基于拓扑结构表示学习的大规模无监督图对齐方法研究

王晨旭[1),2)] 周俊铭[1)] 姜佩京[1)]

[1)](西安交通大学软件学院 西安 710049)

[2)](西安交通大学智能网络与网络安全教育部重点实验室 西安 710049)

摘 要 图数据因其较强的复杂关系表征能力受到广泛关注,在社交网络、学术合作、道路交通、生物信息等多个领域具有重要应用.图对齐技术旨在找出不同图中属于同一实体的节点对,在多个领域具有重要的应用价值,例如,对不同社交网络中属于同一个用户的账号进行关联可以为推荐系统提供更丰富的用户行为画像,对不同生物组织的蛋白质网络进行对齐能够辅助研究人员分析蛋白质的特性和机能.然而,在缺乏人工标注信息的情况下仅使用图的拓扑结构信息实现无监督图对齐一直是图数据挖掘面临的重要难题之一,特别是在大规模图对齐任务中,存在初始种子节点发现难和计算效率低下的问题.针对以上问题,本文提出了一种基于拓扑结构表示学习的大规模无监督图对齐框架.首先,从待匹配图中选取典型子图作为种子节点候选集,利用局部拓扑结构信息计算得到高可靠的种子节点匹配对;然后利用所得种子节点将待匹配图进行融合,并提出一种高效的无监督表示学习算法将融合图映射到统一的向量空间中;最后利用学习得到的节点向量实现大规模图对齐.与已有方法相比,本文所提方法在大规模图对齐任务中用时最短,对齐结果准确率最高,且算法性能受图结构的差异性影响最小.

关键词 图对齐;大规模图;无监督表示学习

中图法分类号 TP301 **DOI 号** 10.11897/SP.J.1016.2023.01350

Large-Scale Graph Alignment Based on Topological Structure Representation Learning

WANG Chen-Xu[1),2)] ZHOU Jun-Ming[1)] JIANG Pei-Jing[1)]

[1)](*School of Software Engineering, Xi'an Jiaotong University, Xi'an* 710049)

[2)](*Ministry of Education Key Lab of Intelligent Network and Network Security, Xi'an Jiaotong University, Xi'an* 710049)

Abstract Graph data has attracted lots of attention due to its strong ability of representing complex relationships. It has been widely used in many fields, such as social networks, academic cooperation, road traffic, and biological information. Graph alignment aims to find node pairs belonging to the same entity in different graphs, which have valuable applications in many fields. For example, associating accounts belonging to the same user in different social networks can provide richer user behavior profiles for recommender systems, and aligning protein networks of different biological tissues can assist researchers in analyzing the characteristics and functions of proteins. However, unsupervised graph alignment using the topological information of graphs has always been one of the important problems faced by graph data mining in the absence of manual annotation information. There are difficulties in finding initial seed nodes and low computational efficiency, especially for large-scale graph alignment tasks. To solve these problems, this paper proposes a large-scale unsupervised graph alignment framework based on topological structure representation learning. Firstly, a typical subgraph is selected from each of the graphs as a candidate

收稿日期:2022-07-18;在线发布日期:2023-01-09.本课题得到国家自然科学基金(62272379)、陕西省自然科学基金(2021JM-018)、国家重点研发计划(2021YFB1715600)、中央高校基本科研业务费专项资金(1191320006)资助.**王晨旭**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为图大数据挖掘、人工智能、网络安全等.E-mail:cxwang@mail.xjtu.edu.cn.**周俊铭**,硕士,主要研究方向为图匹配对齐、网络表示学习等.**姜佩京**,硕士研究生,主要研究方向为图匹配对齐、网络表示学习等.

set of seed nodes. The local topological information is used to retrieve a set of highly reliable seed node pairs. Then, we use the seed nodes to fuse the matching graphs, and propose an efficient unsupervised representation learning algorithm to map the fused graph into a unified vector space. Finally, large-scale graph alignment is realized based on the learned node vectors. Compared with existing methods, the proposed approach uses the least time in large-scale graph alignment tasks and achieves the best performance of alignment accuracy. Moreover, the structural differences of graphs have limited impacts on the performance of the proposed method.

Keywords　graph alignment; large-scale graphs; unsupervised representation learning

1　引　言

图数据结构作为对复杂关系进行建模的重要形式之一，在许多领域有着重要的应用，如在线社交媒体中的社交网络[1]、科学研究中的合作者网络[2]、生物信息领域的蛋白质网络[3]等. 图数据挖掘能够获取很多有价值的信息，如节点分类可以推断科研人员所属的研究领域[4]、节点聚类可以分析社交网络中的兴趣社区[5]、链接预测能够为用户推荐更多的好友关系[6]、节点重要性评估能够分析蛋白质在生物功能中的重要性[7]及社交网络中用户的影响力[8]等. 近年来，旨在识别不同图中的节点是否属于同一实体的图对齐任务受到越来越多的关注[9]，例如，图对齐能够识别不同社交网络中的账号是否来自同一用户，帮助推荐系统构建更为完善的用户画像，提高推荐的准确度[10]；图对齐能够发现不同生物蛋白质网络中相同或相似的蛋白质关联关系，为生物科学研究提供重要的参考价值[11].

图表示学习又名图嵌入[4]，旨在将图中的节点表示为低维、实值、稠密的向量形式，得到的向量表示在空间中的距离反映了节点在图中的亲和度. 这些向量可作为特征输入到传统机器学习算法模型中，进而完成一些常见的下游任务，如节点分类、链接预测及社区发现等. 近年来，利用图表示学习来完成图对齐任务逐渐成为研究的热点.

由于图对齐任务的重要应用价值，已有不少图对齐算法被提出，目前图对齐算法大致可以分为两种，第一种是传统的离散型图对齐算法：如 BigAlign[12]认为图对齐是对源图和目标图的邻接矩阵的行和列进行重新排序，并将图对齐问题转化为置换矩阵的最优求解问题；IsoRank[13]利用节点邻域拓扑结构的相似性实现蛋白质网络的对齐；FINAL[14]提出结构相似度、节点特征相似度和边特征相似度三个衡量标准来解决图对齐问题. 第二类为基于图表示学习的对齐方法：如 REGAL[15]首先通过计算邻居节点度的相似性来学习节点的向量表示，并利用节点向量的相似性来完成图对齐任务；CrossMNA[16]通过图间向量的方式实现多个图的对齐任务，该方法通过图间信息不断优化 2 类节点向量，并通过图间向量的相似性实现节点对齐；GAlign[17]采用一种基于多层向量表示的无监督图对齐框架，该方法通过图卷积网络(GCN)学习节点的向量表示，提出一种数据增强方法确保模型满足图的一致性约束；WAlign[18]使用轻量级的 GCN 结合 Wasserstein 距离鉴别器，确保在训练的过程中源图和目标图的节点具有相似的分布概率，从而找到最佳的节点对应关系. 这些算法在取得成功的同时，也面临着重大的挑战. 首先，随着当前图数据的规模不断扩大，已知的图对齐算法计算效率往往较低，节点之间复杂的关联关系又导致了算法难以并行化，导致很多算法无法应用到大规模图数据的对齐任务上. 其次，传统的图对齐算法往往针对单一领域的图数据而设计，缺乏通用性. 如 IsoRank 算法专门为蛋白质网络对齐而设计[13]，将其直接应用到社交网络的对齐任务时往往出现性能较低、匹配结果不可信的问题.

目前图表示学习在大规模图上进行无监督匹配对齐主要存在两个难点：一是初始种子节点发现难，由于不存在先验的种子匹配节点，直接对全图进行匹配会导致较高的错误率，因此需要先找到一部分高可靠的匹配节点对作为种子. 但如何找到有效的高可靠种子节点以及如何实现初始匹配是研究的难点. 二是将种子节点应用到全图匹配时，当前大部分的图表示学习算法均存在计算效率与表示学习有效性之间的矛盾. 部分图表示学习方法仅使用图的局部结构信息，只保留了节点的 1 阶或 2 阶邻居相似性信息，尽管这种做法时间开销较小，但是难以对全图进行有效的表示学习；部分方法使用了图的高阶

邻居信息，能够学习到距离较远节点之间的相似性关系，相较于仅使用图的局部结构信息而言，这种做法可以提升图对齐的准确率，但在捕获全局结构信息时往往存在计算效率低下、时间开销大的问题.

