## 面向聚类结构感知的网络表征学习算法评估方法

## 摘要 不同于传统的基于拓扑空间的网络分析方法，网络表征学习是近年来逐渐兴起的一种新的网络分析手段，通过保留网络拓扑结构、顶点信息和其他边信息将网络顶点嵌入低维向量空间，进一步解决面向网络的分析任务，如网络重构、链路预测和节点分类等。之前一些研究者针对不同网络表征学习算法，提出了相应的准则来衡量和对比它们的优劣，但是这些方法大都通过统计图表的形式来总结和分析结果，妨碍用户对网络表征学习算法的深入理解和有效探索。为解决这个问题，本文提出一种面向聚类结构感知的网络表征学习算法评估方法，从四个方面探索不同算法在聚类这一重要网络拓扑特征上的视觉表现：（1）表征学习前后聚类结果的一致性；（2）聚类内部节点在学习空间中的分布质量;（3）聚类节点之间在学习空间中的分布质量；（4）聚类中心点在学习空间中的分布质量。针对以上方面，我们分别提出相应的度量指标和可视化设计来分析和展示对比结果。进一步，我们根据所提指标从客观角度探索网络表征算法在具有不同结构特征的聚类上的表现力和适用性，并设计面向聚类结构感知的调查问卷，收集用户对网络表征学习算法的主观评价。最后通过主客观评价的综合对比，我们讨论网络不同表征学习算法的特点，并结合算法原理深入分析潜在原因。

关键词 网络表征学习；视觉表征；评估；聚类

## 1引言

有效的网络分析方法能为用户提供对数据背后知识更深入的了解。然而，大多数网络分析方法仍遭受高计算和空间成本的限制。因此，为了更高效的解决网络分析问题，网络表征学习为近年来网络分析领域的研究热点，它是一种解决网络分析问题的高效的学习范式， 通过保留网络拓扑结构、顶点信息和其他边信息将网络顶点嵌入低维向量空间，其中图形结构信息和图形属性得到最大限度地保留。一般说来，网络表征学习算法大致分为三类[1,2,3]：基于矩阵分解、基于随机游走、基于深度学习。其中基于矩阵分解[7,8,37 ]的核心思想是通过分解描述网络结构的矩阵得到网络节点的表示向量；基于随机游走[38,39]的本质是通过保留网络的局部结构性来估计节点的表示，其模型主要是将节点对应为NLP中的单词，通过随机游走得到的序列对应为NLP中的句子，进而通过skip-gram模型来学习网络节点表征; 基于深度学习[25,26]的有效性已经在计算机视觉以及语音处理领域得到广泛验证，它可以得到一个有效的非线性函数学习模型，非常适合用来拟合高度非线性的结构。

为了验证和对比网络表征学习算法的有效性，很多学者基于不同的角度对网络表征学习算法进行了评估。在低维向量空间中嵌入超大型信息网络在可视化、节点分类和链路预测等许多任务中都有一定的应用价值。并且由于真实信息的不可用性，不同NRL算法学习的顶点表示不能直接用来比较， 一些学者[23-34] 通常在网络重构、链接预测、聚类分析、顶点分类等几个方面用来对比网络表征学习算法。 但是这些传统的评估方法仍然是用文字、图表、统计的形式展示给用户，用户只能通过其中的数值来大致了解各个网络表征学习算法的优缺点，但是并不能详细探索其中的细节和具体表现。这些方法只是从不同的方面提出了一个度量的准则和方法，但是并没有给用户展示这个网络表征学习算法关于网络分析的一些内在机制、细节信息特征，妨碍用户对网络表征学习算法方法的深入理解。

聚类是网络结构最重要的体现，对聚类进行分析有助于明确拓扑结构，已有大量研究专家提出很多聚类的度量方法[18，19，22]，来评估聚类质量的好坏，但是在网络表征学习领域，如何对比和评估学习前后的聚类结构仍是一个开放问题，尚未得到深入研究。 因此为了帮助用户更好的更直观的感知、探索不同网络表征学习算法的内在机制，分析它们在网络聚类结构上的保持和表现，本文提出一种面向聚类结构感知的网络表征学习算法可视化评估方法，主要从四个方面探索不同算法在聚类这一重要网络拓扑结构指标上的视觉表现：（1）学习前后聚类结果的一致性；（2）聚类内部节点分布质量;（3）聚类节点之间的分布质量；（4）聚类中心点的分布质量。针对不同方面，我们分别提出相应的计算指标和可视化设计来量化和展示对比结果，从多个角度帮助用户探索各类网络表征学习算法方法的特点，同时我们使用多个数据集，深入分析不同网络表征学习算法在具有不同结构特征的聚类上的表现力，并结合算法原理深入分析潜在原因。

最后通过两个案例研究来测试我们的技术，并进行专家评审来收集反馈，来证实我们方法的有效性和实用性。总结来说，本文的主要贡献如下：

• 提出一套能够面向对比聚类结构感知网络表征学习算法的概念体系；

• 分别从客观、主观和两者结合上这三个方面对网络表征学习算法进行评估；

• 设计并实现一套支持用户探索式分析和对比聚类结构感知网络表征学习算法的可视化评估工具。

本文的组织结构如下：

第一部分为本文展开的背景，第二部分是本文的相关工作，第三部分是评估流程概述，第四部分为聚类结构特征度量，第五部分是可视化评估分析，并在第六部分做了案例分析（实验结果），最后是总结与展望。

