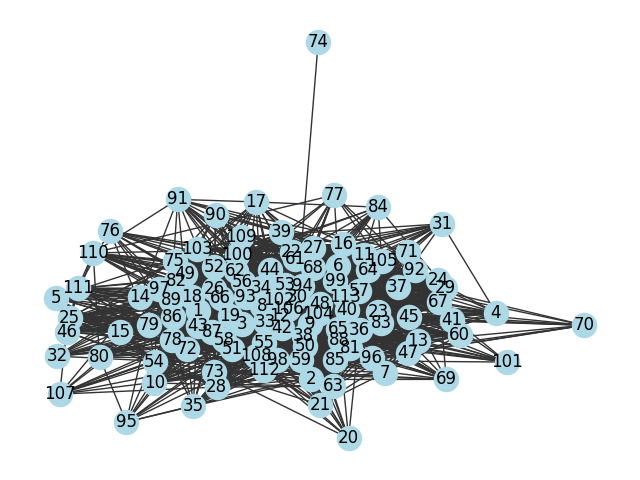
eco-everglades网络结构可视化





ENZYMES-g532网络结构可视化



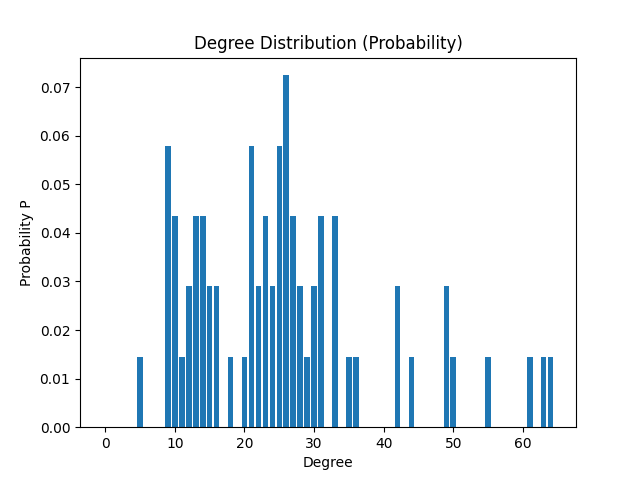


infect-hyper网络结构可视化

## 三、网络的统计特征

### 度分布

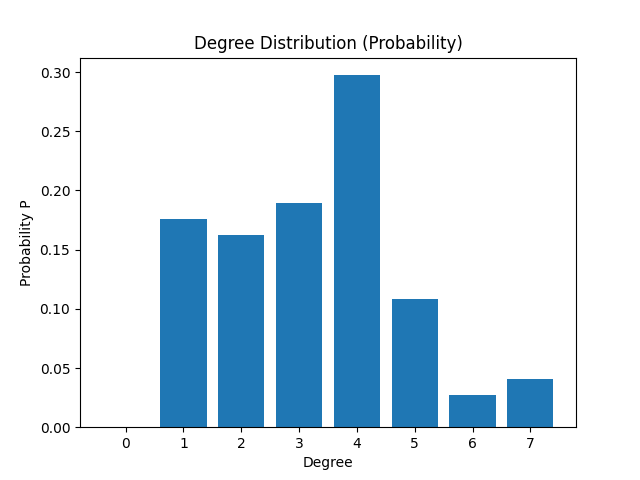




eco-everglades度分布可视化

**实际意义和原因分析**：图中显示的度分布呈现出长尾特征，即大多数节点的度较低，而少数节点的度非常高。在生态网络中，这种分布可能意味着生态系统中存在一些关键物种，它们与许多其他物种有着直接的相互作用。这些关键物种在维持生态系统的稳定性和功能中扮演着重要角色。除此之外，度分布的形状可能反映了食物链和食物网的复杂性。在食物网中，一些物种可能位于食物链的顶端，它们捕食多种其他物种，因此具有较高的度。而许多物种可能只与少数其他物种有直接的相互作用，因此度较低。度分布还可能反映了物种间的共生和竞争关系。具有较高度的物种可能在生态系统中具有竞争优势，因为它们能够与更多的物种建立联系，从而获取更多的资源。另外，度分布的形状也可能受到数据收集方法和分析方法的影响。例如，如果数据收集过程中对某些物种的观察不够全面，可能会导致度分布的偏差。

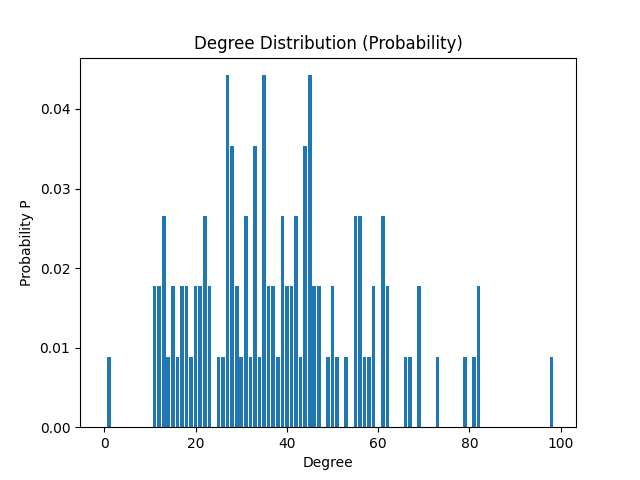




ENZYMES-g532度分布可视化

**实际意义和原因分析**：在化学分子中，原子的连接数受到其化学性质的限制。例如碳原子通常有四个价电子，这意味着它在稳定的分子结构中最多可以形成四个化学键。这种化学键的饱和性导致了度分布的上限。分子的复杂性通常与分子的大小和多样性有关。在ENZYMES数据集中，不同的酶分子可能具有不同的复杂性，这反映在它们的度分布上。度较高的原子可能位于分子的中心或活跃位点，而度较低的原子可能位于分子的边缘。此外，度较高的原子可能与酶的催化活性有关，ENZYMES数据集可能包含多种类型的酶，每种酶的分子结构和度分布可能有所不同，而度分布的形状可能反映了数据集中酶分子的多样性。

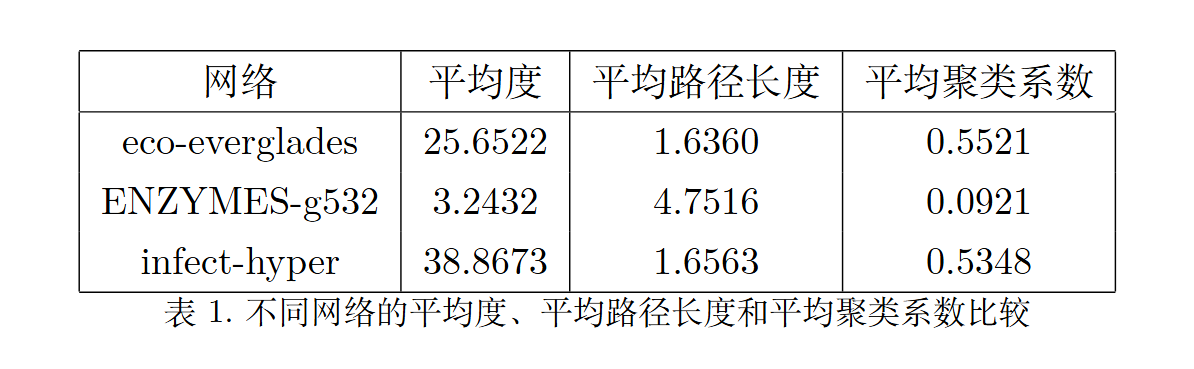




infect-hyper度分布可视化

**实际意义和原因分析**：度分布的长尾特性表明网络中存在一些个体与许多其他个体有接触，而大多数个体的接触相对较少。这反映了现实世界中接触频率的异质性，即某些个体可能有更多的接触机会。度较高的节点可能代表关键的传播节点，这些节点在疾病传播中可能起到关键作用。识别这些节点对于制定有效的公共卫生干预措施至关重要。度分布的形状可以提供网络鲁棒性的线索。如果网络中存在许多高度节点，那么网络可能对节点的移除更为敏感，这可能影响疾病的传播动态。度分布的形状也可能受到数据收集方法的影响。

### 其他统计特征





**平均度（25.6522）**:这个较高的平均度表明，在Everglades生态网络中，每个物种平均与大约25.65个其他物种有直接的相互作用。这可能反映了生态系统中物种间复杂的相互依赖关系。

**平均路径长度（1.6360）**:较短的平均路径长度意味着在网络中，任意两个物种之间的平均“距离”很短，这可能表明生态系统中物种间的相互作用较为紧密，信息或影响可以快速传播。

**平均聚类系数（0.5521）**:较高的聚类系数表明物种倾向于形成紧密的群体或社区结构，这可能与生态系统中的共生关系或食物网的模块化有关。



**平均度（3.2432）**:较低的平均度意味着每个酶分子中的原子平均只与3.24个其他原子形成化学键，这与化学分子的饱和性有关。

**平均路径长度（4.7516）**:较长的平均路径长度表明在分子结构中，原子之间的平均“距离”较远，这可能与分子的三维结构有关，原子间的联系可能需要通过多个中间原子来实现。

**平均聚类系数（0.0921）**:较低的聚类系数可能表明分子结构中原子的连接较为随机，缺乏明显的模块化结构，这与分子中原子的特定化学键合模式有关。



**平均度（38.8673）**:非常高的平均度表明在高阶网络中，每个个体或实体平均与大约38.87个其他个体有接触，这可能反映了传染病传播网络中个体接触的广泛性。

**平均路径长度（1.6563）**:这个较短的平均路径长度意味着在网络中，任意两个个体之间的平均“距离”很短，这可能表明传染病可以迅速在人群中传播。

**平均聚类系数（0.5348）**:较高的聚类系数表明个体倾向于形成紧密的群体，这可能与社会网络中的社区结构或群体行为有关，如家庭、工作场所或社交圈。

## 四、节点重要性

选择的节点重要性指标分别为：

**度中心性、接近中心性、介数中心性、特征向量中心性、Katz中心性、PageRank算法、LeaderRank算法。**

度中心性衡量一个节点与其他节点的直接连接数（即节点的度数）。即一个节点的度数越大，它在网络中与其他节点的连接就越多，意味着它可能在信息传播中更为重要。

度中心性高的节点通常代表网络中的“枢纽”或“中心”，它们直接与其他多个节点连接，因此容易影响整个网络。

接近中心性衡量一个节点到网络中所有其他节点的平均距离。换句话说，接近中心性高的节点能够在较少的步骤中接触到网络中的其他所有节点，信息传播效率较高。

接近中心性高的节点通常在信息传播中更为有效，因为它们可以更快速地到达其他节点。

介数中心性衡量一个节点在其他节点对之间的最短路径上所处的位置。具体来说，介数中心性高的节点在许多最短路径上处于中介位置，起到了信息传递的桥梁作用。

介数中心性高的节点常常在网络中起到关键的中介作用，是信息流通的“瓶颈”节点。

特征向量中心性通过考虑节点的连接质量来衡量节点的重要性。一个节点的特征向量中心性不仅仅依赖于它的度数，还取决于它所连接节点的重要性。如果一个节点与其他重要节点相连，它的中心性值也会较高。

特征向量中心性高的节点通常是网络中“影响力较大的节点”，这些节点往往与其他重要节点连接，从而有更大的影响力。

Katz中心性是特征向量中心性的扩展，考虑了节点与其他节点的连接，但在计算中对远离节点的连接赋予了衰减权重。Katz中心性通过考虑一个节点到其他节点的所有路径的数量，并对长路径赋予较低的权重。

Katz中心性有助于捕捉那些通过较长路径与其他节点连接的节点，从而能够识别出那些具有潜在影响力的节点。

PageRank是由Google提出的一个算法，旨在衡量网页的重要性。它基于链接结构的理念，认为一个网页的影响力不仅仅取决于它本身的质量，还取决于指向它的其他网页的影响力。换句话说，来自重要网页的链接比来自不重要网页的链接更有价值。