当前已有的基于图表示学习的对齐算法往往分别学习 2 个待匹配图的节点向量表示[10,13]，然后利用学习得到的表示向量计算节点间的相似性，相似度高的节点对会被判定为匹配节点. 然而，当前基于拓扑结构的图表示学习算法往往需要对节点表示进行随机的初始化，导致学习得到的待匹配图的节点表示向量未分布在同一向量空间[10,19]，需要将二者映射到同一空间中，特别是在无监督的条件下，往往会导致较低的匹配准确率.

为了解决以上问题，本文提出了一种基于图表示学习的两阶段图对齐框架 LGA（Large-scale Graph Alignment），首先采用高可靠的无监督图匹配算法对两个子图进行匹配得到种子节点，然后利用种子节点对 2 个大图进行融合，然后对融合图进行无监督表示学习得到节点的表示向量，如此可以直接将待匹配图映射到统一的向量空间中，避免了由于表示空间的转换带来的匹配误差. 具体步骤如图 1 所示. 第一阶段为高可靠种子节点发现，先从要匹配的大图中抽取它们的典型子图，利用少量有代表性的节点构成典型子图，然后采用高可靠的无监督图匹配算法对两个子图进行匹配得到种子节点. 第二阶段则利用种子节点对 2 个大图进行匹配，先是根据种子节点对 2 个大图进行融合，将二者合为一个更大规模的图，然后对融合图进行无监督表示学习得到表示向量，最后利用表示向量计算节点之间的相似性进行匹配对齐. 由于融合图的节点数量非常庞大，目前已有的图表示学习算法无法快速高效地表示学习. 针对此问题，本文提出一种基于稀疏矩阵分解和个性化随机游走相结合的高效图表示学习方法，在保证计算效率的同时最大限度地捕获融合图的全局结构信息. 为了验证算法的有效性，本文在 6 个不同类型和不同规模的数据集上进行实验验证，并与多个已知最先进的无监督图对齐算法进行对比，实验结果表明，本文所提方法能够快速高效地实现大规模图的匹配对齐，取得了最高的匹配准确率，且在两个匹配图结构差异较大的情况下依然具有较高的匹配准确率. 本文工作的主要贡献如下：

（1）提出一种基于种子节点融合的大规模图对齐算法，与以往利用种子节点监督图表示学习的范式不同，本文所提方法直接在融合图上进行无监督的图表示学习，直接将待匹配的图数据映射到统一的向量空间中，避免了由于表示空间转换导致的匹

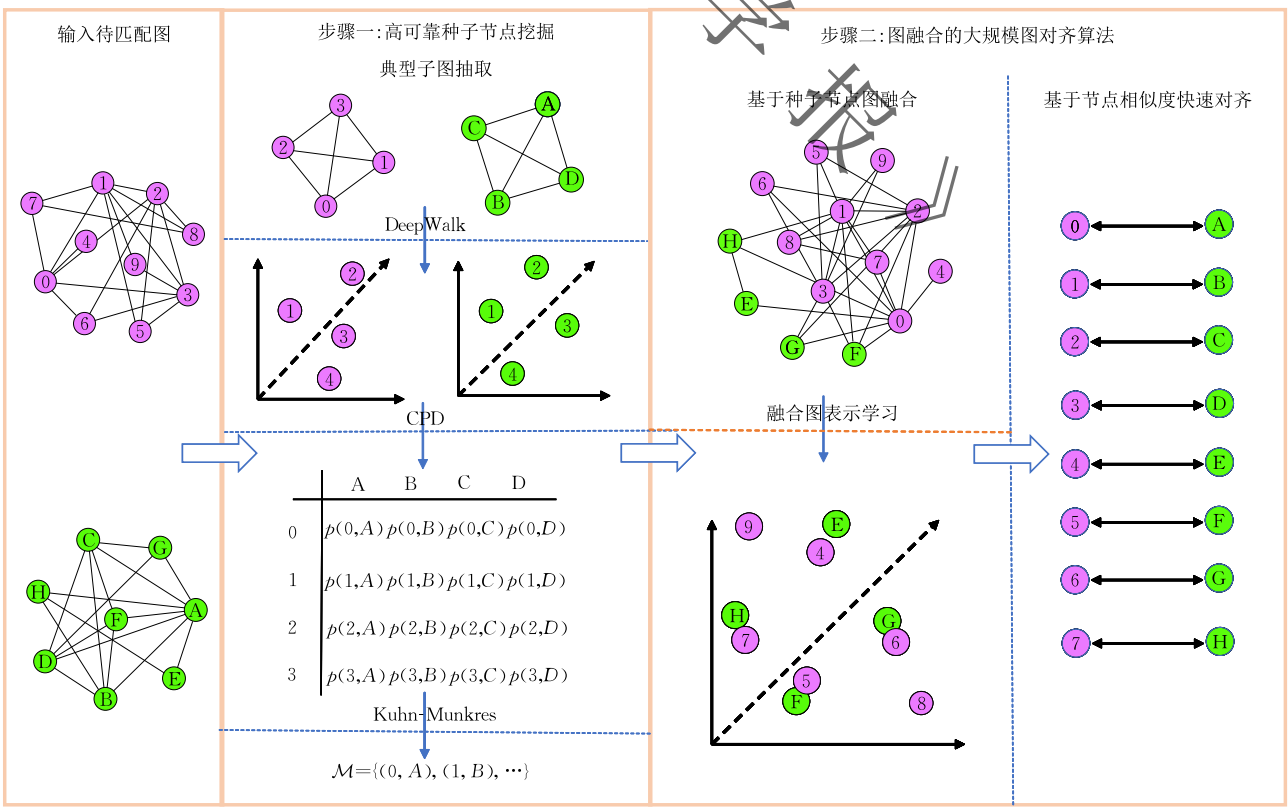


图 1 LGA 大规模图对齐框架示意图

配准确率急剧下降的问题.

(2)提出一种基于稀疏矩阵分解和个性化随机游走相结合的图表示学习方法,实现了对大规模图的高效表示学习,得到的表示向量能够捕获图的全局结构信息,并能够很好地适用于图对齐任务.

(3)在多种类型和不同规模的数据集上进行了广泛的实验验证,结果表明本文所提方法能够应用于大规模图数据的高效对齐,取得了远高于其他方法的匹配准确率和计算效率.本文算法代码已开源:https://github.com/zhoujunming2019/LGA-master.

2　相关工作

2.1　传统图对齐算法

图对齐在多个领域都有着重要应用,如在生物领域可以研究蛋白质之间的关系[11],在社交网络领域[12],跨平台的社交网络用户对齐可以用于用户画像、用户兴趣挖掘;在知识图谱领域可以实现知识图谱的补全和扩展[9].传统图对齐算法大多直接计算对齐矩阵 $\boldsymbol{S}$,通常以邻接矩阵的形式输入源图和目标图,在对齐过程中通过优化损失函数不断地更新对齐矩阵[7].

BigAlign 将图对齐问题建模为目标图邻接矩阵的行与列的重新排列问题[12],通过优化损失函数找到置换矩阵 $\boldsymbol{H}$ 使得排列后的目标图邻接矩阵接近于源图的邻接矩阵.为了解决求解过程中存在的 NP(Non-deterministic Polynomial-time)难问题,BigAlign 使用更为宽松的约束,并在优化的过程中使用迭代优化提高计算效率;IsoRank 认为如果两个图中节点的邻域相似,那么它们将大概率是匹配的[13].因此,IsoRank 认为不同蛋白质网络中的节点在它们各自的序列和邻域拓扑结构相似的情况下将获得较高的匹配率,基于该思路提出了一种专门针对蛋白质网络的节点对齐算法.然而传统图对齐方法除了对结构噪声较为敏感外,其计算效率也不尽如人意.