## **2相关工作**

**2.1网络表征学习**

网络表征学习是一种解决网络分析问题的有效且高效的学习范式， 通过保留网络拓扑结构， 顶点内容和其他边信息将网络顶点嵌入低维向量空间。网络表征学习算法大概分为以下三类：

**矩阵分解** 基于矩阵分解的方法的核心思想是以矩阵的形式表示网络顶点之间的连接，并使用矩阵分解来获得嵌入。LLE[7] 是局部线性嵌入，一种无监督学习算法，用于计算高维输入的低维保邻域嵌入。 Belkin等学者[10] 提出了一种几何激励算法Laplacian Eigenmaps，用于构造从嵌入高维空间的低维流形中采样的数据的表示；[Ahmed](http://xueshu.baidu.com/s?wd=author:(Amr%20Ahmed)%20&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)等学者[36] 提出了一种基于流算法的因式分解技术Graph Factorization，它依赖于对一个网络进行分区，从而使相邻顶点的数目最小化，而不是使跨分区的边最小化； Cao[8] 提出了一种学习加权网络顶点表示的新模型GraRep; HOPE[37] 是一种新的图形嵌入算法，即高阶近似保持嵌入，它具有可扩展性，能够保持大尺度图形的高阶近似，并能够捕获非对称传递性。

**随机游走** 基于随机游走的模型主要是由word2vec启发而来，将节点对应为NLP中的单词，通过随机游走得到的序列对应为NLP中的句子，这些算法本质上是通过保留网络的局部结构性来估计节点的表示。Perozzi[5] 提出了一种学习网络顶点潜在表示的新方法DeepWalk， 但是在现实中， 网络顶点含有丰富的信息(如文本) ， 不能很好地应用于典型的表示学习方法的算法框架，所以Yang[12] 通过证明最先进的网络表示方法Deepwalk实际上等价于矩阵分解(MF) ， 提出了文本关联的Deepwalk(TADW)； struc2vec[38] 使用层次结构来度量不同尺度下的节点相似性，并构造多层网络来编码结构相似性并生成节点的结构上下文;Grover等学者[39] 提出了node2vec，一个学习网络节点连续特征表示的算法框架，将节点映射到特征的低维空间，从而最大限度地保留节点的网络邻居; Dong等学者[40] 开发了metapath2vec，将基于元路径的随机行走形式化，以构造节点的异类邻域，然后利用异类跳图模型执行节点嵌入； HARP[16] 是一种新的图形节点低维嵌入学习方法， 该方法保持了图形的高阶结构特征，通过在嵌入输入网络之前对其进行压缩， 有效地避免了复杂的嵌入配置（即局部极小值）。

**深度学习Deep learning** 基于深度学习的网络表征学习能够对非线性结构进行建模，因此可以使用深度自动编码器来降低数据的维数。为有效地捕获多层非线性网络结构并保持全局和局部结构的方法Wang[26] 提出了一种结构化的深度网络嵌入方法，即SDNE；Chang[13] 设计了一种网络数据的深度嵌入算法，利用一个高度非线性的多层嵌入函数来捕获网络中异构数据之间的复杂交互；GCN[41] 图卷积神经网络是一种新的深度学习架构，它能够利用包含在数据以及数据之间关系中的信息；DNGR[25]深层神经网络将随机冲浪与深度自动编码器结合在一起，支持模型的健壮性，并捕获更高阶的接近度; GraphSAGE[15] 利用节点特征信息学习一个函数，该函数通过从节点的本地邻域中抽样和聚合特性来生成嵌入。

**2.2网络表征学习算法评估**

针对网络表征学习算法的评估方法，一些学者[23-34] 通常在网络重构、链接预测、聚类分析、顶点分类几个方面用来做比较：

**2.2.1网络重构**

Feder[42] 定义了网络重构的概念，主要思想是利用网络的链接结构对节点和边进行分组。 多年来，许多研究人员使用基于聚集的方法[43-45]来重构网络， Navlakha等人[46]使用信息理论中的最小描述长度（MDL）[47]将图形汇总为图形汇总和边缘校正。并且通过大量的实验对比，在网络重构方面，尽管方法的性能依赖于数据集，但是保持高阶近似性的嵌入方法通常优于其他方法，其中SDNE在所有数据集上都表现良好、node2vec学习的嵌入具有较低的重建精度、HOPE学习了线性嵌入，但保留了高阶近似，在没有任何附加参数的情况下很好地重建了图形；

**2.2.2链路预测**

链接预测旨在预测网络中丢失的边，或者未来可能会出现的边。 在社交网络中，可以用于推荐可能的好友关系，达到更好的用户体验。链路预测普遍存在于生物网络分析中，验证节点间链路的存在需要昂贵的实验测试。将实验限制在存在可能性排序的链接上已经证明是非常经济有效的。Liben等人[55]、Lu等人[56]和Hasan等人[57]调查了这一领域的最新进展，并将算法分为：（a）基于相似性（局部和全局）[59、60、58]、（b）基于最大似然性[61、62]和（c）概率方法[63、64]。另外，嵌入可以明确或隐含的捕获网络固有动态，使得应用程序可以链路预测。 Wang等人[26]和Ou等人[29]预测了公共可用协作和社交网络上已知节点表示的链接。经大量研究显示，在这些数据集上，使用嵌入进行预测的链路比上述基于相似性的传统链路预测方法更准确，并且node2vec在BlogCatalog上取得了良好的性能， 但在其他数据集上却表现不佳，HOPE在所有数据集上都取得了良好的性能。另外，在PPI中，HOPE在更高维度上优于其他方法，而SDNE生成的嵌入在低维度上实现了更高的链路预测图。