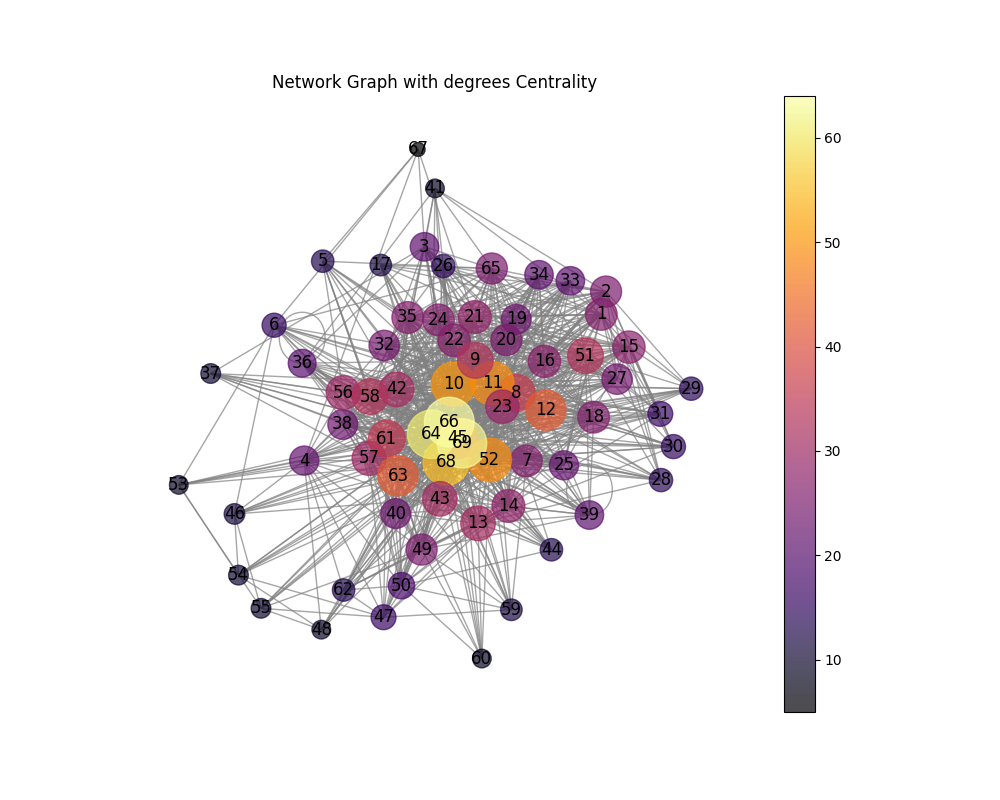
PageRank被广泛应用于网页排名，能够评估网页的影响力、受欢迎程度以及权威性。

LeaderRank算法是PageRank的变体，旨在解决PageRank中可能存在的“首位集中”问题。在LeaderRank中，节点的权重不仅由连接数量决定，还会考虑节点的相对影响力和连接的质量。算法通过迭代更新节点的排名，强调长期影响力而非短期连接。

LeaderRank算法改进了PageRank的局限性，尤其是在信息传播过程中，能够更好地识别出具有持久影响力的节点，而不仅仅是依赖于直接链接的数量。

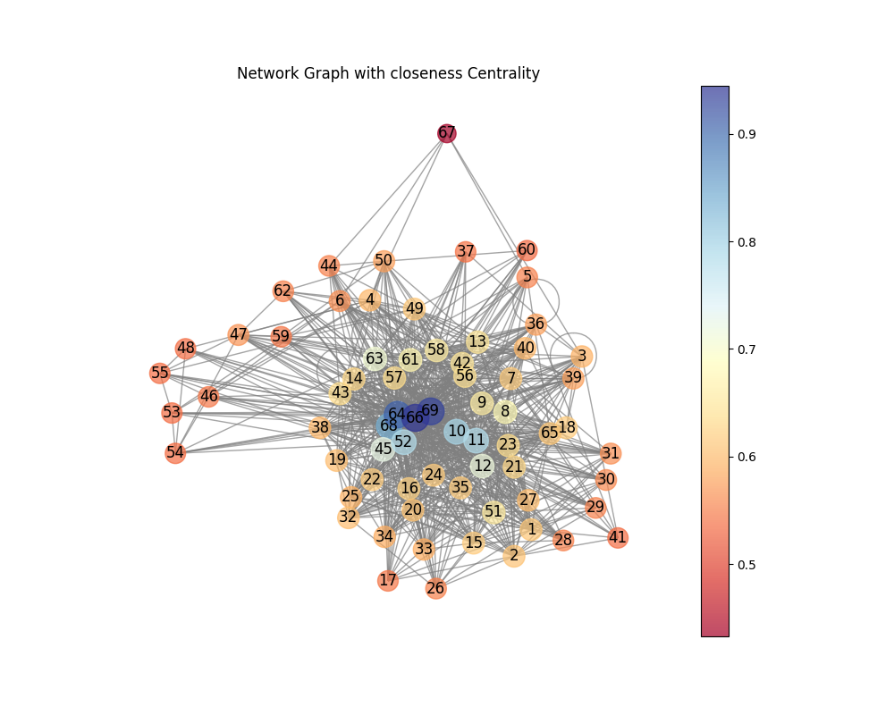
**对每一个网络根据每一个指标将节点重要性进行可视化分析，越重要的节点颜色越深，可视化的结果如图所示。**





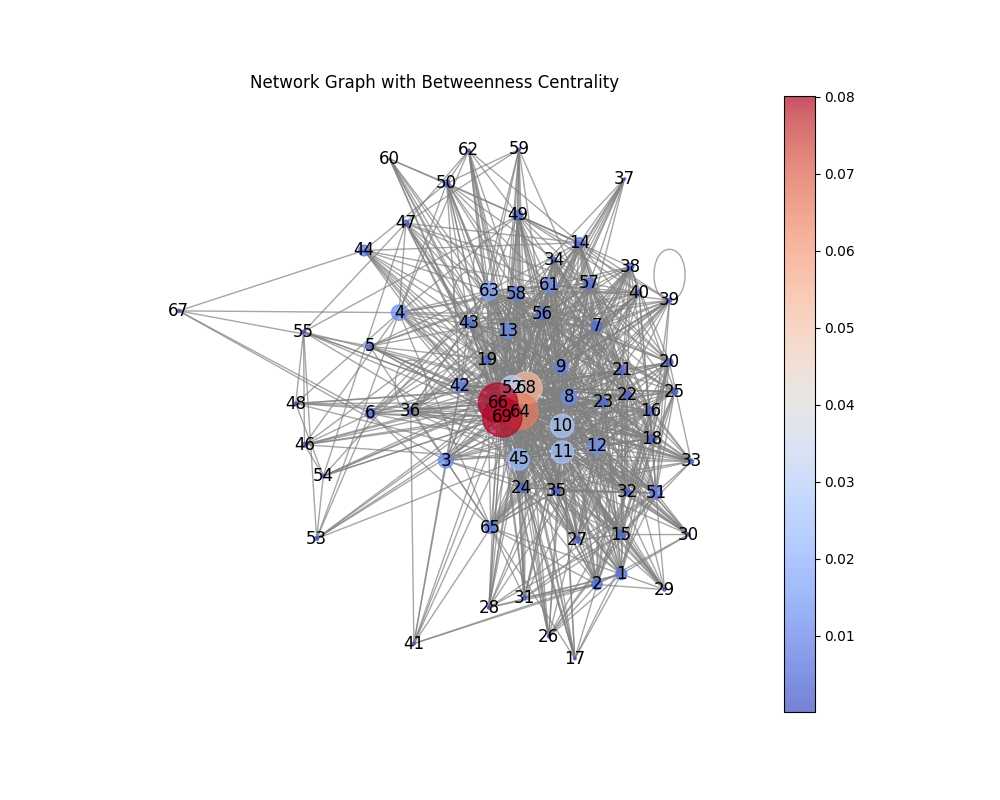
度中心性

在eco-everglades网络中，度中心性高的节点通常代表与其他物种有更多相互作用的物种。例如，在食物链中，度中心性高的物种可能是捕食者或被捕食者，它们与其他物种之间的相互作用非常频繁。此外，度中心性高的节点可能是生态网络的“关键物种”或“关键功能”，它们在维持生态平衡和稳定性方面起着重要作用。



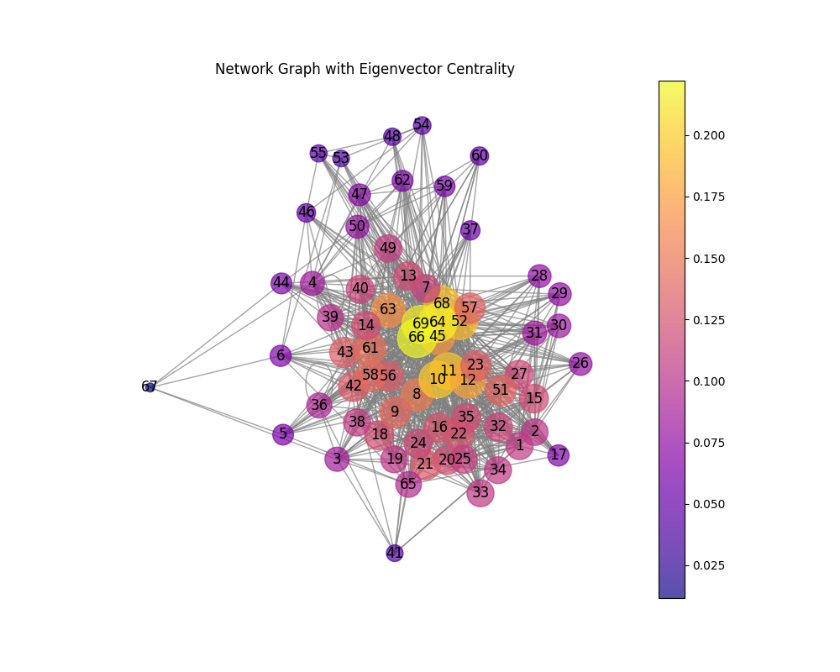
接近中心性

在eco-everglades网络中，接近中心性高的节点通常表示该节点能够在较短的时间内与其他节点建立联系。对于某些生态过程，比如物种间的能量流动、物质循环或信息传播，接近中心性高的节点可能是关键的传播者或桥梁。



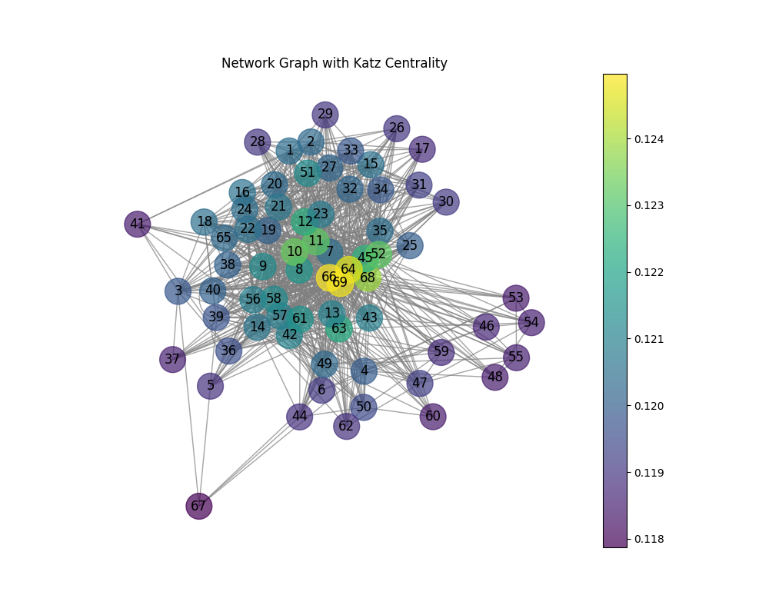
介数中心性

在eco-everglades网络中，介数中心性高的节点通常扮演着生态网络中的“关键物种”或“枢纽”角色。这些节点连接着网络的不同部分，通常在物种间的相互作用中起着“桥梁”作用。一个顶级捕食者可能通过控制多个物种的种群数量而影响整个生态系统的稳定性，或者一个植物物种可能为不同的动物提供食物来源。