2.2　基于图表示学习的图对齐方法

随着图表示学习技术的发展,基于节点表示的图对齐算法逐渐成为研究的热点.IONE[8]在研究社交网络用户对齐问题中提出将每个用户的粉丝(follower-ship)/关注者(followee-ship)建模为输入/输出的上下文表示,以保持拥有相似粉丝/关注者的用户的表示向量在嵌入空间中距离较近.为了进一步解决不同图中节点所在嵌入空间不一致的问题,IONE 选取具有较大匹配概率的节点作为锚节点,使其对相邻的节点在空间中提供的上下文嵌入表示权值更高,最后采用随机梯度下降和负采样算法解决扩展性问题.REGAL[11]通过对节点一阶近邻和二阶近邻的度信息进行加权计算获取节点初始的表示向量,根据初始表示向量的相似性计算得到节点的相似度矩阵,使不同图中具有相似结构的节点具有更高的相似度,最终对相似度矩阵进行分解从而得到节点的低维向量表示.此方法仅考虑了图的局部拓扑信息,通常情况下对结构噪声的较为敏感.

DeepLink[20]基于随机游走策略生成结点序列后使用 Skip-Gram 模型对目标图和源图分别训练得到两个不同的嵌入向量空间,在解决特征空间不一致性的问题时引入深度神经网络,利用种子节点训练得到两个特征空间的非线性变换关系.为了提高图对齐的准确率,DeepLink 在设计损失函数时考虑了对偶学习,在有监督的方式下使损失函数最小化来学习源图和目标图的向量表示,需要大量的人工标注信息,往往难以获得.Chu 等人[16]针对多个图对齐问题提出了 CrossMNA 算法,该模型利用图之间的交叉信息不断优化节点的向量表示,将图间向量用于图对齐任务,图内向量用于其他下游图分析任务.实验结果表明 CrossMNA 能够取得较好的图对齐效果,并在较少内存使用的情况下实现更好的连接预测任务.Trung 等人[17]提出了 GAlign 模型,该模型利用图卷积网络(Graph Convolutional Networks,GCN)的多层嵌入并结合拓扑一致性原则实现无监督图对齐任务.在训练过程中,通过比对 GCN 各层之间的嵌入向量来获取锚节点,通过锚节点实现 GCN 的加权传播.此外,该模型引入基于扰动的数据增强方法和自适应的损失函数来增强模型的鲁棒性.Gao 等人[18]提出了 WAlign 模型,该模型能够捕获图的局部和全局结构信息,并解决了传统 GCN 由于卷积层数的增加而导致的过度平滑问题.该模型采用一种轻量级的基于 GCN 的图对齐框架,在优化过程中使用 Wasserstein 距离鉴别器,确保在训练的过程中源图和目标图的节点具有相似的概率分布,并引入新的损失函数来避免训练过程中得到平凡解,找到最优的节点对应关系.

Derr 等人[21]提出了一种基于深度对抗学习的无监督图对齐方法 DANA(Deep Adversarial Network Alignment),该算法通过对齐节点的表示向量分布来发现复杂的节点映射关系,并在此基础上实现快

速的最近邻节点对齐.此外,该模型还具备一定模型选择能力.Nguyen 等人[22]提出了 NAWAL 模型,该模型将源图和目标图分别投影到各自的向量空间中,并使用一种基于对抗学习的无监督对齐方法来优化源图和目标图之间的映射函数,从而找到最佳的匹配映射.实验表明 NAWAL 能够在真实数据集上取得不错的对齐效果,并且可以与现有的有监督方法相结合,利用先验信息更好地指导对齐.Zheng 等人[23]提出了 JORA 模型来解决社交网络间用户身份的弱监督对齐问题,该模型同时优化了表示学习和图对齐模型,同时保留了图内和图间节点的相似性信息,并采用注意力机制自适应地学习节点的相似性,极大地减少了模型对标签数量的依赖.

与以上方法相比,本文提出一种新的基于融合图表示学习的大规模无监督图对齐框架.首先采用典型子图获取高可靠的种子匹配节点对,然后利用种子节点对目标图和源图进行融合,并提出了一种高效的图表示学习算法直接将融合图映射到统一的向量空间,避免了由于特征空间转换带来的误差,且不需要对深度神经网络模型进行训练,在保证计算效率的同时尽可能地捕捉图的结构信息,大大提高了准确性和鲁棒性.

3 高可靠种子节点挖掘

设待匹配图为 $G_1(V_1,E_1)$ 和 $G_2(V_2,E_2)$,其中 V_1 和 V_2 分别为两个图的节点集合,E_1 和 E_2 为边的集合,为了表达上更清晰方便,本文在后续章节均采用符号 u 代表图 G_1 中的节点,采用 v 表示图 G_2 中的节点.基于拓扑结构的无监督图对齐可以定义为:在两个节点子集 $V_1^s\subset V_1$ 和 $V_2^s\subset V_2$ 找到一个映射 $\mathcal{M}=\{(u,v)\mid u\in V_1^s,v\in V_2^s\}$,使得由 V_1^s 和 V_2^s 构成的两个子图之间具有最相似的拓扑连接结构.在缺乏先验信息和节点属性信息的情况下,对 G_1 和 G_2 进行无监督对齐主要依赖于二者之间的拓扑结构信息,在图规模较大的情况下直接进行全图匹配往往计算效率较低且会存在较高的错误率.为此,本文首先在图中选取典型子图作为候选种子节点集,对候选种子节点进行对齐,得到源图 G_1 与目标图 G_2 中高可靠的匹配节点对作为种子节点,为下一步的匹配计算奠定基础.

3.1 典型子图抽取

由真实世界数据构建的图结构,其节点的度往往遵循幂律分布[4],即只有少部分节点拥有大量的邻居节点而大部分节点只有较少的连接.图 2 为对本文后续实验中涉及到的 6 个公开数据集节点度的分布情况,由图可知,所有数据集均呈现出少部分节点拥有大量连接的情况,这部分节点包含了更加丰富的结构信息,首先对这些节点进行匹配对齐所得结果的准确性和可靠性会更高.因此,本文首先分别从待匹配图中选取度最大的前 K 个节点构成典型子图,然后对典型子图进行无监督图对齐得到种子节点.对待匹配图 $G_1(V_1,E_1)$ 来说,选取度数最大的前 K 个节点构成集合 $V_{1'}=\{u_1,u_2,\cdots,u_K\}$,则 $V_{1'}$ 为 V_1 的一个子集,设由 $V_{1'}$ 中的元素构成的子图为 $G_{1'}(V_{1'},E_{1'})$,其中 $E_{1'}$ 为原始边集合 E_1 的子集,且仅保留 $V_{1'}$ 中节点之间的边,如此得到的子图 $G_{1'}(V_{1'},E_{1'})$ 即为 $G_1(V_1,E_1)$ 的典型子集,由于 $V_{1'}$ 中的节点在原始图中度最大,因此典型子图 $G_{1'}(V_{1'},E_{1'})$ 保留了原始图的大部分拓扑结构信息.同样地,设 $G_2(V_2,E_2)$ 的典型子图为 $G_{2'}(V_{2'},E_{2'})$.本节所要挖掘的高可靠种子节点则是由 $V_{1'}$ 和 $V_{2'}$ 的子集构成的节点对的集合,该集合也是 $\mathcal{M}$ 的一个子集.

为了使 K 能够与原始图规模相适应,设 $K=\beta\sqrt[2]{n}$,其中 n 代表图中节点数量,β 为可调参数.种子节点数 K 的选择主要考虑了算法的普适性,即无论图规模的大小如何,算法均能够高效准确地选择合适的种子节点数.如果按照百分比选取种子节点,对于大规模图来说会造成种子匹配准确率下降以及预匹配效率低下的问题,而对小型图来说则会导致种子节点过少难以发挥预匹配的作用.本文提出的计算公式既不会在大规模网络上选择过多的节点进行预匹配,也不会在小型图上选择过少的节点来构建典型子图,从而达到算法适配的效果.