**2.2.3聚类分析**

网络聚类（也就是网络分区）可以有两种类型：（a）基于结构的聚类和（b）基于属性的聚类。前者可以进一步分为两类，即基于社区的和结构上等价的集群。

基于结构的方法[50，51，52]旨在寻找聚类内边缘数量多、聚类间边缘数量少的密集子网络，相反，结构等价性集群[53]的设计目的是识别具有类似角色的节点（如异常值）； 另外White等人[54]在嵌入过程中使用k-means对节点进行聚类，并可视化WordNet和NCAA数据集上获得的聚类，验证获得的聚类具有直观的解释；为解决关系的建模和预测问题，Tu等[14] 提出了TransNet模型，利用平移机制来解决社会关系抽取问题。经过大量的实验表明，在社区分类方面，HOPE和SDNE生成的嵌入物保留了较高的接近度，很好地分离了社区，但LE、GF和LLE只有到达某种程度时才能捕捉到社区结构；

**2.2.4节点分类**

在进行网络数据的分析时，一个最常见的场景就是对网络中的节点进行合理的划分。使用标记的节点和网络中的链接推断缺少的标签，预测这些缺失标签的任务称为节点分类。Bhagat等人[64] 调查了文献中用于此任务的方法。他们将这些方法分为两类，即基于特征提取和基于随机行走。基于特征的模型[65，66，67]的方法根据其邻域和本地网络统计为节点生成特征，然后应用到如Logistic Regression[78] 和Naive Bayes[79] 这样的分类器来预测标签。基于随机行走的模型[70，71]通过随机行走传播标签。嵌入可以解释为基于网络结构自动提取的节点特征，因此属于第一类。 一些最近的研究[23，24，26，29，39]评估了嵌入各种信息网络（包括语言、社会、生物学和协作图）的预测能力。他们表明，嵌入可以高精度地预测丢失的标签，并且node2vec在节点分类任务上保留了节点之间的同源性和结构等效性，PPI中的节点可能在功能上相关，并与类似的节点相互作用。

最后Daokun学者[4] 总结了用于评估现有NRL算法学习表示质量的信息网络类型和网络分析任务并且得出顶点分类在无监督和半监督环境中最常用作为评估方法。

## 3 评估实验概述

**3.1实验数据集**

根据我们当前的实验要求，我们在研究中使用了三个不同类型和规模的数据集(表 1)：一个社交网络的Hamiltonc[72]，一个生物网络的Citeseer-M10[73]，一个引文网络的Protein-Protein Interaction[74]和一个模拟数据集。其中，本文选择的这三个数据集每种网络的所属领域和顶点规模（*V*）都是不同的，我们根据其边的规模由小到大排列的原则在所有数据集中选择。同时我们使用的这些网络所得到的聚类数量都是控制在五至十以内。在本文的试点研究中，我们选择的网络为无向网络，并别它们都是二进制网络，权重赋值为1。我们重点要探索和对比不同网络表征学习算法分别在不同的网络拓扑结构的数据集中所表现出来的学习结果。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Type** | ***V*** | ***E*** | ***Y*** | **ML** | **VA** |
| Hamiltonc | undirected, binary | 2,118 | 87,486 | 15 | YES | YES |
| Protein-Protein Interaction | undirected, binary | 4,777 | 184,812 | 40 | YES | NO |
| Citeseer-M10 | directed, binary | 10,310 | 77,218 | 10 | NO | YES |

表1

表1.三个数据集的网络性质，其中*V*为顶点数，*E*为边数，*Y*为标签数，ML表示多标签，VA为是否有附加属性。

**3.2网络表征实验算法**

因为数据集过于巨大，本文考虑到在现有技术上无法对所有数据都进行评估，所有本文在进行相关调研工作后选择了六种网络表征学习算法：LE[10]，LLE[7]， GF[36]，Deepwalk[5]，node2vec [39]和SDNE[26] 。重要的是，我们筛选的出的这些网络表征学习算法，涵盖了网络表征学习领域的三个主要网络表征学习算法种类：

**基于矩阵分解** 基于矩阵分解的学习方法是最早的也是最经典的网络表征学习算法，我们选取了LE，LLE和GF这三个最常用的。

Laplacian Eigenmaps目的是在权重wij较高时保持两个节点的嵌入紧密。它的最小化目标函数为公式：



其中，L是网络G的拉普拉斯函数，为了消除一般解，目标函数受约束，其中*YTDY=I*。

LLE假设每个节点都是嵌入空间中相邻节点的线性组合，计算高维输入的低维保邻域嵌入。

GF是第一种获得嵌入O(|E|)的学习方法，它通过对网络的邻接矩阵进行因式分解，来最小化目标函数：



**基于随机游走** 在基于随机游走方面，Deepwalk和node2vec是最经典的两个方法。Deepwalk能学习网络顶点潜在表示，它通过最大化观测以vi为中心的随机游动中最后k个节点和下一个k个节点的概率，保持节点之间的高阶邻近性最大化对数如下，其中2k+1是随机游动的长度。



node2vec可以最大限度地保留节点的网络邻居，它通过最大化固定长度随机游动中后续节点出现的概率来保持节点之间更高的近似度并采用了有偏差的随机游动，能够在在广度优先（BFS）和深度优先（DFS）网络搜索之间进行权衡，其产生的嵌入信息比DeepWalk更丰富。