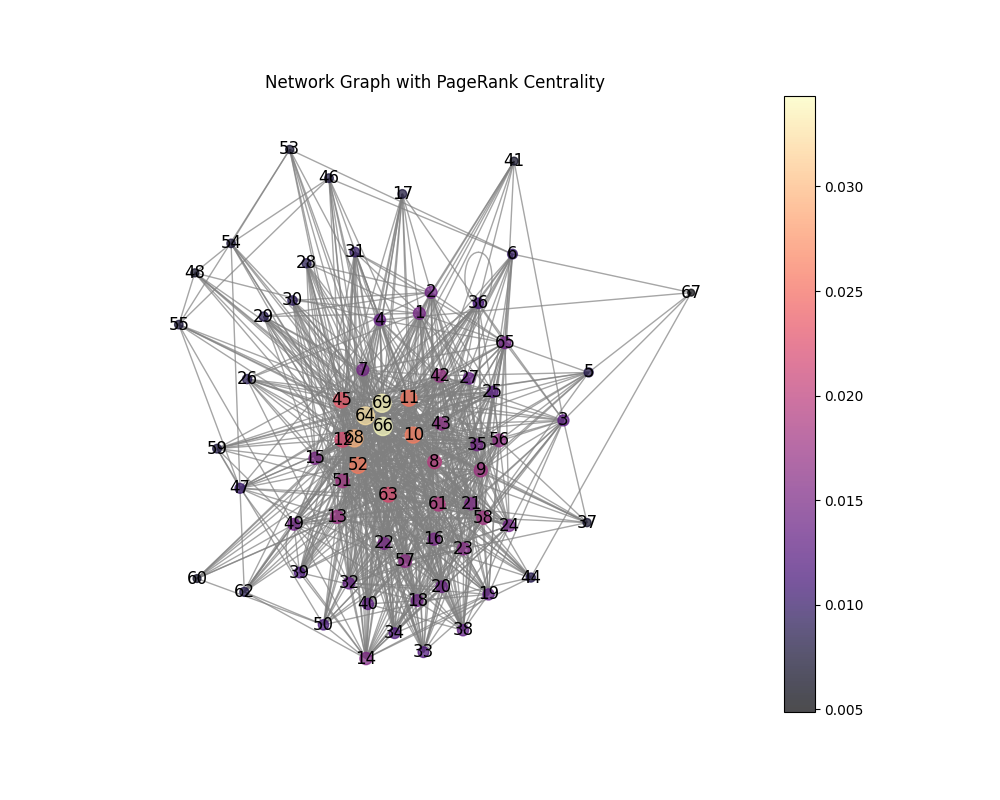
特征向量中心性

在eco-everglades网络中，高特征向量中心性的节点与网络中其他重要节点有较强的联系，影响着其他物种或资源的分布与流动。低特征向量中心性的节点即使连接了多个节点，但与生态网络中的其他重要节点联系较少，它们对系统的调节作用较弱。



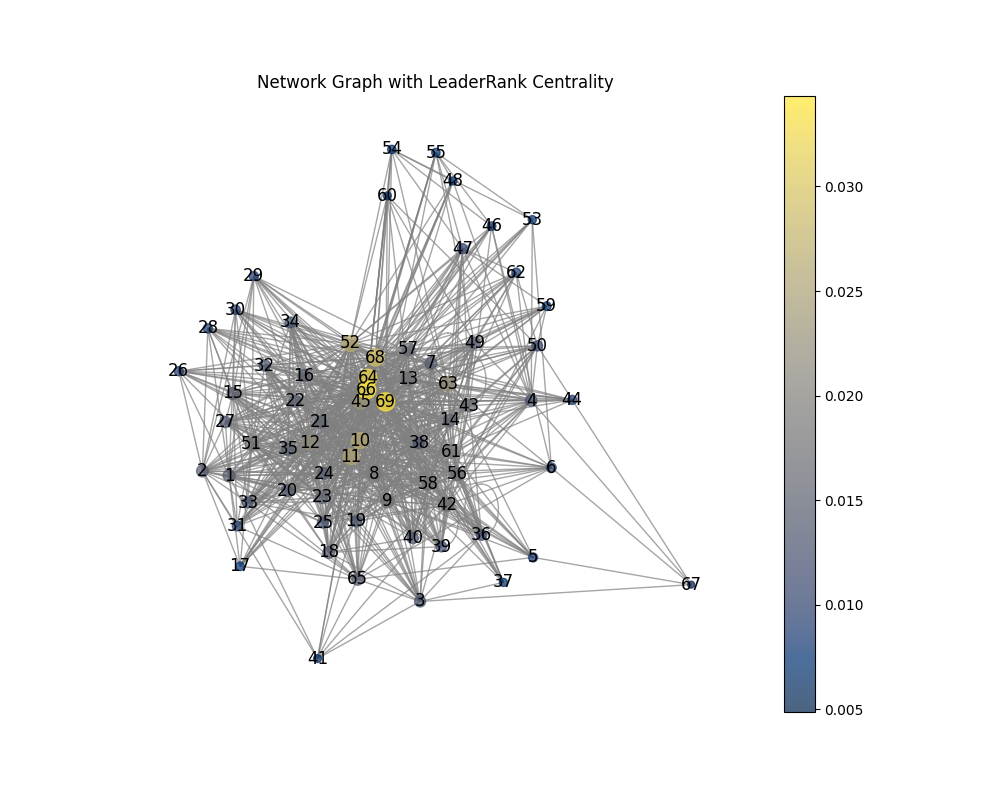
Katz中心性

在eco-everglades网络中，高Katz中心性的节点通过直接或间接的连接与网络中的其他节点建立广泛联系，代表了在传播过程中更容易接触到更多节点。这些节点在资源流动、信息传播和物种间相互作用中通常占据重要地位。低Katz中心性的节点连接较少、影响较小，通常处于生态网络的边缘，可能对生态系统的功能有较小的影响。



PageRank算法

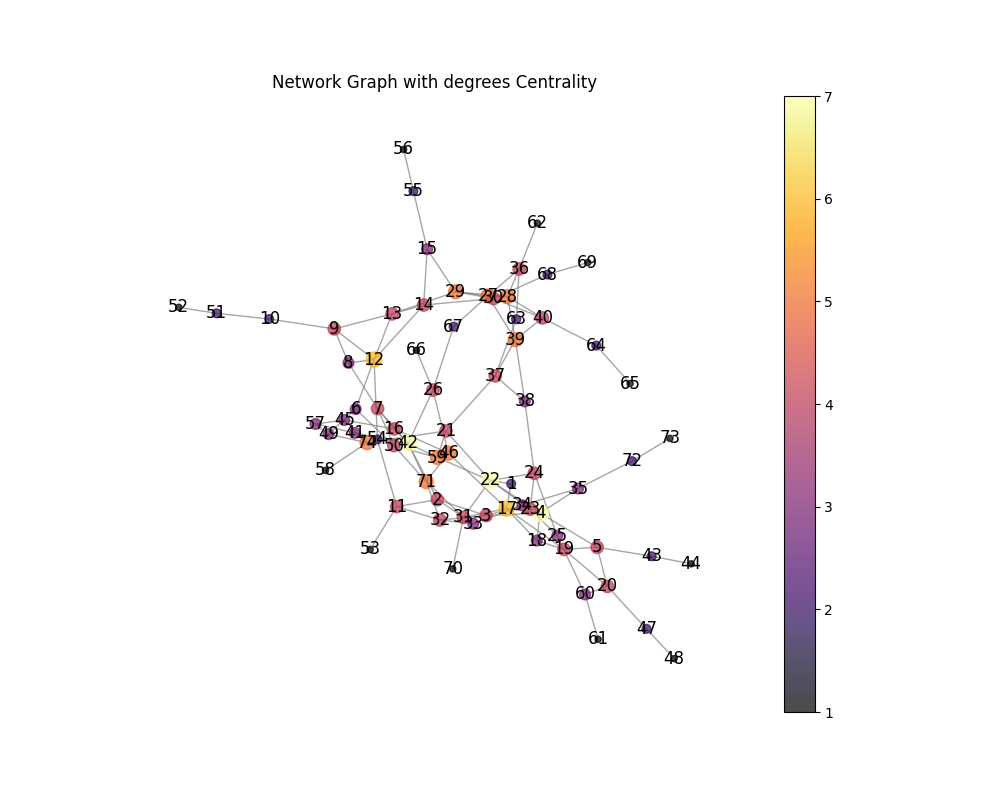
在eco-everglades网络中，高PageRank的节点通过与多个重要节点的连接，具有较强的“声望”或“影响力”。这些节点在生态系统中的作用类似于“枢纽物种”，它们能够有效影响其他节点的行为或生存。低PageRank的节点通常连接较少，且没有与其他高影响力节点建立强连接，因此它们的影响力和重要性较低。



LeaderRank算法

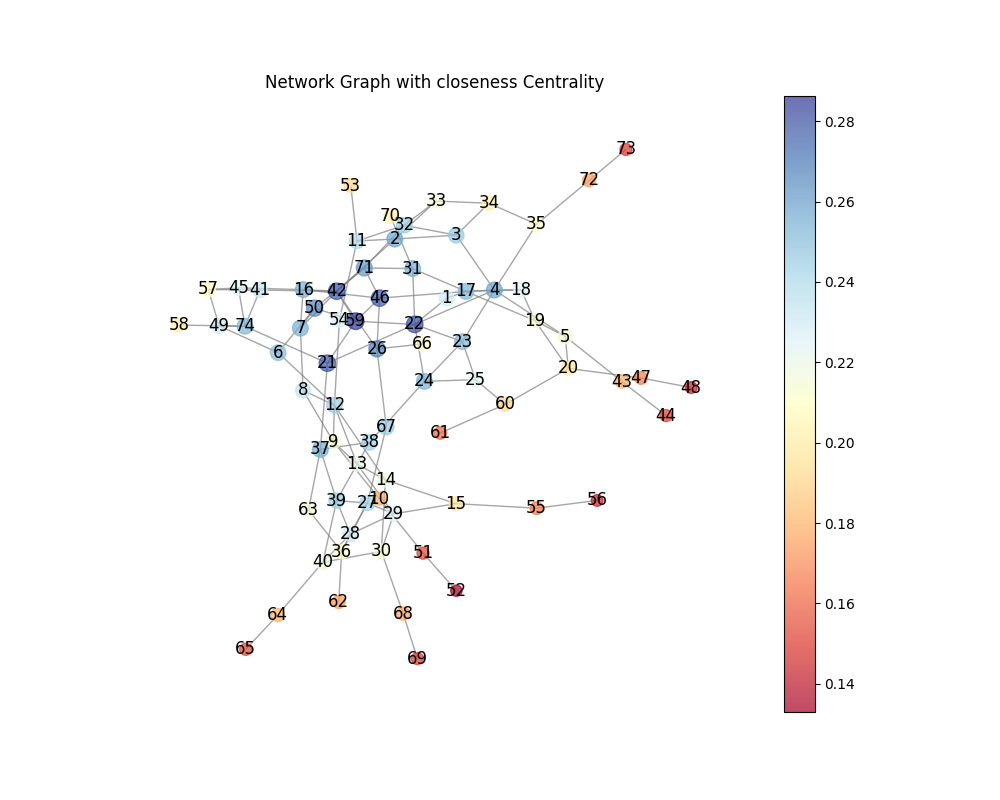
在eco-everglades网络中，高LeaderRank的节点是那些能有效地将资源或信息传递到网络其他部分的节点，代表着“领导者”角色。这些节点通常在生态网络中起着协调或引导作用。低LeaderRank的节点通常是跟随者，它们的影响力相对较弱，更多依赖于其他重要节点进行资源交换或信息传播。