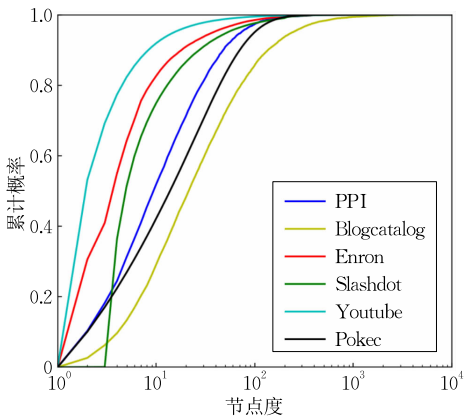


图 2 不同图中节点度的累计概率分布

3.2 典型子图匹配

本文采用 DeepMatching 算法[19]对典型子图进

行无监督对齐以得到高可靠种子节点，首先采用 DeepWalk[24] 图表示学习算法得到两个典型子图的节点向量表示，然后采用一致性点漂移算法（Coherent Point Drift，CPD）[25] 得到待匹配节点之间的匹配概率矩阵，最后采用最大权匹配算法得到高可靠的种子节点.

3.2.1　典型子图表示学习

设抽取得到的典型子图为 $G_{1'}(V_{1'},E_{1'})$ 和 $G_{2'}(V_{2'},E_{2'})$，首先使用 DeepWalk[24] 算法对两个子图进行编码得到节点的向量表示，DeepWalk 算法包括 2 个步骤. 第一步采用随机游走算法生成节点的上下文邻接节点信息，随机游走过程中每次从所在节点的邻居中随机选择一个节点作为下一跳，不断循环直到满足一定的路径长度后停止游走. 第二步采用 word2vec 技术中的 Skip-Gram 模型来学习节点的向量表示，在此过程中将图中的节点视为 word2vec 语言模型中的单词，将随机游走所得的路径视为自然语言中的句子，该模型通过最大化节点的上下文概率来学习节点的向量表示. 设由 DeepWalk 得到的两个典型子图的节点向量表示分别为 $\boldsymbol{R}^{1'}$ 和 $\boldsymbol{R}^{2'}$.

3.2.2　一致性概率矩阵计算

得到节点的向量表示后，本文使用一致性点漂移算法（Coherent Point Drift，CPD）[25] 计算种子节点之间一致性的概率矩阵. 一致性点漂移算法将两个点集 $\boldsymbol{R}^{1'}$ 和 $\boldsymbol{R}^{2'}$ 中节点之间的对应关系建模为关联概率问题，使用概率值描述两个点集中节点之间存在对应关系的确定性程度，即节点之间的匹配概率. CPD 使用混合高斯模型（Gaussian Mixture Model，GMM）描述点集 $\boldsymbol{R}^{1'}$ 和 $\boldsymbol{R}^{2'}$ 中所有节点对之间的确定性关系，然后对高斯混合模型的参数进行求解获得节点之间的匹配关系，通过最大似然法拟合高斯混合模型的质心，使概率密度函数的期望最大化，求解混合高斯模型中参数可得到节点间的匹配概率. GMM 模型采用 EM 算法来计算两个图中种子节点之间的匹配概率矩阵 $\boldsymbol{P}$，在期望求解阶段（E-step）中，其中的每个元素 p_{ij} 通过以下方式计算：

$$p_{ij}=\frac{\exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}\left\|\boldsymbol{r}_j^1-(\boldsymbol{r}_i^2+\boldsymbol{G}_{i,*})\boldsymbol{W}\right\|^2\right)}{\left(\sum_{k=1}^{n_2}\exp\left(-\frac{1}{2\boldsymbol{\sigma}^2}\left\|\boldsymbol{r}_j^1-(\boldsymbol{r}_k^2+\boldsymbol{G}_{k,*})\boldsymbol{W}\right\|^2\right)+\frac{w}{1-w}\frac{(2\pi\boldsymbol{\sigma}^2)^{d/2}n_2}{n_1}\right)},$$

其中 $\boldsymbol{W}$ 为系数矩阵，初始化为全 0 矩阵，$\boldsymbol{\sigma}^2$ 为各向同性斜方差矩阵，初始化值为 $\boldsymbol{\sigma}^2=-\frac{1}{dn_1n_2}\sum_{i,j=1}^{n_1,n_2}\left\|\boldsymbol{r}_j^1-\boldsymbol{r}_i^2\right\|^2$，$\boldsymbol{G}$ 为核对角矩阵，其元素值为 $g_{ij}=\exp(-\left\|\boldsymbol{r}_j^1-\boldsymbol{r}_i^2\right\|^2/2\beta^2)$，参数 $0\leq w\leq 1$，向量 $\boldsymbol{r}_j^1$ 与 $\boldsymbol{r}_i^2$ 为学习得到的节点表示向量，分别为 $\boldsymbol{R}^{1'}$ 和 $\boldsymbol{R}^{2'}$ 的第 j 行和第 i 行，d 为向量维度，n_1 与 n_2 为典型子图中节点的数量，β 为大于 0 的参数. 最大化阶段（M-step）主要更新系数矩阵，其计算方式如下：

$\boldsymbol{W}=(\boldsymbol{G}+\lambda\boldsymbol{\sigma}^2\mathrm{diag}(\boldsymbol{P1})^{-1})^{-1}(\mathrm{diag}(\boldsymbol{P1})^{-1}\boldsymbol{PM}_1-\boldsymbol{M}_2)$，其中 diag() 为对角化操作，$\boldsymbol{P}$ 为匹配概率矩阵，$\boldsymbol{1}$ 为全 1 的列向量，$\lambda>0$ 为参数. 经过迭代收敛可以得到匹配概率矩阵 $\boldsymbol{P}$. 即通过 CPD 可以得到节点 u_i 与 v_j 之间的匹配概率 $p(u_i,v_j)\in(0,1)$，其中 $u_i\in V_{1'}$，$v_j\in V_{2'}$.

在理想情况下，如果节点 u_i 与 v_j 存在实际的匹配关系，那么它们之间概率应该输出为 1，否则应该输出 0，概率值越大意味着两个节点的对应关系的确定性就越大.

3.3　种子节点生成

本文方法为基于拓扑结构信息的无监督图对齐算法，目前大多数无监督算法或多或少都利用到拓扑一致性约束，对两个图分布差异较大情况下的对齐任务均是基于有监督的，例如 Wang 等人[26] 利用双重约束机制用来平衡两个图中不同节点度差异十分大的情况. 目前大多数无监督算法都假设对齐的两个图之间具有拓扑一致性，当两个待匹配图的拓扑结构差异较大时，算法对齐的准确率往往较低，且节点之间的相似性也不会太高.

为了获得高可靠的种子节点对，本文根据种子节点的一致性概率矩阵，利用二分图最大权匹配算法使种子节点之间的匹配概率最大化. 二分图匹配的目标是在二分图中找出一个子图，要求任意属于该子图的两条边都不会交汇于同一个顶点，且该子图中所有边的权值之和最大. 本文使用 Kuhn-Munkres[27] 算法得到种子节点集合 $\mathcal{M}_s$. 本文选取两个图中度数最大的 K 个顶点来构建典型子图是因为其最大程度地保留了原始图的拓扑结构信息，从而确保所得的匹配种子节点具有较高的可靠性. 值得注意的是，并非度数较大的节点一定在另一个图中存在对应的匹配节点，为了确保种子节点的可靠性，仅保留匹配概率大于一定阈值的种子节点对，即 $p(u_i,v_j)>\theta$，$(u_i,v_j)\in\mathcal{M}_s$，$\theta$ 为可调参数，为了保障所得种子节点的可靠性，本文取 $\theta=0.9$，确保所得种子节点的准确性，以免对后续匹配造成较大的负面影响. 当两个图的拓扑结构差异较大时，匹配所得的大部分种子节点的匹配概率会低于该阈值，若所

得种子节点的数量少于 $0.5K$ 时，则认为源图和目标图之间存在较大的拓扑结构差异，不会进行后续的图融合匹配对齐，从而确保最终所得结果可靠有效.