**基于深度学习** SDNE是基于深度学习的网络表征学习算法中最优的一个，它通过高度非线性函数来获得嵌入，它由无监督和监督两部分组成，前者寻找一个可以重构其邻域的节点的嵌入，后者则监督相似顶点在嵌入空间是否远离，这个模型能够有效地捕获多层非线性网络结构并保持全局和局部结构的方法。

**3.3评估流程概述**

网络表征学习是近年来逐渐兴起的一种新的网络分析手段。之前一些研究者针对不同网络学习方法，提出了相应的准则来衡量和对比它们的优劣，但是这些方法大都通过统计图表的形式来总结和分析结果，妨碍用户对网络表征学习的深入理解和有效探索。在另一方面，社区发现是网络结构最重要的体现，对聚类进行分析有助于明确拓扑结构，已有大量研究专家提出很多聚类的度量方法[18，19，22]，来评估聚类质量的好坏，但是在网络表征学习领域，如何对比和评估学习前后的聚类结构仍是一个开放问题，尚未得到深入研究。为了解决这些问题，本文设计了一种面向聚类结构感知的网络表征学习算法可视分析系统。其中本文提出了四个在聚类这一重要网络拓扑结构指标上的度量指标：学习前后聚类结果的一致性；聚类内部节点分布质量;聚类节点之间的分布质量；聚类中心点的分布质量。针对这四个度量指标，我们设计可视化视图来探索网络表征方法在具有不同结构特征的聚类上的表现力和适用性，从多角度帮助用户探索网络表征学习算法的特点。并结合原理深入分析潜在原因。最后通过案例分析和专家反馈来验证本文所提方法的有效性和实用性。

本文主要关注各类网络表征学习算法在不同网络拓扑结构方面在聚类这一指标上的适应性。如图2流程图所示，我们选择涵盖了网络表征学习算法三个大类的六种网络表征学习算法，并设计了一个客观系统，通过可视视图来客观表示不同网络表征学习算法在学习前后聚类结果的一致性、聚类内部节点分布质量、聚类节点之间的分布质量和聚类中心点的分布质量上面的适应性和优劣性。但由于网络表征学习内部极其复杂且结果具有随机性，所以我们还设计了一个主观问卷，选择志愿者对我们的主观问卷进行回答，并记录结果，来收集志愿者的主观经验，并设计了一个精密、完整的实验过程以确保实验的正确性。然后我们将用户在问卷系统上做出的选择进行统计和指标设定。最后，我们将主观问卷指标结果和记录的客观系统指标进行对比分析，选择最适合社区发现的学习方法。。

图1

本文设计了三个视图模块和一个参数调整模块组成我们的可视分析系统。其中系统的第一个模块为原始网络拓扑结构模块，它通过颜色映射展示原始结构的力导向图、散点图、社区发现图；其二为聚类质量测试模块，包含桑基图和径向坐标图，桑基图可以衡量学习前后不同学习方法同原始网络结构的聚类一致性，而径向坐标图能够展示密度和中心值变化，即可以展示六个不同的学习方法在一个社区的差异，以对比学习方法学习前后聚类的质量；其三是六个网络表征学习算法模块，我们通过散点图、荷包图、聚类图和热力图展示六个不同的学习方法的结果；最后一个就是参数调整模块，这个模块可以让用户交互的选择和改变聚类的数量、社区之间的散点分布相似度最大值和最小值等参数。

系统界面如图3所示。

图2

## **4 聚类结构特征度量**

对于评价聚类结果的好坏，学者们在不同的应用中,为了选择最恰当的聚类方法,会从不同的角度提出聚类评价指标。网络聚类质量评价体系是评价聚类算法聚类结果好坏的重要指标。通过评价我们不仅可以度量聚类算法的有效性,而且可以很好地了解给定网络的动态关系。在聚类结构特征方面，我们通过学习前后聚类结果的一致性、学习后聚类内部节点分布质量、学习后聚类之间节点分布质量和聚类中心点的分部质量这四个度量指标对多种网络表征学习算法进行评估。利用模块度聚类写到第一段

在对原始网络进行聚类时，本文选择用基于模块度的聚类方法，得到原始聚类视图，再通过力导向图和散点图直观展示原始网络拓扑结构聚类后的节点分布

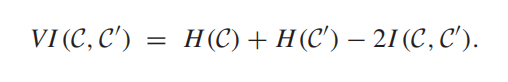


*i*表示社区，*Tre*表示迹，每一个社区内部的边数 ||||，*ai*表示社区*i*内部的点所关联的所有的边的数目与总边数的比例。

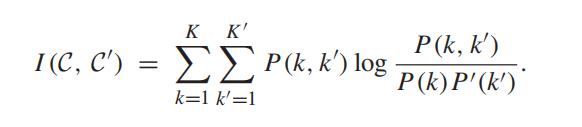
在对网络表征实验算法进行聚类时，我们选择用k-means聚类方法。K-Means是聚类算法中的最常用的一种，通过k-means算法将选定的网络表征实验算法进行社区发现。我们将数据集设置为*k*个分组，即得到*k*个社区，接下来从数据集中随机选择k个数据点作为初始社区中心点，对集合中每一个节点，计算与每一个初始社区中心点的距离，离哪个初始社区中心点，就将其划分到那个社区，当第一轮的所有节点都确认完再通过（平均）算法选出新的社区中心点。如果新社区中心点和上一轮的社区中心点之间的距离少于设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定），认为我们进行的聚类已经达到期望的结果，算法终止。