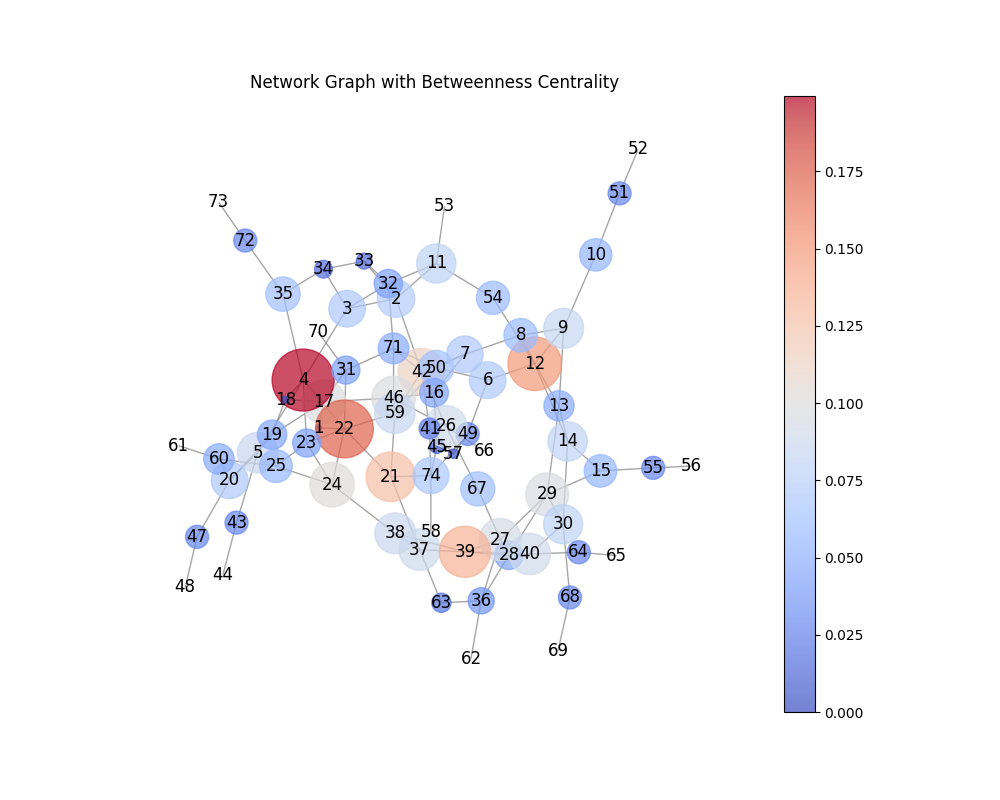
度中心性

高度中心性的节点与许多其他分子直接相互作用，可能是多条生化反应路径的关键酶或“枢纽酶”。这些酶通常参与多种生化途径，调节网络中其他酶的活性。低度中心性的节点表示该酶与其他分子的联系较少，可能是局部反应中的专一性酶，影响较为局限。



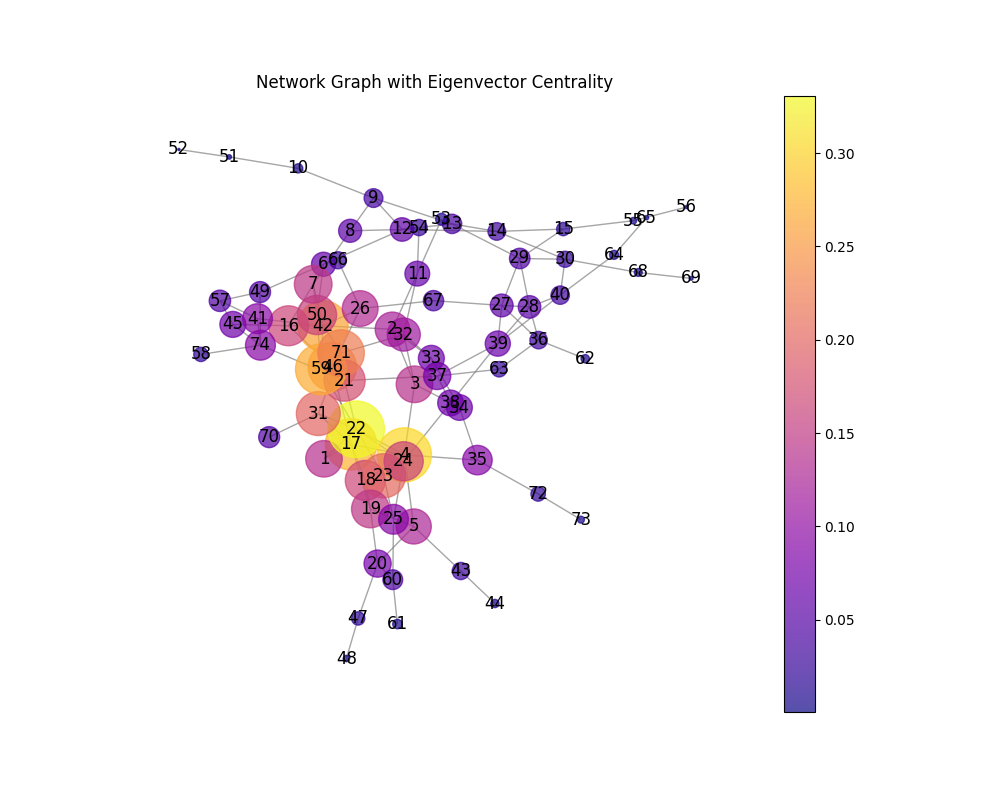
接近中心性

高接近中心性的节点与网络中其他节点的平均距离较短，意味着这些酶能够快速影响或传播其作用至其他反应或分子。它们在生化反应中通常起到快速调节的作用。低接近中心性的节点表示该酶与其他反应途径中的分子之间的联系较远，可能对网络的快速响应和调控作用较小。



介数中心性

高介数中心性的节点处于多个反应路径之间的“桥梁”位置，承担着在不同分子群体之间转导信息或物质的功能。这些酶通常是网络中的关键调节者或控制点，决定着多个反应的流向。低介数中心性的节点表示该酶较少充当网络中其他反应路径之间的中介，通常参与局部反应，不直接调控跨越多个路径的反应。



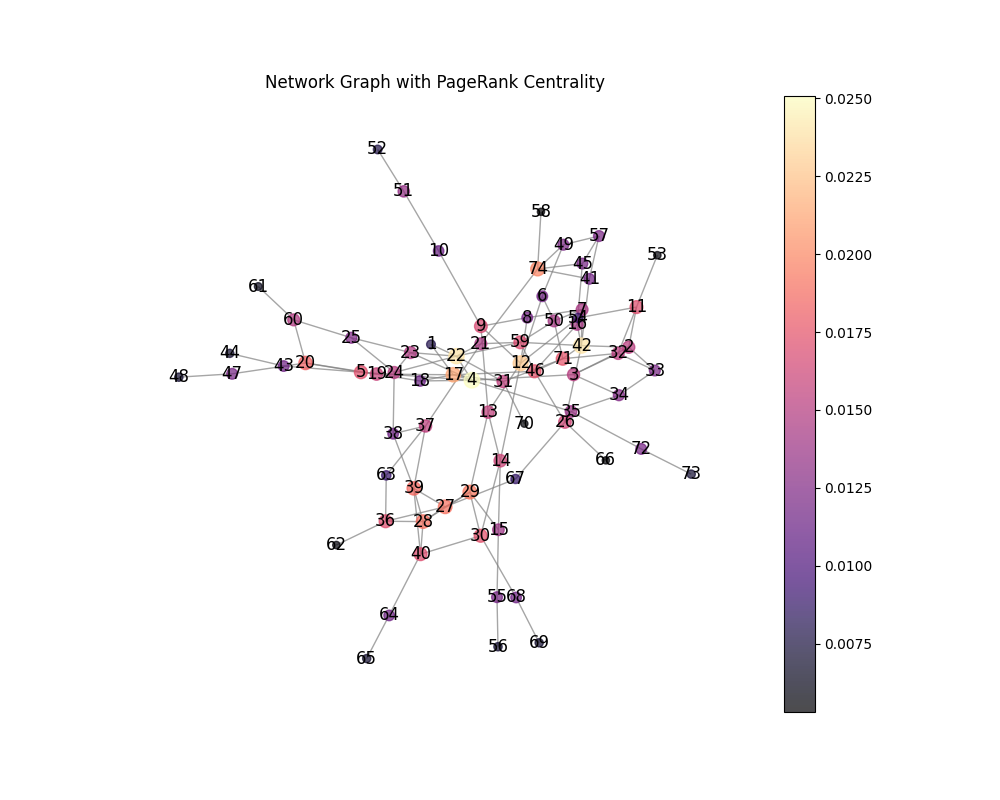
特征向量中心性

高特征向量中心性的节点不仅连接多个分子，而且与其他重要酶有强大的联系。这些酶通常在酶促反应中起到核心作用，影响其他重要酶的活动或反应速率。低特征向量中心性的节点与重要酶的联系较少，可能在网络中发挥较小的作用，或仅在某些特定的反应中起作用。



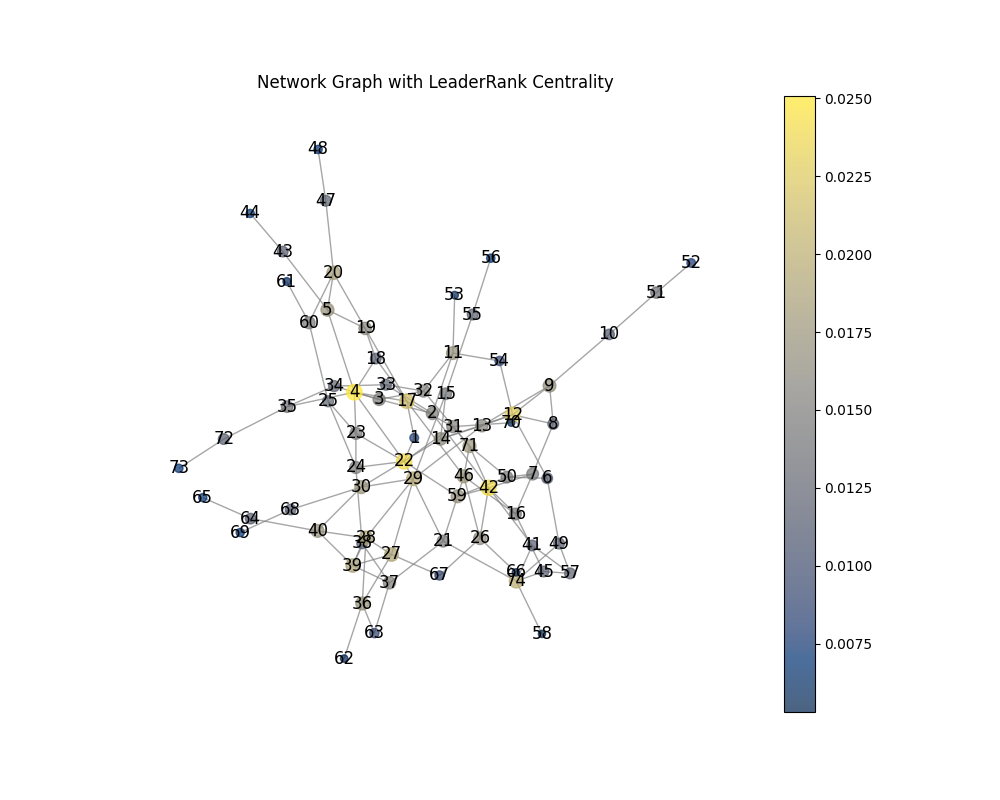
Katz中心性

高Katz中心性的节点能够通过直接和间接的连接影响其他多个酶或分子。它们在反应网络中扮演着重要角色，能通过长链反应或跨越多个路径的影响调节系统。低Katz中心性的节点影响较为局限，通常只与网络中的一小部分分子或反应途径相关，不具备广泛的调节能力。