4 基于图融合的大规模图对齐算法

在得到种子节点集 $\mathcal{M}$ 后，可以利用已有的有监督图对齐算法对 G_1 和 G_2 进行匹配对齐. 已有方法大多先采用图表示学习算法分别获取 G_1 和 G_2 节点的向量表示，然后根据种子节点将 G_1 和 G_2 的向量表示转换到同一向量空间中. 这一过程由于种子节点分布与全图节点分布的不一致性可能导致转换后的向量表示存在较大误差，影响匹配结果的准确性. 且由于 G_1 和 G_2 的规模较大，目前大多数的图表示学习算法如 DeepWalk[24]、GCN[28] 等均无法实现对结构信息的快速高效提取. 针对以上问题，本文提出一种基于图融合的大规模图对齐算法，首先根据种子节点对待匹配图融合为一个大图，然后提出一种高效的融合图表示学习算法获得节点的向量表示，最后利用节点向量计算它们之间的相似性并进行匹配对齐.

4.1 基于种子节点的图融合

为使已配对的高可靠种子节点在融合图中具有更多的结构信息权重从而降低整体结构差异带来的负面影响，本文基于已匹配的种子节点将 G_1 和 G_2 进行融合，构造成新的大图 G_3. 具体地，将 G_1 和 G_2 中已匹配的种子节点进行聚合形成一个新的节点，其余在 G_1 和 G_2 中的节点按照原有节点之间的连接关系添加到 G_3 中.

此过程可以看作将网络 G_1 和 G_2 基于种子节点进行连通，种子节点为连接纽带.

4.2 融合图表示学习

目前大多数表示学习方法难以捕获图中的全局结构信息，该问题在大规模图上尤为突出. 为此，提出一种基于个性化随机游走的图表示学习算法并应用于 G_3，可以快速得到节点的高效向量表示. 个性化随机游走能捕获到全图的结构信息，并能够高效地表达节点之间的亲近关系，已经被广泛应用于图数据挖掘中. 具体地，对于给定 G_3 中的一个源节点 u，从节点 u 开始进行随机游走，每到一个节点都以概率 α 停止游走并且回到源节点 u 重新开始，或者以概率 $(1-\alpha)$ 随机选择一个邻居节点继续游走. 经过多轮游走之后，图中每个顶点被源节点 u 访问到的概率会逐步收敛趋于稳定，稳定后的概率可以作为源节点 u 与其他节点之间的亲和度得分. 把图中所有节点分别当作源节点进行个性化随机游走，可获得如下的亲和度矩阵：

$$\boldsymbol{M}=\sum_{i=0}^{\infty}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^i \tag{1}$$

其中 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{A}$ 为概率转移矩阵，$\boldsymbol{A}$ 为图的邻接矩阵，$\boldsymbol{D}$ 为对角矩阵，对角线上元素为该节点在图中的度大小，即 $\boldsymbol{D}[i,i]=[deg(i)]$. 在大规模图中，由于节点数量庞大，直接计算个性化随机游走矩阵 $\boldsymbol{M}\in\mathbb{R}^{n\times n}$ 的效率低下. 此外，为了得到节点的低维度向量表示，还需要对个性化随机游走矩阵进行分解以达到降维的目的. 但是个性化随机游走矩阵 $\boldsymbol{M}$ 包含着全图的高阶结构信息，$\boldsymbol{M}$ 为稠密矩阵，对该矩阵进行奇异值分解的时间复杂度为 $O(n^3)$，在大规模图上会存在严重的计算效率问题，甚至不可行.

为了提高图表示学习的可扩展性，本文提出一种近似快速分解方法，在保证有效性的前提下提高表示学习的效率. 通过分析发现，图表示学习要求输出相似度矩阵分解后的节点低维表示，因此个性化随机游走矩阵 $\boldsymbol{M}$ 仅为中间产物，并不要求直接计算得出. 基于这一思想本文采用 ApproxPPR[29] 算法的思想，将矩阵 $\boldsymbol{M}$ 的计算和分解集成到每一次迭代运算中. 相似度矩阵 $\boldsymbol{M}$ 可以看成是图中节点从低阶到高阶亲和度的加权和. 首先对图 G_3 的邻接矩阵 $\boldsymbol{A}$ 进行奇异值分解：

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

式中 $\boldsymbol{U}$ 为左奇异矩阵，$\boldsymbol{\Sigma}$ 为奇异值对角矩阵，$\boldsymbol{V}$ 为右奇异矩阵. 此时，可以矩阵 $\boldsymbol{A}$ 看作矩阵 $\boldsymbol{X}_1$ 和 $\boldsymbol{Y}_1$ 的乘积，将 $\boldsymbol{X}_1$ 看作节点的初始表示向量为

$$\boldsymbol{X}_1=\boldsymbol{D}^{-1}\boldsymbol{U}\sqrt{\boldsymbol{\Sigma}} \tag{3}$$

而 $\boldsymbol{Y}_1=\boldsymbol{V}\sqrt{\boldsymbol{\Sigma}}$ 则为节点的上下文向量表示. 初始节点表示 $\boldsymbol{X}_1$ 只包含 G_3 中的一阶邻居信息，难以描述整个网络节点间的拓扑结构. 为了逐步聚合图中的高阶结构信息，对初始嵌入向量 $\boldsymbol{X}_1$ 进行迭代计算可以得到：

$$\boldsymbol{X}=\sum_{i=1}^{\infty}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\boldsymbol{X}_1 \tag{4}$$

迭代后的向量表示 $\boldsymbol{X}$ 包含了图中更高阶的结构信息，其中参数 α 的取值决定了表示向量包含的图结构信息的范围，α 取值过低则表示向量能够包含更多的高阶信息，但需要较多的迭代次数才能收敛；取值过低则会包含更多的局部结构信息，比如在极端情况下 α 取值为 1 时，则学习得到的表示向量只包

含节点的一阶邻居结构信息. 为了综合算法的有效性和时间效率，本文中 α 取值为 0.15.

由于邻接矩阵 $\boldsymbol{A}$ 为稀疏矩阵，可以采用 BKSVD[30] 进行近似奇异值分解得到 $\boldsymbol{A}\approx\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ 的近似解. 该奇异值分解的时间复杂度与网络中边的数量成线性关系，远低于分解稠密矩阵 $\boldsymbol{M}$ 需要的 $O(n^3)$ 复杂度，大大提高了图表示学习方法的可扩展性.

最终得到的图表示向量为 $\boldsymbol{R}=[\boldsymbol{X}\oplus\boldsymbol{Y}]$，$\oplus$ 为对表示向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行拼接操作.

4.3　基于社区结构的高阶信息快速传播

在对式(4)进行高阶信息计算与聚合过程中，需要大量的大规模矩阵乘法迭代运算，导致计算效率依旧不高. 为了提高这一部分的计算效率，本文利用图中普遍存在的社区结构对高阶信息传播过程进行加速计算. 在真实世界形成的图上往往会存在一定的聚集现象[31]，也称为社区结构，在同一社区内的节点之间连接关系更加紧密，而属于不同社区的节点其连接关系相对而言比较稀疏[32-35]，社区结构在真实图数据中广泛存在，如社交网络中具有相同兴趣爱好的用户之间会有更多的连接关系，从而形成兴趣社团[36]；而在生物医药领域，研究人员往往将具有相似生物功能的蛋白质和药物用边连接起来，从而会形成生物蛋白质网络中的社区结构[37]. 如图 3 所示. 图中出现两个社区结构，可以明显看出社区内的节点相比社区间的节点之间的连接更加紧密. 信息的传播依赖于节点之间的连接，其连接密集程度对传播效率会产生极大的影响，因此节点信息在社区内与社区间的传播存在较大差异，本文利用图中存在社区结构这一特点对高阶信息传播进行加速计算.