**4.1学习前后聚类结果的一致性写直接用互信息和信息熵，加公式**

为了保证学习前后对比聚类结果的客观性，我们将实验算法中*k*的数值与原始网络聚类的数量保持一致，即模块度聚类出的簇数量为6，实验算法后的网络进行社区发现时设定的*k*值也要为6。通过对网络表征实验算法学习前后的聚类，得到7个聚类后的结果，根据其聚类结果将它们的拓扑结构二维投影出来，观察其投影结果，将学习后的网络表征实验算法结果依次跟原始聚类结果进行交集对比。我们认为两者聚类结果越相似的越好，而不是聚类更优的网络表征实验算法更好。其中，我们使用两个聚类C的数量作为比较聚类结果的相似性标准。



VI（C，C’）表示两个进行比较的聚类之间的一致性的量化。H（C’）来度量其C’簇的不确定性。假设现在我们现在知道这个点属于C中的哪个簇。I（C，C’）指的是这种不确定性的减少，在所有点上的平均值。我们定义了I（C，C’）聚类之间的相互信息，C，C’等于相关的随机变量之间的相互信息:



H（c）的公式，没有个c和c‘的意义，k也介绍一下

C表示学习前的，C‘是学习后的，并引用参考文献，再写

**4.2学习后聚类内部节点分布质量**

聚类系数*Clu*是一个聚类质量好坏的重要评定指标，是我们聚类内部节点分布质量的评定指标。类的内部分散程度，通过对聚类后获得的簇进行密度计算，即将簇的面积除以簇内节点的数目，我们将这个数值表示为类内的分散程度的量化。得到的密度值越大我们则认为这个网络表征实验算法在这个网络图中学习后聚类的越好，即聚类系数越高，我们认为这个实验算法更好，适应性更强。其中聚类系数的计算公式如：



*e*指的是这个社区中所包含的所有点的面积，表示这个社区中选择节点的数量。，

这里用Pearson相关系数将聚类系数与点的分布情况进行相关性计算。其中点的分散情况我们用节点数量除以面积n/S。这里得到的P值越接近1越说明这个算法学习效果好。

P=

**4.3类节点之间的分布质量**

类间紧密性主要是指聚类后获得的类之间是否距离更近。我们将原始的聚类结果中获得的类进行聚类之间两两进行举例计算，其中距离越小，紧密性越好。类与类之间的距离，我们通过他们的社区中心点进行计算，两个类的社区中心点之间的距离就代表着两个类之间的距离。同时，我们认为进行比较的两个类，类间链接的线越多越紧密，就表示两个类的节点之间分部质量越好。其中节点之间的线的链接我们通过视图直接展示，由用户主观评定实验算法在类间节点的分布上的适应性。

学习之前的每两个类之间的连线数量*L*，学习后重心间距离作为类间距离d计算（加重心计算公式），一般来说L越大，d越小。Pear系数把两个维度L与d进行关联计算，这个应该是越接近-1越好

P=

**4.4聚类中心点的分布质量**

在通过对网络表征实验算法进行K-means聚类的过程中对网络迭代求解，其中通过设定K的数值得到规定数额的初始社区中心点。然后计算每个节点与各个初始社区中心之间的距离，把每个节点分配给距离它最近的社区中心。社区中心以及分配给它们的节点就代表一个社区。每分配一个样本，社区的社区中心会根据社区中现有的对象被重新计算。这个过程将不断重复直到满足没有节点被重新分配给不同的社区。最后得到的社区中心点的位置是否依然也在聚类的中心是我们评价网络表征实验算法在社区划分上的又一指标。在本实验中，我们认为社区中心点的位置越处于中间越好，其中我们将社区的重心点和社区中心点位置进行距离比较，两者距离越相近越符合。

我们将六个网络表征学习算法方法有六个盒须图，

重心点到盒包围边点的所有距离的平均值用社区的半径R表示

## 5 可视化评估系统设计

根据任务和设计目标, 本文提出并实现了一个基于网络表征学习算法的视觉表征评估的系统, 包含控制面板、网络结构视图、学习投影视图和聚类质量评估视图和调查问卷视图四个模块。 另外本文也添加用户调查

**5.1 控制面板**

加载数据，并显示数据的详细内容：数据名称、数据背景、点数、边数、聚类数

**5.2网络结构视图**

使用力引导算法构造原始网络拓扑结构的布局，节点和线条都使用黑色，可以选择某个聚类进行高亮，其聚类中心点使用红色进行显示。

我们一共使用三个不同大小规模的网络数据，系统的左下角视图如图2a展示当前选择数据规模的原始散点图和力引导图，其中数据的规模在系统的参数调整模块（如图2b所示）中选择。原始网络拓扑结构模块主要包含原始网络散点图和原始力引导图，如图3所示是分别在大中小三个规模中的网络拓扑结构展示图：