PageRank算法

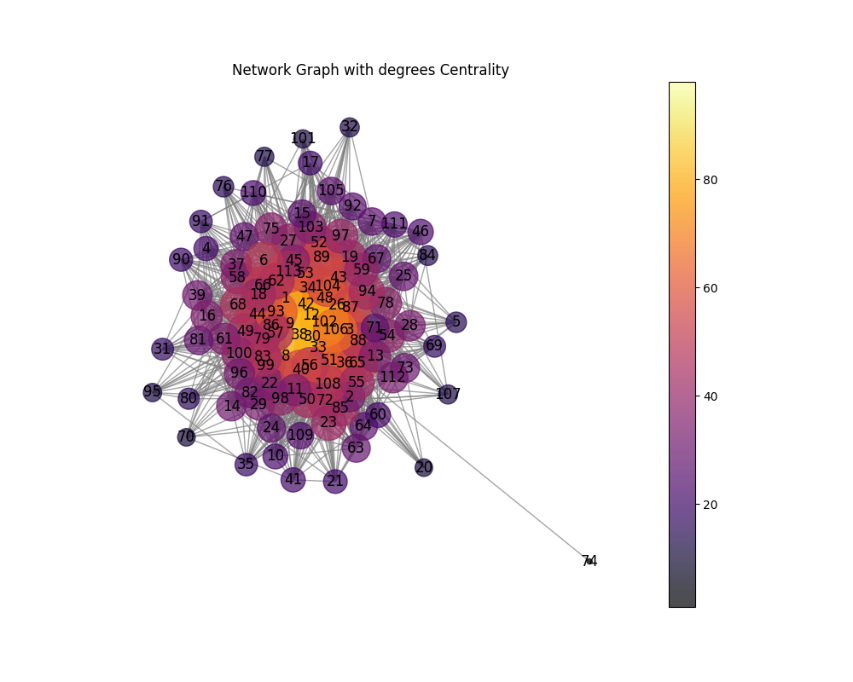
高PageRank的节点与多个重要节点有联系，并且这些联系也是高质量的，意味着这些酶在网络中有较高的“声望”。它们是网络中的关键酶，可能是反应途径的核心或控制点。低PageRank的节点联系较少且质量较低，它们对整个反应网络的影响力较弱，通常是网络中的边缘酶。



LeaderRank算法

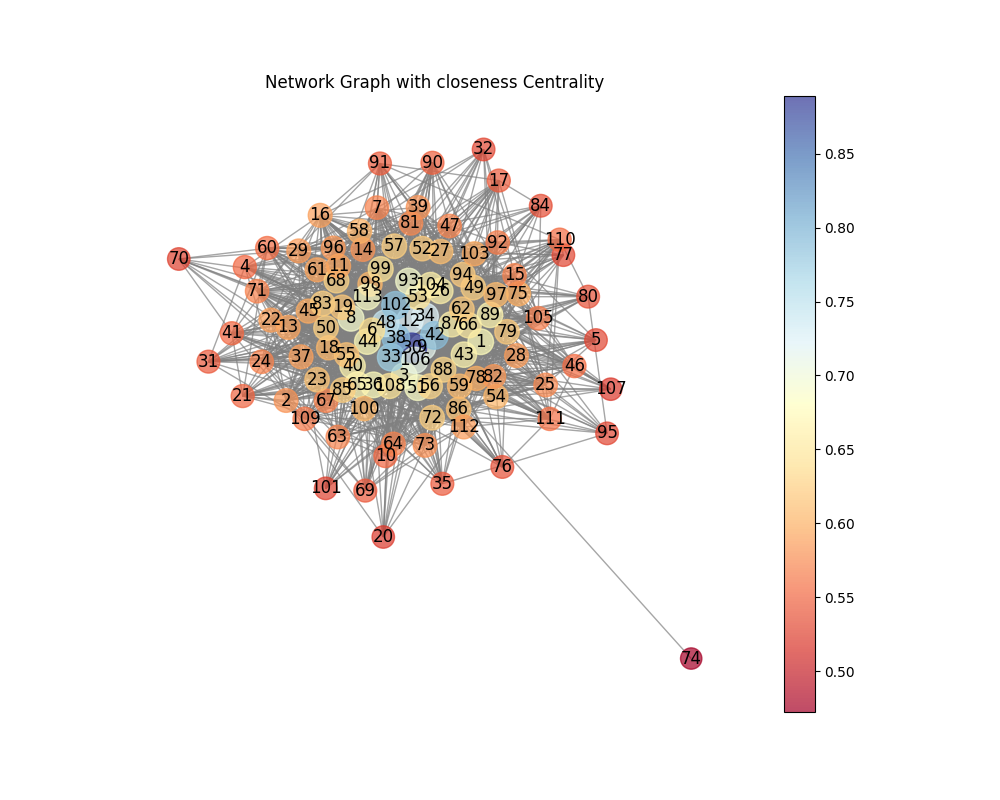
高LeaderRank的节点在信息或资源流动中起到了“领导”作用，能够有效地引导和调节其他酶或反应的进行。这些酶可能在网络中发挥着调控或协调作用，影响多个反应路径的进程。低LeaderRank的节点通常是跟随者，它们的影响力较小，通常在局部反应中起作用，缺乏在全局反应网络中的主导地位。





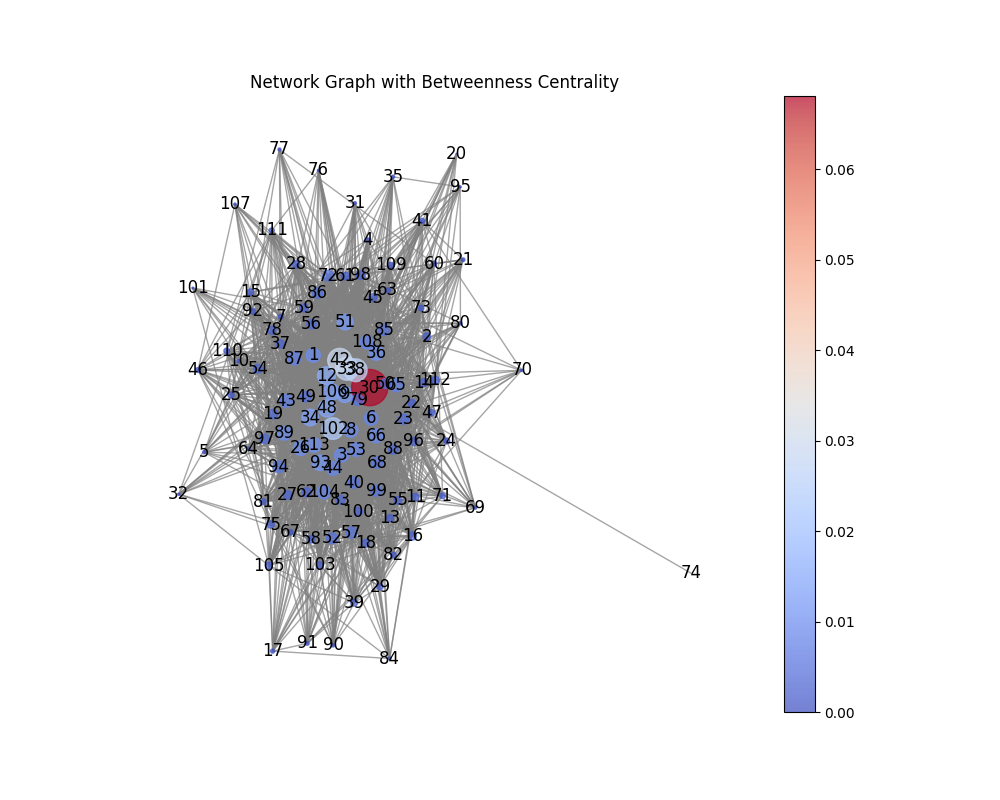
度中心性

高度中心性的节点有更多的直接连接，意味着这些个体与更多的其他个体接触或传播病原体，是传播的“超级传播者”。它们通常处于网络的核心，可能是疾病传播的关键节点。低度中心性的节点与少数个体接触，它们的传播潜力较小，对网络中的传播影响较为局限。



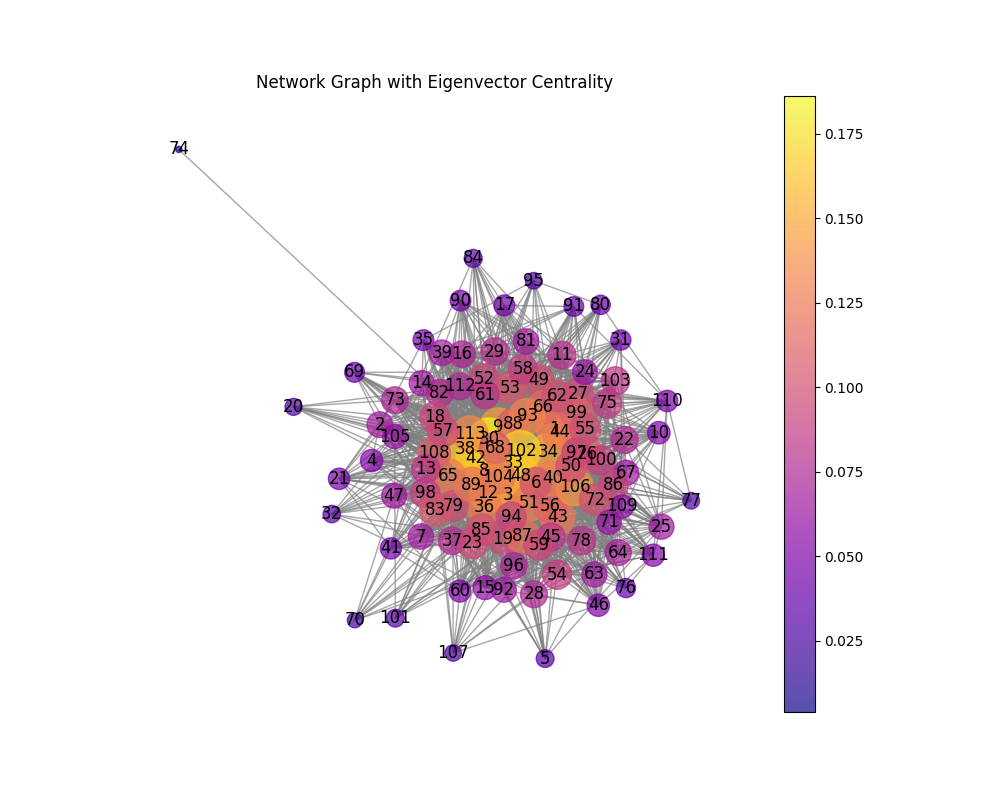
接近中心性

高接近中心性的节点与网络中的其他节点距离较短，能够在最短时间内接触到网络中的所有其他节点。它们在传染病的传播中具有“快速传播”能力，能迅速感染网络中的其他个体。低接近中心性的节点距离其他节点较远，意味着它们的传播速度较慢，在传染病传播中起到的作用相对较小。