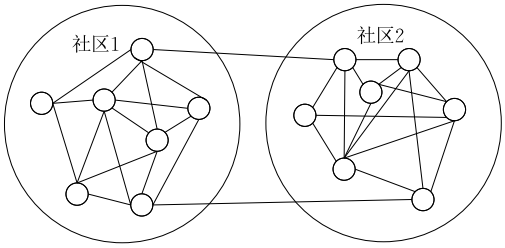


图 3　社区聚集现象

在式(4)中为了得到图的表示向量 $\boldsymbol{X}$，需要大量迭代使图中的高阶信息融入初始表示向量 $\boldsymbol{X}_1$ 中，该过程可以看作信息在图上不断的传播和聚合. 由于图中存在社区结构，社区内的节点之间连接较为紧密，导致节点信息在其相应的社区内反复传播，难以扩散到社区外，只有经过一定次数的迭代传播后，节点信息才能逐步扩散到全图上[38]. 因此，该信息在图中的传播过程可以分为社区传播和逃逸传播 2 个阶段，基于发现可以对表示向量 $\boldsymbol{X}$ 进行快速计算.

4.3.1　社区传播阶段

假定社区传播阶段的迭代次数为 s，在对社区传播阶段进行计算时，首先对式(4)中的前 $l<s$ 项进行求和计算，可以得到 $\boldsymbol{X}_{\mathrm{Reality}}$：

$$\boldsymbol{X}_{\mathrm{Reality}}=\sum_{i=1}^{l}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\boldsymbol{X}_1 \tag{5}$$

$\boldsymbol{X}_{\mathrm{Reality}}$ 包括了初始表示向量 $\boldsymbol{X}_1$ 前 l 次迭代，其计算结果为社区传播阶段中前 l 次节点信息在网络中传播得到的向量表示，其中 l 为可调参数. 设 $\boldsymbol{X}_{\mathrm{Community}}$ 为整个社区传播阶段得到的向量矩阵：

$$\boldsymbol{X}_{\mathrm{Community}}=\sum_{i=1}^{s}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\boldsymbol{X}_1 \tag{6}$$

由于图中存在社区结构，因此在一定次数内的信息传播其实是在社区内反复进行的，属于社区内的节点大概率会将其信息传播到同一社区内的节点上，在前 l 次传播中获得信息的节点将在接下来的 $(s-l)$ 次迭代中有很大概率会再次获得该信息，因此在社区传播阶段接下来的第 $l+1$ 次迭代到第 s 次迭代，节点获得的信息可以看作与前 l 次迭代所得信息成比例，即 $\boldsymbol{X}_{\mathrm{Community}}$ 可以从 $\boldsymbol{X}_{\mathrm{Reality}}$ 中计算得到：

$$\boldsymbol{X}_{\mathrm{Community}}\approx\delta\cdot\boldsymbol{X}_{\mathrm{Reality}} \tag{7}$$

式中 δ 为比例系数，对比前 l 次迭代与整个社区传播阶段的总得分可以求得比例系数 δ：

$$\delta=\frac{\left\|\sum_{i=1}^{s}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\right\|_1}{\left\|\sum_{i=1}^{l}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\right\|_1} \tag{8}$$

其中 $l<s$，由于概率转移矩阵 $\boldsymbol{P}$ 的每一项都是非负实数，可以将式(8)如下表示：

$$\delta=\frac{\sum_{i=1}^{s}\alpha(1-\alpha)^i\left\|\boldsymbol{P}^{i-1}\right\|_1}{\sum_{i=1}^{l}\alpha(1-\alpha)^i\left\|\boldsymbol{P}^{i-1}\right\|_1} \tag{9}$$

概率转移矩阵 $\boldsymbol{P}$ 为行随机矩阵，其每一行元素之和为 1，可以得出 $\boldsymbol{P}^{i-1}$ 也为行随机矩阵，且 $\left\|\boldsymbol{P}^{i-1}\right\|_1=n$，由此可以得出 δ 的表达式为

$$\delta=\frac{\sum_{i=1}^{s}\alpha(1-\alpha)^i}{\sum_{i=1}^{l}\alpha(1-\alpha)^i}=\frac{1-(1-\alpha)^s}{1-(1-\alpha)^l} \tag{10}$$

基于以上计算可以快速得到初始表示向量 $\boldsymbol{X}_1$ 在社区传播阶段进行高阶信息聚合后的空间向量矩阵

$\boldsymbol{X}_{\text{Community}}$.

4.3.2 逃逸传播阶段

在得到 $\boldsymbol{X}_{\text{Community}}$ 后，信息的传播将不再受限制于图中的社区结构，而将进一步扩散到全图的节点上，此阶段称为逃逸阶段，设经过逃逸阶段传播后的表示向量矩阵为 $\boldsymbol{X}_{\text{Escape}}$. 在社区传播阶段，距离源节点越近，节点从源节点获得的信息将会越多. 而在逃逸阶段由于经过多次迭代，在系数 $0<1-\alpha<0$ 的影响下，此时图中节点经过高阶传播后信息与源节点的相关性不断被削减，即在逃逸阶段节点所获得的信息与到传播源的远近无关，此时，度越大的节点越有机会接收到邻居节点传播的信息，并且所获得的信息与邻居节点所拥有信息正相关，此思想与 PageRank 的基本假设一致，

因此可以用原始的 PageRank 算法对逃逸阶段的信息传播进行计算. 首先计算图 G_3 的 PageRank 值，得到每个节点的权重，算法的累计幂迭代形式如下：

$$pagerank(G)=\sum_{i=0}^{\infty}\alpha(1-\alpha)^i\boldsymbol{e}\boldsymbol{P}^i \tag{11}$$

其中 $\boldsymbol{e}\in R^{1\times n}$，其中每一项元素值都为 $\frac{1}{n}$，n 为网络中节点的个数，最终聚合后的表示向量为

$$\boldsymbol{x}^i_{\text{Escape}}=\sum_{i=s+1}^{\infty}\alpha(1-\alpha)^i\boldsymbol{e}\boldsymbol{P}^i\boldsymbol{X}_1 \tag{12}$$

当 $\|\boldsymbol{x}^{i+1}_{\text{Escape}}-\boldsymbol{x}^i_{\text{Escape}}\|_1\leq 10^{-6}$ 时停止迭代，最终得到表示矩阵 $\boldsymbol{X}\approx\boldsymbol{X}_{\text{Community}}+n\boldsymbol{e}^{\mathrm{T}}\boldsymbol{x}_{\text{Escape}}$. 值得注意的是，式(12)中 $\boldsymbol{e}$ 为行向量，因此在计算时向量 $\boldsymbol{e}$ 先与矩阵 $\boldsymbol{P}$ 相乘得到新的行向量再继续迭代运算，而不是先计算矩阵 $\boldsymbol{P}$ 的多次幂，从而避免了大型矩阵的乘法运算.

4.4 基于节点相似度的快速匹配对齐

在得到节点的向量表示后，可以计算图 G_1 和 G_2 中所有节点两两之间的相似性，然后按照 Kuhn-Munkres[27] 算法进行匹配. 但由于节点数量庞大，该算法计算效率较低，为此，本文采用一种快速的匹配对齐算法进行计算. 首先，计算节点表示向量在空间中的欧氏距离来描述它们之间的相似性，假设节点 $u\in G_1$，$v\in G_2$，那么它们之间距离为 $d_{u,v}=\|\boldsymbol{v}_u-\boldsymbol{v}_v\|_2$，对于每一个属于 G_2 的节点，在 G_1 中寻找与其距离最近的节点进行匹配，一旦匹配则移除该节点，该算法虽牺牲了一定的准确性，但相比 Kuhn-Munkres 算法而言，其时间复杂度从 $O(n^3)$ 降为了 $O(n^2)$，提高了计算效率.