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

图3

根据每个规模的数据集，我们对六个不同的算法进行，如图4是根据模拟数据对六个网络表征学习算法进行

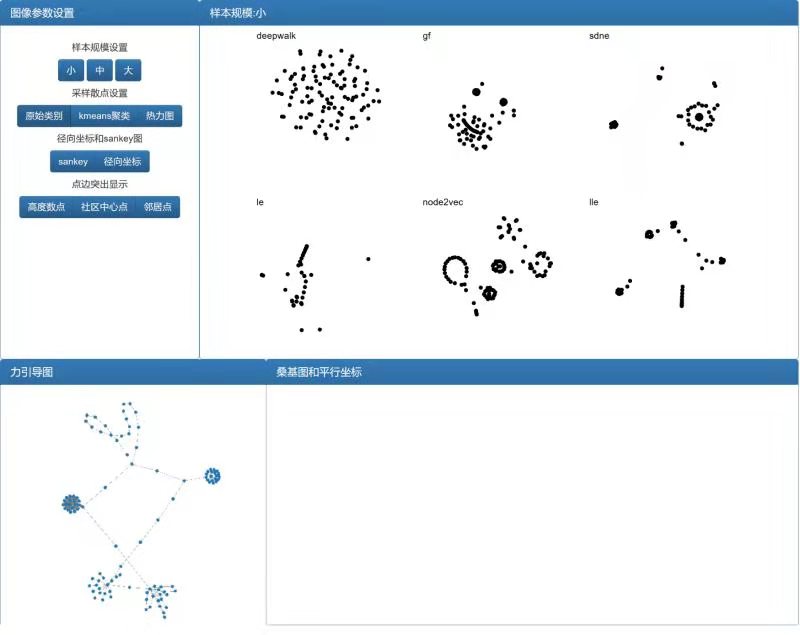
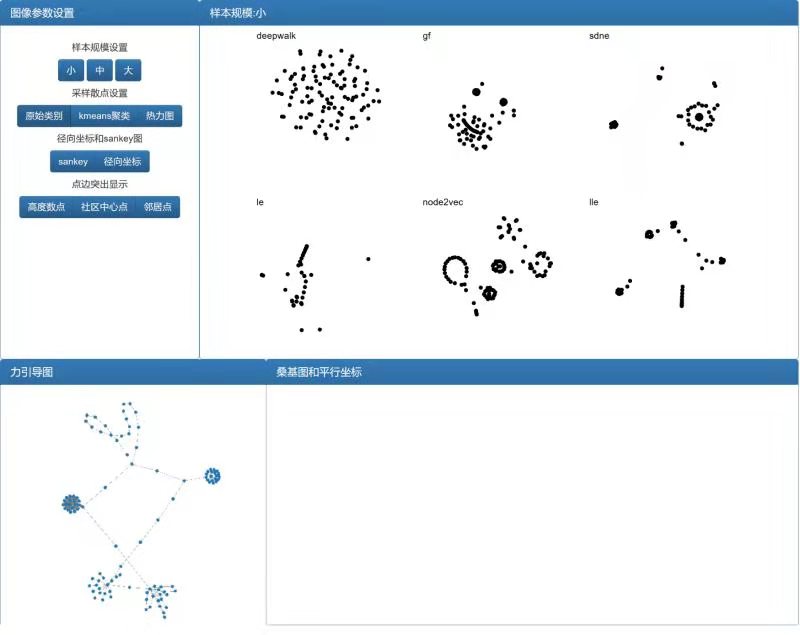


图 5

**5.3学习投影视图**

这个模块主要是展示实验中选择的六个网络表征学习算法在学习前后针对聚类结构感知变化对比。主要包括如图5所示是六个网络表征学习算法在学习后展现社区划分的力引导图

|  |  |  |
| --- | --- | --- |
|  |  |  |
| LE | LLE | GF |
|  |  |  |
| Deepwalk | node2vec | SDNE |

图5

**5.4聚类质量评估视图**

4个选项框，对应4个聚类指标，点击某个选项框，展示相应的聚类质量评估可视化结果

1. 学习前后的一致性：桑基图
2. 类内：环状平行坐标
3. 类间：环状平行坐标
4. 聚类中心点：盒须图

这部分主要是关于六个网络表征学习算法在学习前后社区划分的变化和学习后聚类的质量测试对比展示，包含平行坐标图和雷达图。如图\*所示，是使用Protein-Protein Interaction数据集进行社区划分时的径向坐标图和桑基图，其中平行坐标图中横轴表示为。。。

|  |  |
| --- | --- |
|  |  |
| 径向坐标图 | 桑基图 |

图6

图\*是一个sankey图，以deepwalk为原始模板 计算聚类相似度 然后从左到右依次排开 带颜色的块 即是每个类的代表 连线代表了不同方法过程中某一类的点到另外一种方法变成另外某一类或者没变

本文根据相似性的高低将每个投影后的聚类情况通过桑基图展示。本文将不同学习方法的聚类与原始网络结构进行比较，第一排是原始聚类的，后面依次是不同学习方法的聚类效果，其中，在一排里一个方块表示一个类，每个类的高度表示聚类的数量，本文把根据相似性好坏对所有，排的越在后面的聚类效果越混乱。目的：衡量学习前后不同学习方法同原始网络结构的一致性。