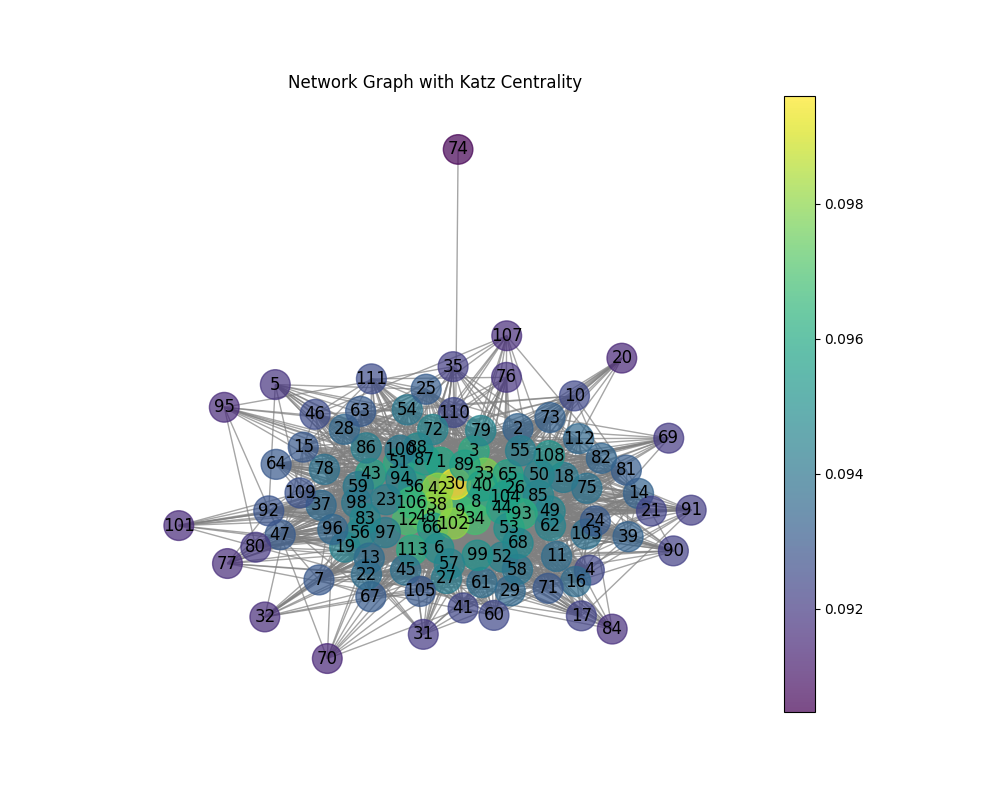
介数中心性

高介数中心性的节点处于多个传播路径的中介位置，控制着疾病传播的流向。这些节点是传播的“关键枢纽”，一旦这些节点被控制或隔离，可能会显著减缓整个网络中的传播速度。低介数中心性的节点在传播路径中作用较小，它们通常不充当传播的“桥梁”，因此对整体传播过程的控制力较弱。



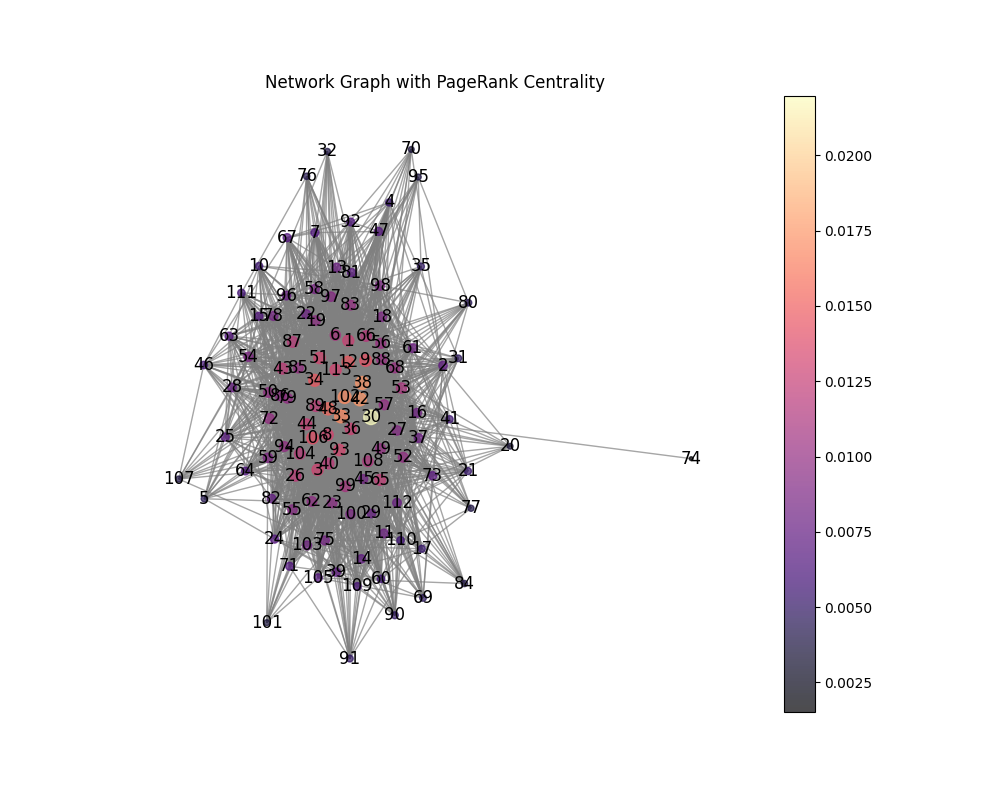
特征向量中心性

高特征向量中心性的节点与其他重要节点有紧密的联系，这些节点不仅直接接触许多个体，而且这些个体本身也具有较高的重要性。它们在疾病传播中具有较强的间接影响力，是“超级传播者”的关键。低特征向量中心性的节点即使直接连接一些个体，但这些个体的影响力较低，因此它们在传播中的作用相对较小。



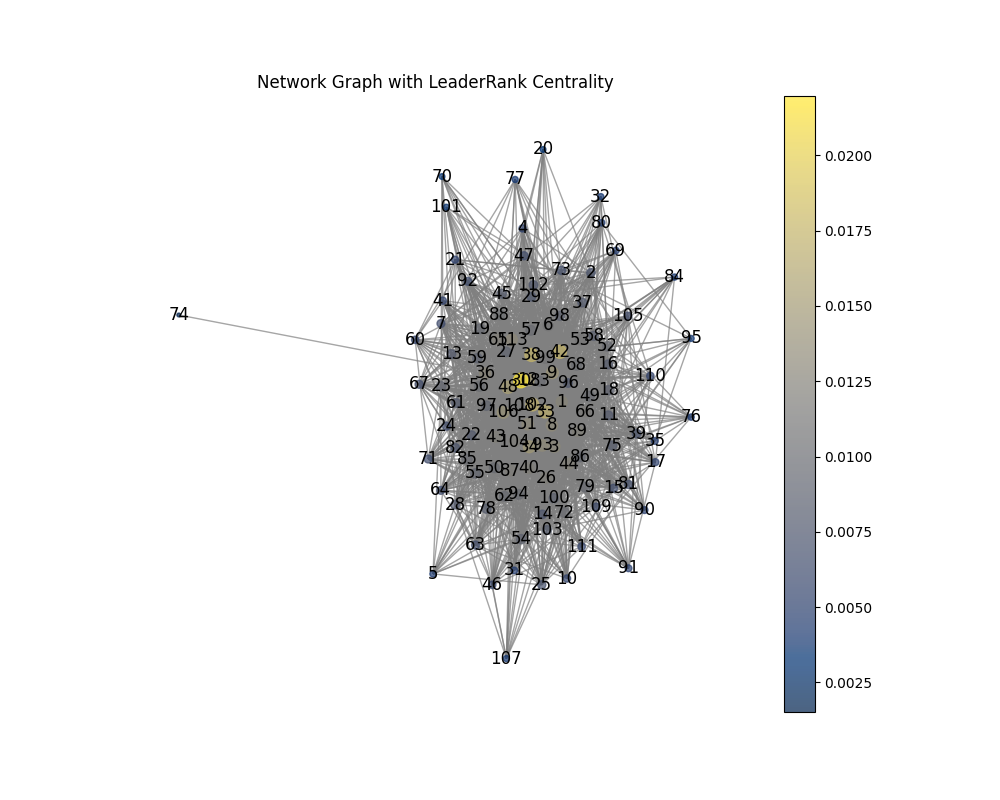
Katz中心性

高Katz中心性的节点通过直接和间接的连接影响着更多的个体。它们在传播过程中不仅能通过自己的直接接触传播疾病，还能通过较长链条的间接联系来扩散病原体。低Katz中心性的节点影响较为局限，通常只通过少数的直接连接进行传播，传播效果较弱。



PageRank算法

高PageRank的节点与多个重要节点连接，且这些连接在传播过程中起到了较高的“价值”。这些节点在传染病传播中具有较大的影响力，可能是传播的核心节点或“关键传播者”。低PageRank的节点连接较少，且连接的质量较低，它们对传染病的传播作用较小，通常是传播路径中的边缘节点。



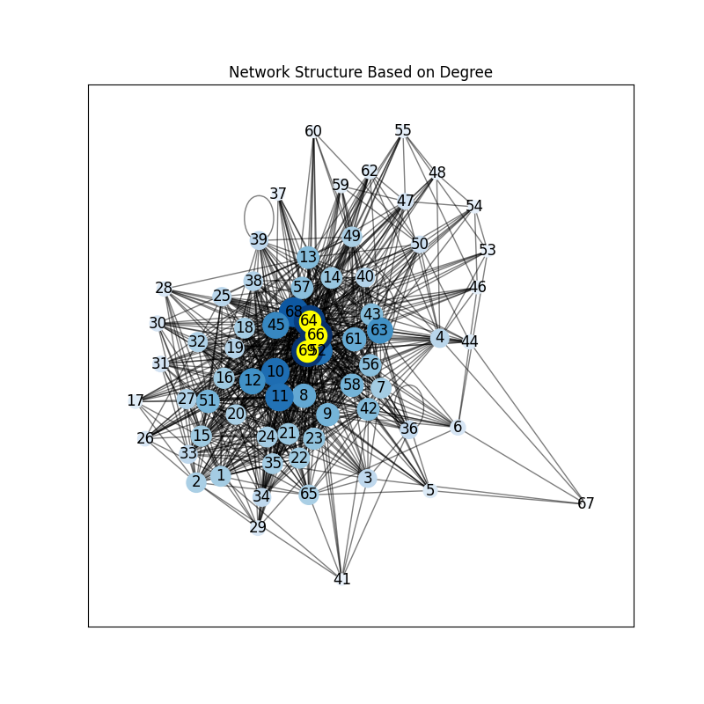
LeaderRank算法

高LeaderRank的节点在传播过程中起到引导和调节作用，能够在网络中引领疾病传播的方向。它们可能是关键的“传播引导者”，其位置使得它们能够控制传播的广度和速度。低LeaderRank的节点影响力较小，通常不作为传播的主导力量，更多依赖于其他高影响力节点的传播。

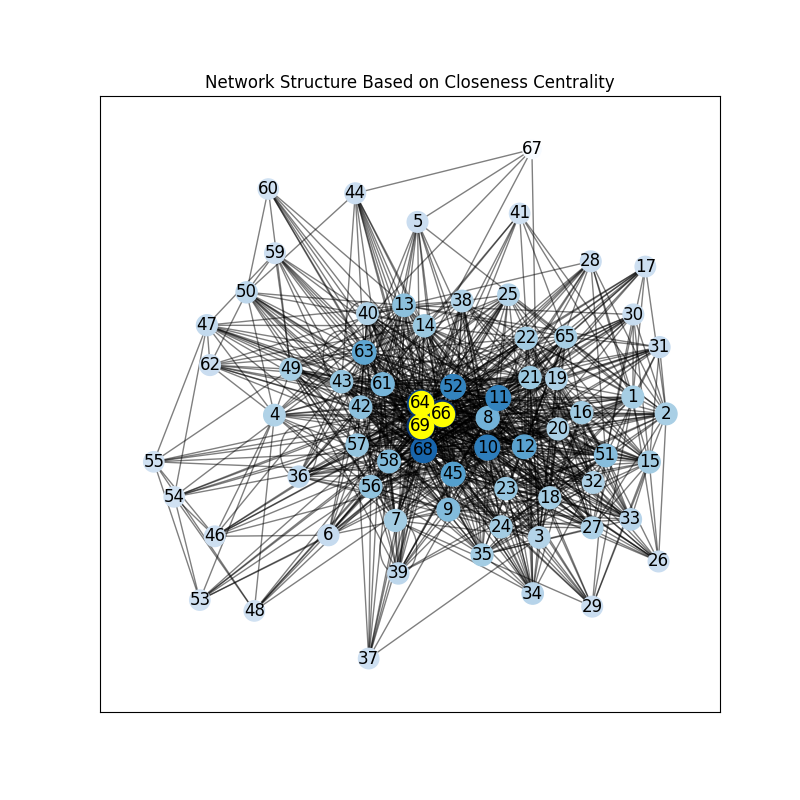
## 五、节点组重要性

以下为结合度算法、接近中心性算法、k-核算法与删后矩阵最小特征值算法，计算的网络中由3个节点组成的节点组的重要性，并进行可视化与对比分析，高亮显示的为该算法下最重要的由三个节点组成的节点组。