值得注意的是，本文核心算法虽然间接借鉴了 ApproxPPR 算法的思想和直接采用 Kuhn-MUnkres 等算法，但本文的核心贡献在于提出一种基于种子节点融合的大规模图对齐算法，与以往基于图表示学习的图对齐方法不同，本文所提方法直接在融合图上进行无监督的图表示学习，避免了表示空间的转换，从而大大提高了匹配准确率. 此外，为了提高图表示学习算法在大规模融合图上的计算效率，本文算法利用图的社区结构来加速高阶结构信息的传播，在保证有效性的同时大大提高了图表示学习的计算效率，使得本文算法能够较好地解决大规模图对齐问题.

5 实验与结果分析

5.1 实验数据与运行环境

本文在 6 个被广泛使用的公开数据集上进行了实验，表 1 列出了数据集的详细信息. PPI[39] 为人类蛋白质网络，BlogCatalog[39] 为社交网络，由不同博主与其社会关系构成，Enron[40] 是由联邦能源监管委员会在调查期间公布并发布的邮件通信网络，Slashdot[41] 为资讯科技网络，其所有新闻由用户提供，用户之间可以形成朋友或者敌人关系，Youtube[42] 为视频共享网站，用户之间可以建立友谊关系并进行视频分享，Pokec[43] 是斯洛伐克最受欢迎的在线社交网络. 所有的数据集都可以在公开网站或者论文中获得. 本文主要解决基于拓扑结构信息的无监督图对齐问题，由于算法仅依赖图的拓扑结构，参照以往类似算法的实验假设[44-46]：待匹配的两个图往往共享共同的连接关系，例如微信社交网络与 QQ 好友关系网络都是以用户之间真实的熟人关系为基础的，但由于人类活动的随机性和数据采集不完全等原因会导致此 2 个社交网络的拓扑结构具有一定的差异性，但都可以看作是从用户真实的社会网络按照一定的比例采样连接关系得到的. 因此，在实验时，将原始图表示为 G，对原始图 G 随机删除一定比例的边后得到最大连通子图 G_1 和 G_2，设删除边的比例为 γ，然后对 G_1 和 G_2 进行对齐. 此实验设置符合大多数的图对齐任务的应用场景. 由于随机裁剪边生成待匹配图具有一定的不确定性，为了获得更加稳定的实验结果，所有实验重复执行 20 次. 从运行结果来看，随机裁剪对运行结果并没有产生明显的影响，因此该操作具备普适性. 在所有实验中仅使用图结构信息来完成对齐任务，且初始时均不包含任何先验的匹配信息.

表 1　实验数据集

数据集名称	类型	节点数量*	边数量
PPI	蛋白质网络	3890	76584
BlogCatalg	社交网络	10312	333983
Enron	邮件通信网络	36692	183831
Slashdot	新闻网络	77360	905468
Youtube	视频共享网络	1134890	2987624
Pokec	社交网络	1632803	30622564

为了验证本文所提图对齐算法的有效性，与 7 种目前被广泛使用的无监督图对齐算法进行对比，分别是 BigAlign[12]、IsoRank[13]、FINAL[14]、REGAL[15]、GAlign[17]、WAlign[18] 和 NAWAL[22]. 本文所有实验代码均运行在系统为 Unbuntu 18.04、CPU 为 Intel(R) Xeon(R) Gold 6266C CPU@3.00GHz 和内存为 128GB 的服务器上. 表 2 列出了本文算法的参数设置表. 在分析不同参数对模型效果的影响时，为控制实验变量，仅改变待分析参数的取值，其余参数按照表 2 中的默认值进行设置并且保持不变. 本文以图对齐匹配准确率作为不同算法的评价指标，在计算时将匹配概率最大的节点对作为对齐节点，计算正确对齐节点占图 G_1 中总节点数量的百分比. 所有结果均为 20 次实验的平均值.

表 2　实验参数设置表

参数名称	参数描述	取值范围	默认值
α	个性化随机游走的停止概率	0～1.0	0.15
s	社区传播迭代次数	6～18	9
l	社区传播初始迭代次数	2～7	4
β	典型子图节点数量调节系数	2～12	4
d	嵌入向量的维度	40～18	80
θ	种子节点匹配概率阈值	0～1.0	0.90

5.2　实验结果及分析

5.2.1　种子节点匹配结果分析

图 4 为本文所提算法在种子节点匹配时的实验结果. 实验结果表明本文方法得到的种子节点具有较高的匹配准确度和可靠性，虽然随着移除边比例的增加，种子节点的匹配准确度也在逐步下降. 其中

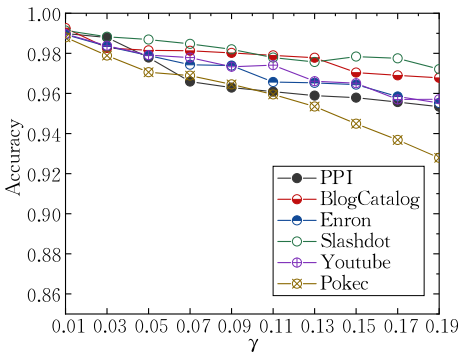


图 4　种子节点匹配准确率

Pokec 数据集上准确性下降最快，但即使在移除边的比例达到 20%的情况下依然取得了 0.92 以上的匹配准确率，确保了种子节点的可靠性，为后续基于图融合的匹配奠定了基础.

5.2.2　全图匹配结果分析

图 5 为在 6 个数据集上各算法的实验结果对比. 本文所提方法在 PPI 和 BlogCatalog 两个数据集上在所有移除边比例相同的情况下均取得了最好的匹配准确率. 与其他方法相比，本文所提方法的性能随移除边比例增加呈缓慢下降趋势，这是因为随着移除边比例的增加，待匹配图中包含的结构信息会随之减少，匹配准确率有所下降也在情理之中. 与本文方法形成鲜明对比的是，其他对比方法的匹配准确率随移除边比例的增加呈现急剧下降的趋势. 特别是 REGAL 与 WAlign 2 个方法，当移除边的比例不高时，待匹配图的拓扑结构较为相似，二者的匹配准确率除低于本文所提算法外，远高于其他 3 种方法. 但当移除边的比例大于 15%时，二者的匹配准确率急剧下降，其匹配准确率几乎与其他 3 种方法持平.

值得注意的是，当移除边的比例小于 3%时，REGAL 方法在数据集 Enron、Slashdot 和 Youtube 上的匹配准确率要高于本文所提算法，WAlign 方法在数据集 Enron 上的匹配准确率高于本文所提出的算法. 得到该匹配结果是因为 REGAL 方法通过计算待匹配节点邻居的度大小信息，在待匹配图结构差异不大的情况下实际匹配节点具有相似的结构信息，学习到的表示向量在特征空间上更加接近，从而能够取得较高准确率. 而 WAlign 方法最小化不同图节点表示向量之间的 Wasserstein 距离找到最优对齐结果，相当于将问题转化为最小代价的二分图最优匹配问题，在移除边比例较小的情况下能够使节点表示向量在不同图中具有相似的概率分布，从而达到较高的图对齐准确率，然而 WAlign 方法随着图规模的扩大，其计算所耗费时间会急剧上升. 同时这两种方法对待匹配图的拓扑结构差异性较为敏感，二者的性能随移除边比例的增加会急剧降低.