**5.5调查问卷视图**

为了调查用户对系统中六种网络表征学习算法的各类表现进行主观评价，我们分别对4种聚类指标提出了一些问题来获取用户的意见，在系统最右侧为调查问卷视图，如图2e所示。

## 6 案例分析---实验结果

**6.1客观结果对比**

一阶邻居二阶邻居（如何表示，点击高亮）

**6.2场景分析---用户实验**

**6.3主观评价对比（专家反馈）**

针对于某某类网络结构更适合某某类网络表征实验算法收集主观的用户分析

**6.4不足之处讨论**

## 7 总结与展望

本文提出了一种基于网络表征学习算法的视觉表征评估的系统，通过对多个网络表征学习算法在关键点、最短路径和社区三个网络结构方面的变化对比对网络表征学习算法进行多维度的评估。同时设计了一个客观网络表征学习算法可视评估系统，通过可视视图来客观表示不同网络表征学习在关键点、关键路径和社区发现上面的适应性和优劣性, 并进行了实验验证。结果表明 算法在学习前后度中心性和社区中心点等关键点的一致性较好，同时也在最短路径位置变化上 。 算法在学习前后聚类结构的保持和分布情况较为稳定

缺点和不足

[1] Hamilton W L , Ying R , Leskovec J . Representation Learning on Graphs: Methods and Applications[J]. 2017.