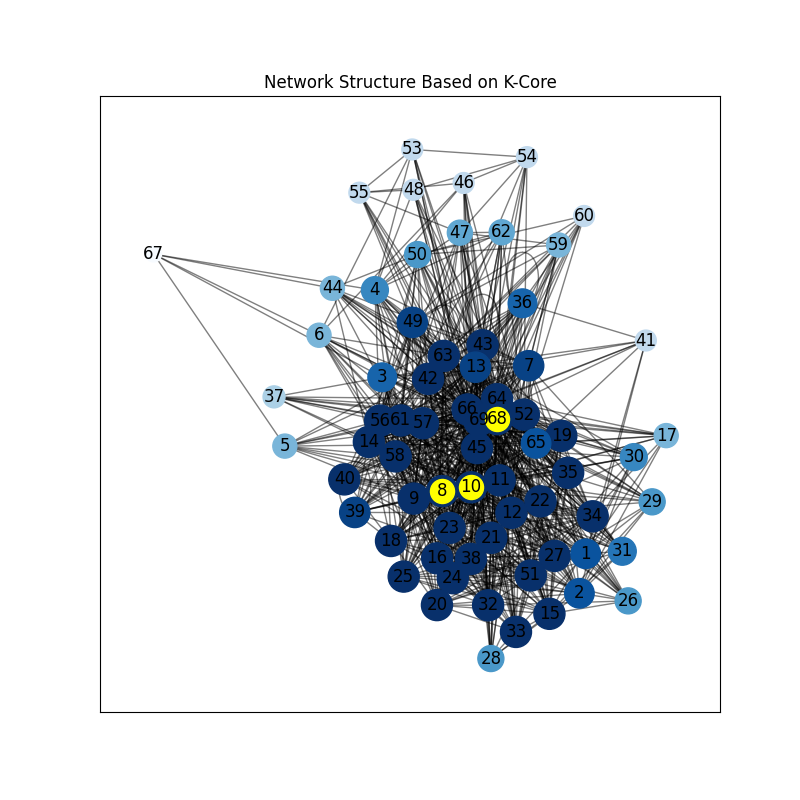




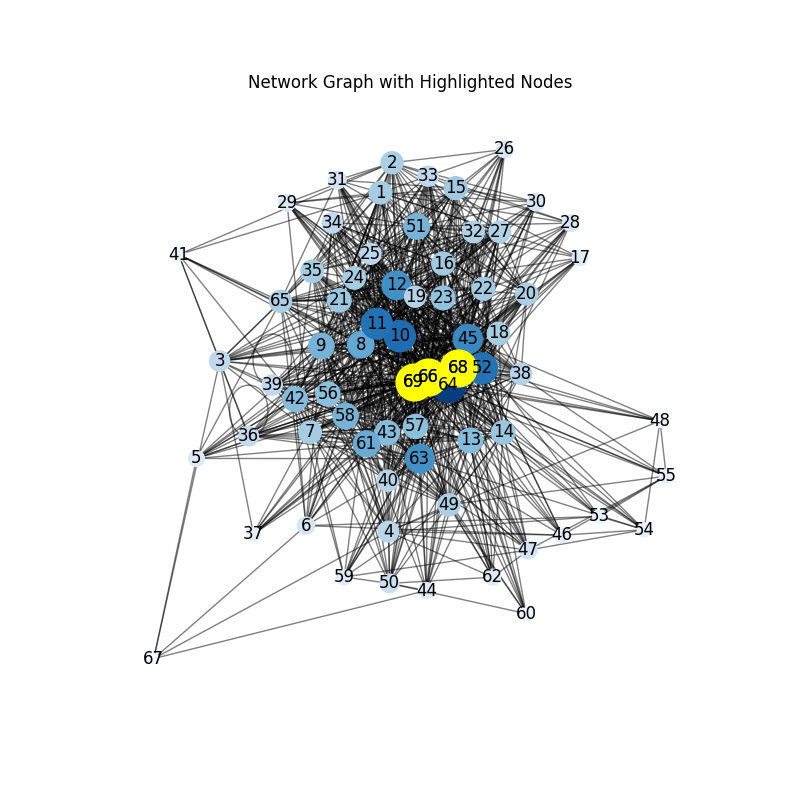
度算法



接近中心性算法

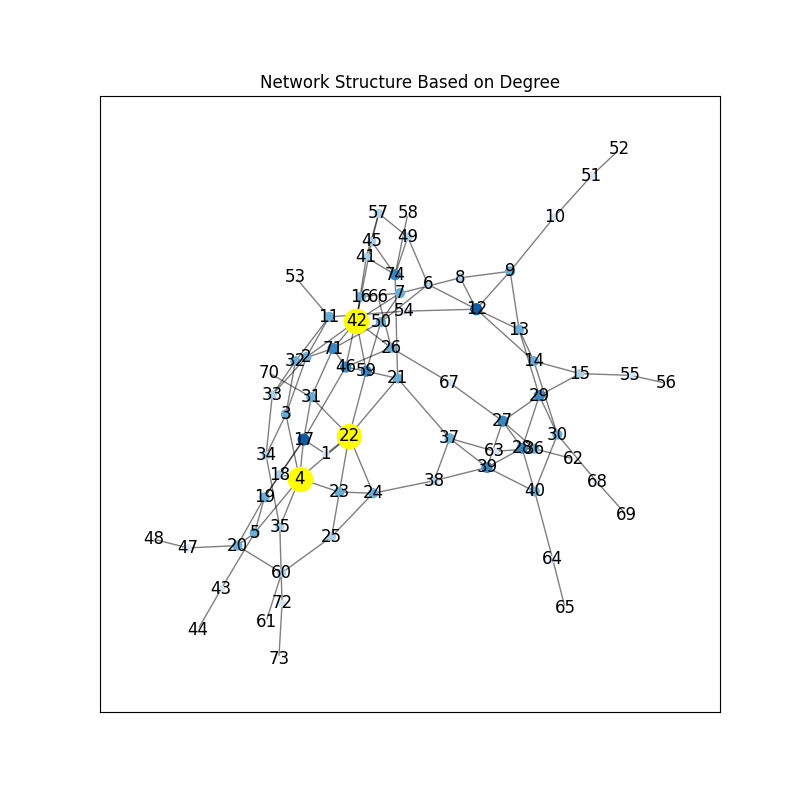


k-核算法

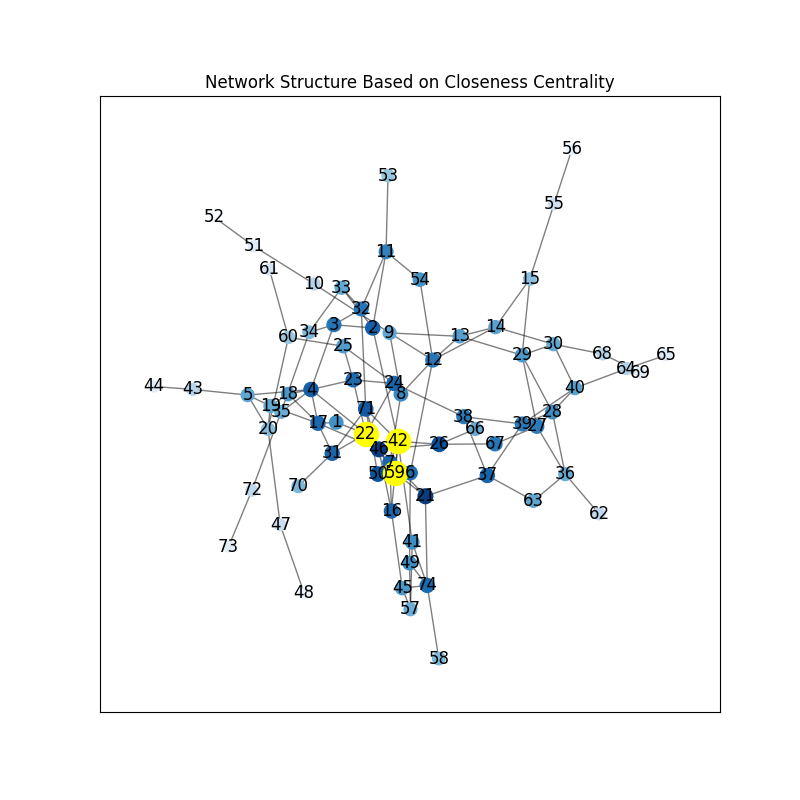


删后矩阵最小特征值算法

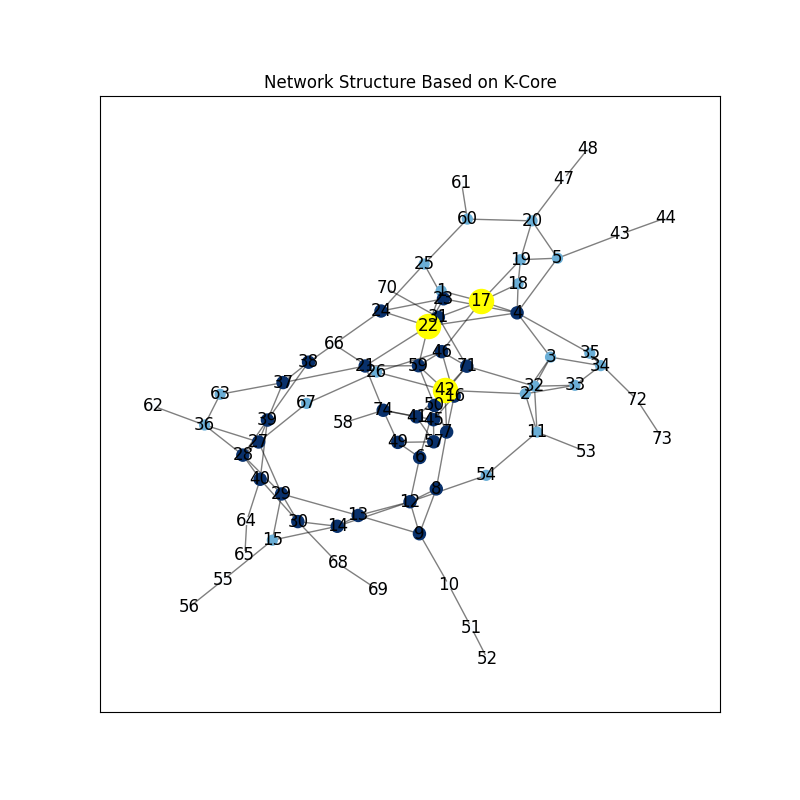




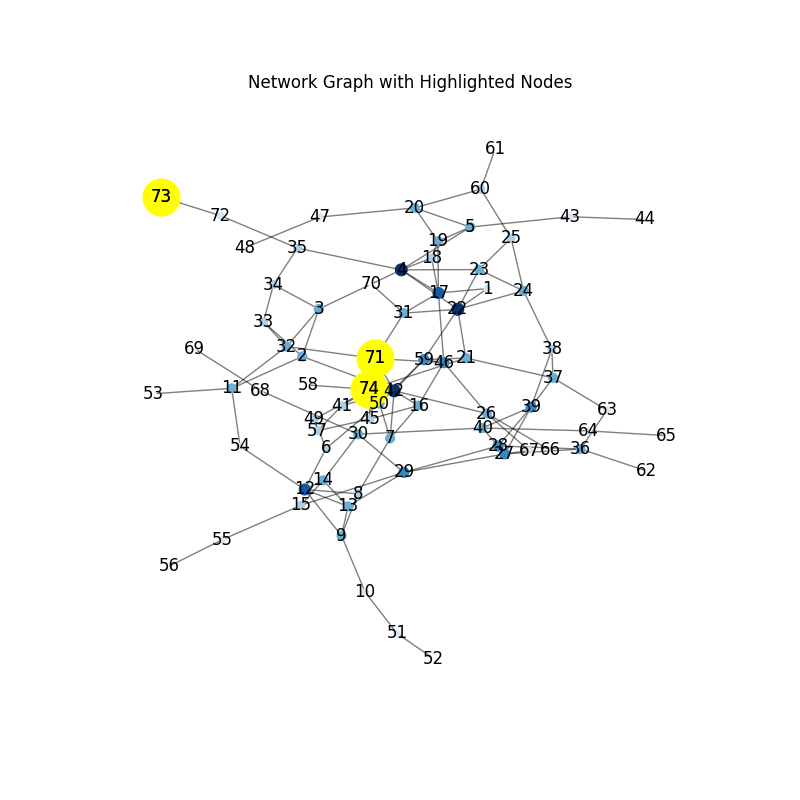
度算法



接近中心性算法

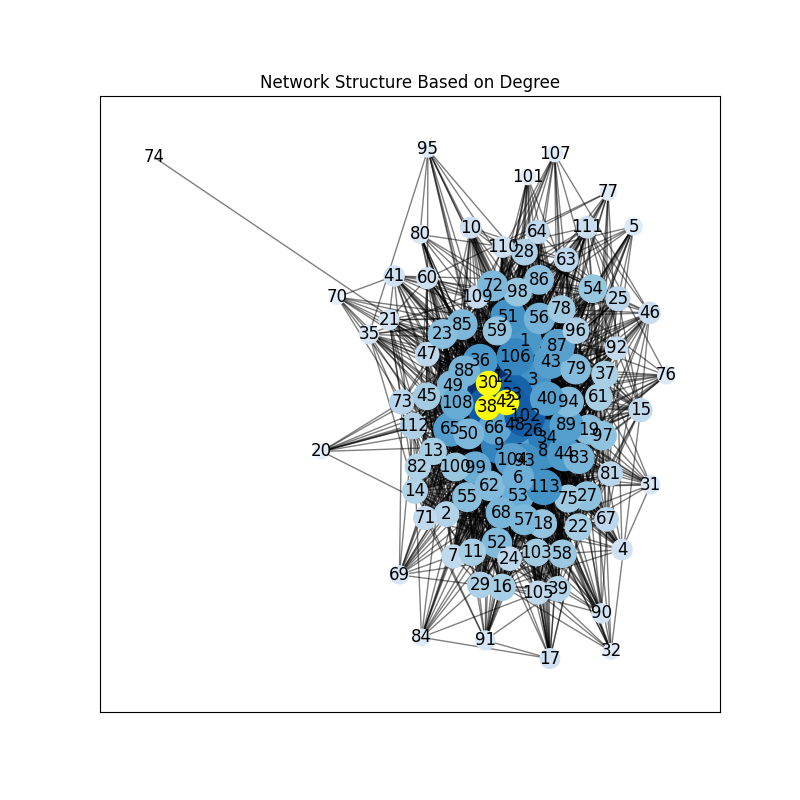


k-核算法

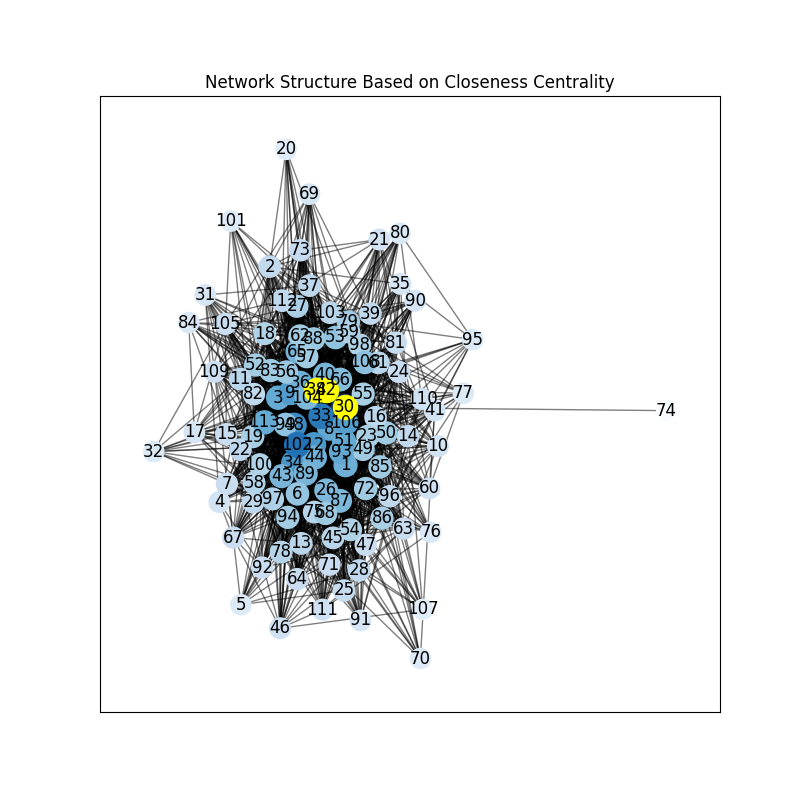


删后矩阵最小特征值算法

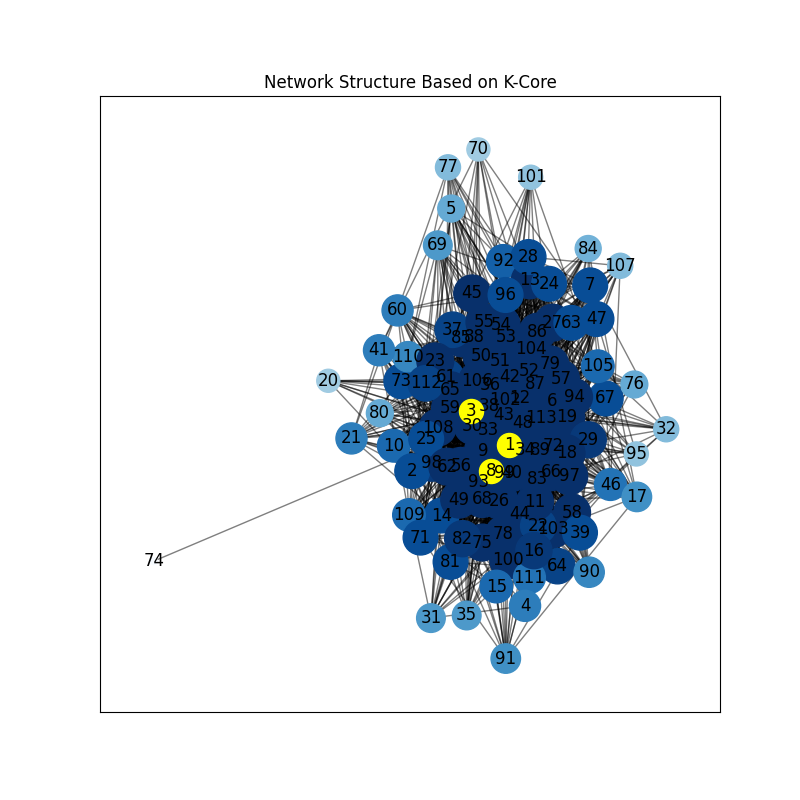




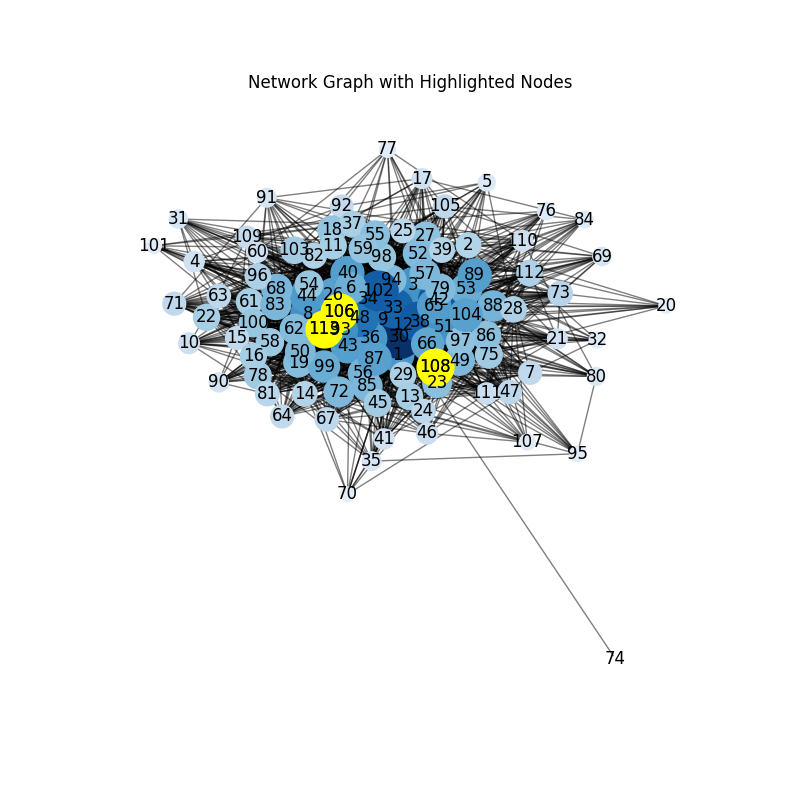
度算法



接近中心性算法



k-核算法



删后矩阵最小特征值算法