1. Goyal P , Ferrara E . Graph Embedding Techniques, Applications, and Performance: A Survey[J]. Knowledge-Based Systems, 2017.
2. Cai H Y , Zheng V W, Chang K . A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2018:1-1.
3. Daokun Z, Jie Y, Xingquan Z, et al. Network Representation Learning: A Survey[J]. IEEE Transactions on Big Data, 2018:1-1.
4. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations [J]. 2014.
5. Wang Y , Shen Q , Archambault D , et al. AmbiguityVis: Visualization of Ambiguity in Graph Layouts[J]. IEEE Transactions on Visualization & Computer Graphics, 2016, 22(1):359-68.
6. S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
7. Cao S , Lu W , Xu Q . GraRep: Learning Graph Representations with Global Structural Information[C]// Acm International on Conference on Information & Knowledge Management. ACM, 2015.
8. Tang J , Qu M , Wang M , et al. LINE: Large-scale information network embedding[J]. 24th International Conference on World Wide Web, WWW 2015, 2015.
9. M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: NIPS, Vol. 14, 2001, pp. 585–591.
10. A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pages 59–66. IEEE, 2009
11. Yang C , Liu Z , Zhao D , et al. Network Representation Learning with Rich Text Information[C]// International Conference on Artificial Intelligence. AAAI Press, 2015.
12. Chang S , Han W , Tang J , et al. Heterogeneous Network Embedding via Deep Architectures[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2015.
13. Tu C , Zhang Z , Liu Z , et al. TransNet: Translation-Based Network Representation Learning for Social Relation Extraction[C]// Twenty-Sixth International Joint Conference on Artificial Intelligence. AAAI Press, 2017.
14. Hamilton W L , Ying R , Leskovec J . Inductive Representation Learning on Large Graphs[J]. 2017.
15. Chen H , Perozzi B , Hu Y , et al. HARP: Hierarchical Representation Learning for Networks[J]. 2017.
16. Cunchao T U , Yang C , Liu Z , et al. Network representation learning: an overview[J]. Scientia Sinica(Informationis), 2017.
17. M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. Computer Graphics Forum, 28(3):831–838, 2009
18. A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, pages 59–66. IEEE, 2009.
19. P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987
20. T. Calinski and J. Harabasz. A dendrite method for cluster analysis. ´ Communications in Statistics-theory and Methods, 3(1):1–27, 1974.
21. C. Dunne, S. Ross, B. Shneiderman, and M. Martino. Readability metric feedback for aiding node-link visualization designers. IBM Journal of Research and Development, 59(2/3):14–1, 2015.
22. J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, LINE: Large-scale information network embedding, in Proceedings of the 24th International Conference on World Wide Web. ACM, 2015, pp. 1067–1077.
23. B. Perozzi, R. Al-Rfou, and S. Skiena, Deepwalk: Online learning of social representations, in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014, pp. 701–710.
24. S. Cao, W. Lu, Q. Xu, Deep neural networks for learning graph representations, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1145–1152.
25. D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1225–1234.
26. S. Cao, W. Lu, and Q. Xu, “Grarep: Learning graph representation swith global structural information,” in Proceedings of the 24th ACM International Conference on Information and Knowledge Management. ACM, 2015, pp. 891–900
27. A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 855–864
28. M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, Asymmetric transitivity preserving graph embedding, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 1105–1114
29. C. Zhou, Y. Liu, X. Liu, Z. Liu, and J. Gao, “Scalable graph embedding for asymmetric proximity,” in Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 2942–2948
30. C. Tu, W. Zhang, Z. Liu, and M. Sun, “Max-margin deepwalk: discriminative learning of network representation,” in Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 3889–3895.
31. X. Zhang, W. Chen, and H. Yan, “Tline: scalable transductive network embedding,” in Information Retrieval Technology. Springer, 2016, pp. 98–110.
32. J. Chen, Q. Zhang, and X. Huang, “Incorporate group information to enhance network embedding,” in Proceedings of the 25th ACM International Conference on Information and Knowledge Management. ACM, 2016, pp. 1901–1904.
33. S. Wang, J. Tang, C. Aggarwal, and H. Liu, “Linked document embedding for classification,” in Proceedings of the 25th ACM International Conference on Information and Knowledge Management. ACM, 2016, pp. 115–124.
34. M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. Computer Graphics Forum, 28(3):831–838, 2009
35. A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, A. J. Smola, Distributed large-scale natural graph factorization, in: Proceedings of the 22nd international conference on World Wide Web, ACM, 2013, pp. 37–48.
36. M. Ou, P. Cui, J. Pei, Z. Zhang, W. Zhu, Asymmetric transitivity preserving graph embedding, in: Proc. of ACM SIGKDD, 2016, pp. 1105–1114.
37. Ribeiro L F R , Savarese P H P , Figueiredo D R . struc2vec: Learning Node Representations from Structural Identity[J]. 2017.
38. A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 855–864.
39. Dong Y , Chawla N V , Swami A , et al. metapath2vec: Scalable Representation Learning for Heterogeneous Networks[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2017.
40. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907.
41. T. Feder, R. Motwani, Clique partitions, graph compression and speeding-up algorithms, in: Proceedings of the twenty-third annual ACM symposium on Theory of computing, 1991, pp. 123–133.
42. P. M. Pardalos, J. Xue, The maximum clique problem, Journal of global Optimization 4 (3) (1994) 301–328.
43. Y. Tian, R. A. Hankins, J. M. Patel, Efficient aggregation for graph summarization, in: Proceedings of the SIGMOD international conference on Management of data, ACM, 2008, pp. 567–580.
44. H. Toivonen, F. Zhou, A. Hartikainen, A. Hinkka, Compression of weighted graphs, in: Proc. 17th international conference on Knowledge discovery and data mining, 2011, pp. 965–973
45. S. Navlakha, R. Rastogi, N. Shrivastava, Graph summarization with bounded error, in: Proceedings of the international conference on Management of data, ACM, 2008, pp. 419–432
46. J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471
47. K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572.
48. L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.
49. C. H. Ding, X. He, H. Zha, M. Gu, H. D. Simon, A min-max cut algorithm for graph partitioning and data clustering, in: International Conference on Data Mining, IEEE, 2001, pp. 107–114
50. J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on pattern analysis and machine intelligence 22 (8) (2000) 888–905
51. M. E. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical review E 69 (2) (2004) 026113
52. X. Xu, N. Yuruk, Z. Feng, T. A. Schweiger, Scan: a structural clustering algorithm for networks, in: Proceedings 13th international conference on Knowledge discovery and data mining, 2007, pp. 824–833.
53. S. White, P. Smyth, A spectral clustering approach to finding communities in graphs, in: Proceedings of the 2005 SIAM international conference on data mining, SIAM, 2005, pp. 274–285
54. D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, journal of the Association for Information Science and Technology 58 (7) (2007) 1019–1031.
55. L. Lu, T. Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications 390 (6) (2011) 1150–1170.
56. M. Al Hasan, M. J. Zaki, A survey of link prediction in social networks, in: Social network data analytics, 2011, pp. 243–275.
57. L. Katz, A new status index derived from sociometric analysis, Psychometrika 18 (1) (1953) 39–43
58. P. Jaccard, Etude comparative de la distribution florale dans une portion des Alpes et du Jura, Impr. Corbaz, 1901.
59. L. A. Adamic, E. Adar, Friends and neighbors on the web, Social networks 25 (3) (2003) 211–230 A. Clauset, C. Moore, M. E. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98– 101.
60. H. C. White, S. A. Boorman, R. L. Breiger, Social structure from multiple networks. i. blockmodels of roles and positions, American journal of sociology 81 (4) (1976) 730–780.
61. N. Friedman, L. Getoor, D. Koller, A. Pfeffer, Learning probabilistic relational models, in: IJCAI, 1999, pp. 1300–1309.
62. D. Heckerman, C. Meek, D. Koller, Probabilistic entity-relationship models, prms, and plate models, Introduction to statistical relational learning (2007) 201–238
63. S. Bhagat, G. Cormode, S. Muthukrishnan, Node classification in social networks, in: Social network data analytics, Springer, 2011, pp. 115–148.
64. S. Bhagat, I. Rozenbaum, G. Cormode, Applying link-based classifi- cation to label blogs, in: Proceedings of WebKDD: workshop on Web mining and social network analysis, ACM, 2007, pp. 92–101.
65. Q. Lu, L. Getoor, Link-based classification, in: ICML, Vol. 3, 2003, pp. 496–503.
66. J. Neville, D. Jensen, Iterative classification in relational data, in: Proc. Workshop on Learning Statistical Models from Relational Data, 2000, pp. 13–20.
67. D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons, 2013.
68. A. McCallum, K. Nigam, et al., A comparison of event models for naive bayes text classification, in: AAAI-98 workshop on learning for text categorization, Vol. 752, Citeseer, 1998, pp. 41–48
69. A. Azran, The rendezvous algorithm: Multiclass semi-supervised learning with markov random walks, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 49–56.
70. S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, M. Aly, Video suggestion and discovery for youtube: taking random walks through the view graph, in: Proc. 17th int. conference on World Wide Web, 2008, pp. 895–904
71. A. L. Traud, P. J. Mucha, and M. A. Porter, “Social structure of facebook networks,” Physica A: Statistical Mechanics and its Applications, vol. 391, no. 16, pp. 4165–4180, 2012.
72. <http://citeseerx.ist.psu.edu/>
73. B.-J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D. H. Lackner, J. Bahler, V. Wood ¨ et al., “The biogrid interaction database: 2008 update,” Nucleic Acids Research, vol. 36, pp. D637–D640, 2008.
74. Li C , Baciu G , Wang Y . Module-based visualization of large-scale graph network?data[J]. Journal of Visualization, 2016.