**列表进行数值分析：**

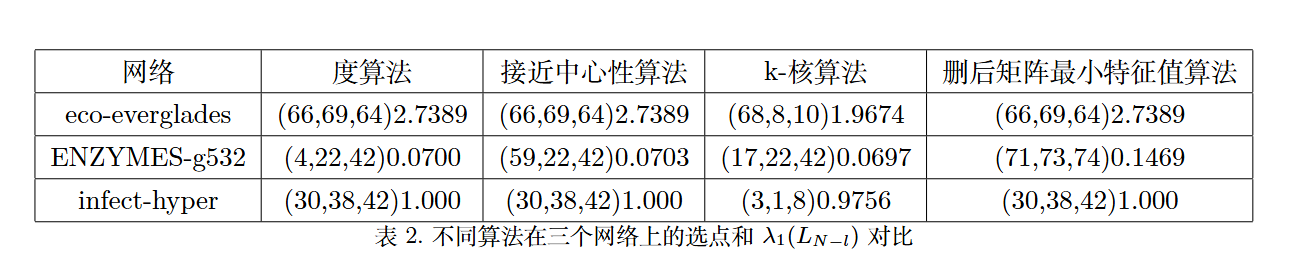


表2中，括号内的是依据算法选取的最重要的包含三个受控节点的节点组，括号外的是计算得到的删后矩阵最小特征值指标。从指标的数值可以看出来，删后矩阵最小特征值算法相比于其他算法更优，k-核算法得到的结果相对其他算法来说更差。

## 六、同步能力比较分析

分别假设网络的同步化区域为有界和无界两种类型，根据以下公式计算：

Laplace矩阵最大特征值与最小非零特征值之比：有界同步同步化区城R越小，网络同步能力越强。

Laplace矩阵最小非零特征值：无界同步化区域，越大，网络同步能力越强。



表格中的有界同步化区域计算的是，而无界同步化区域计算的是。

从表格中的结果可以看出来，当我们假设网络的同步区域都为有界时，eco-everglades网络的同步化能力最强，远超过另外两个网络，而ENZYMES-g532网络和infect-hyper网络的同步能力相似，前者比后者的同步能力更强一些。假设网络的同步区域都为无界时，eco-everglades网络的同步能力依然最强，远超另外两个网络，但此时ENZYMES-g532网络的同步能力相较infect-hyper网络更差一些。