图 5(a)到图 5(d)的结果表明，NAWAL 模型较其他方法取得了相对稳定的匹配准确率，其性能未出现随着删除边比例的增加而急剧下降的情况. 这是因为 NAWAL 采用对抗训练的方式学习源图和目标图之间的映射关系，对两个图之间的结构差异变化具有较高的适应性. 然而 NAWAL 模型的准

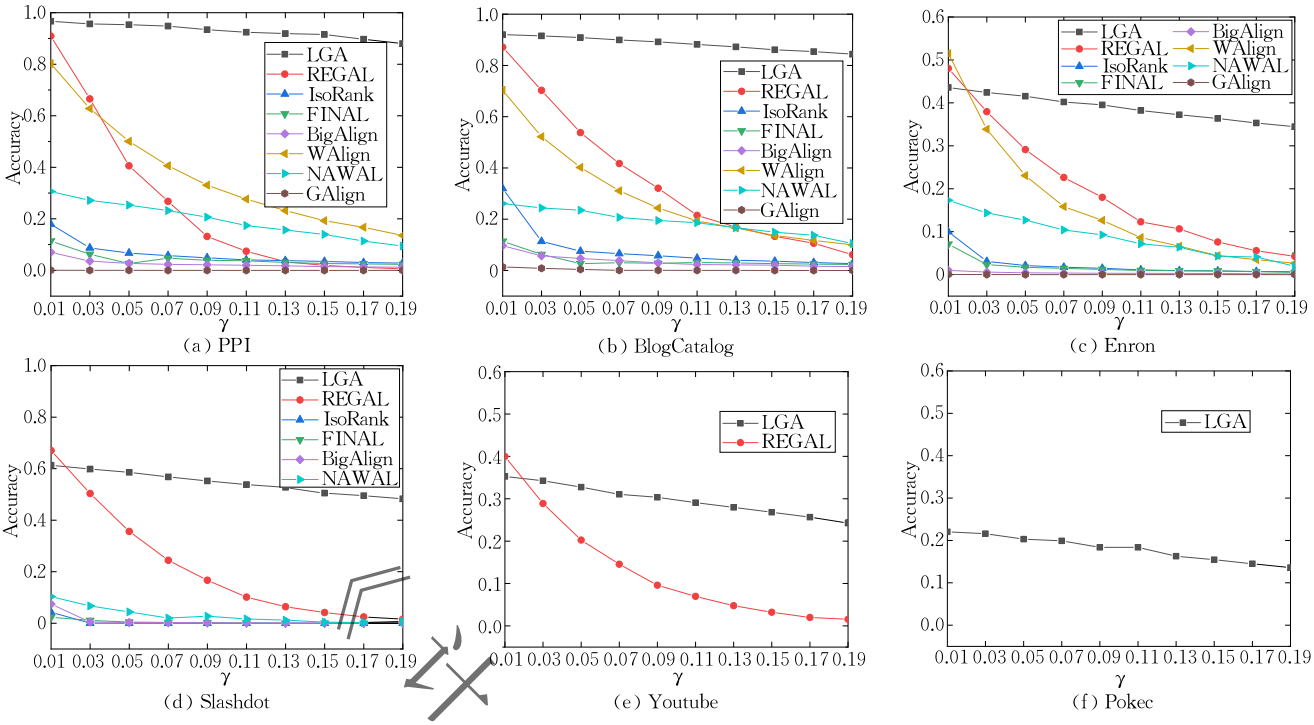


图 5 图对齐实验结果图

确率远低于本文所提方法. 此外,GAlign 模型在 PPI、BlogCatalog、Enron 上取得了很低的匹配准确率,这是因为该模型在节点对齐时非常依赖节点的属性信息,而本实验中仅使用节点的结构信息,导致了较高的错误匹配率.

特别说明,在数据集 Slashdot 上 WAlign、GAlign 等方法运行时间大于三天而无法得出匹配结果,在数据集 Youtube 上,除了本文提出的方法和 REGAL 方法,其他对比方法皆因效率或者内存原因而无法完成对齐实验,因此得到如图 5(d)和(e)所示的实验结果. 在拥有 3 千多万条边的大型数据集 Pokec 上,只有本文提出的算法在三天内得到匹配结果且不超出内存限制,实验结果如图 5(f)所示.

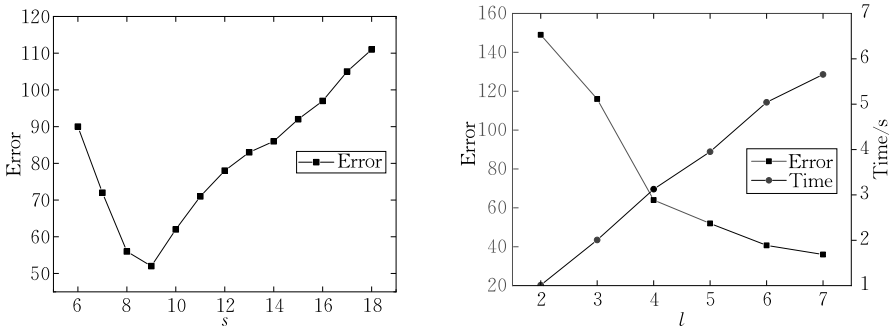
5.3 参数对实验影响

本小节验证信息传播参数 s 和 l、种子节点参数 β、向量表示维度 d 等参数对实验结果的影响. 其中参数 s 和 l 分别控制着社区传播阶段和逃逸传播阶段的迭代次数,二者的大小会对表示向量 $\boldsymbol{X}$ 计算的准确度与计算效率造成影响,将初始化向量 $\boldsymbol{X}_1$ 进行高阶信息传播所造成的计算误差定义为

$$error=\left\|\boldsymbol{X}_{\text{Community}}+\boldsymbol{X}_{\text{Escape}}-\sum_{i=1}^{\infty}\alpha(1-\alpha)^i\cdot\boldsymbol{P}^{i-1}\boldsymbol{X}_1\right\|_1.$$

在研究参数 s 对 $error$ 的影响时,为了控制变量,将 l 值固定为 4,改变参数 s 的大小,在大型数据集 Youtube 上进行实验,结果如图 6(a)所示. 实验结果表明计算误差 $error$ 在初始阶段会随着参数 s 的增大而减少,但当 s 值增长到 9 左右时,误差开始随着参数 s 增大. 基于社区结构进行高阶信息传播所产生的误差由社区传播 $\boldsymbol{X}_{\text{Community}}$ 和逃逸传播 $\boldsymbol{X}_{\text{Escape}}$ 构成,随着参数 s 的增大,社区传播阶段高阶信息的聚合需要估计的迭代次数会增多,从而造成此阶段产生的误差增长,而逃逸传播阶段需要估计的迭代次数会减少,因此误差减小,在 s 刚开始增长时逃逸阶段的误差减小量大于社区传播阶段误差增长量,而随着 s 的进一步增大,误差的增长量会逐渐大于减小量,从而造成整体误差随着参数 s 的增长先减小继而再上升的现象. 根据实验结果,本文设置参数 $s=9$. 图 6 表明参数 s 和 l 都会对误差造成影响,其中参数 l 会对参数 s 的拐点值的产生造成影响,参数 s 对 l 无影响. 此外,由式(10)可知,比例系数 δ 可由 s 的值直接计算得到,此部分计算时间与 s 的取值大小无关,因此本文所提算法的运行时间主要由参数 l 的取值大小决定,因此在图 6(a)中未画出运行时间.

在研究参数 l 对误差与计算效率的影响时,将 s 值固定设置为 9,改变 l 值大小,实验结果如图 6(b)所示. 随着 l 值的增长,社区传播阶段对向量 $\boldsymbol{X}_1$ 的传播迭代次数增加,因此耗费的时间会随之增长,但会减小该阶段带来的计算误差,考虑到准确性和运行时间之间的平衡,本文设置 $l=4$.

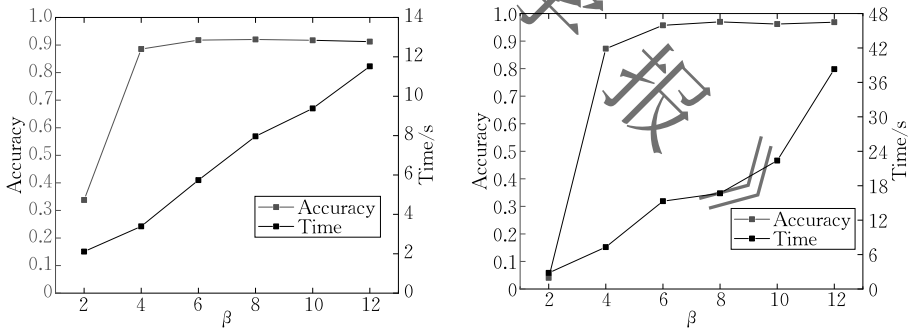


(a) 参数s对误差的影响　　(b) 参数l对误差和效率的影响

图 6　参数 s 和 l 对实验的影响

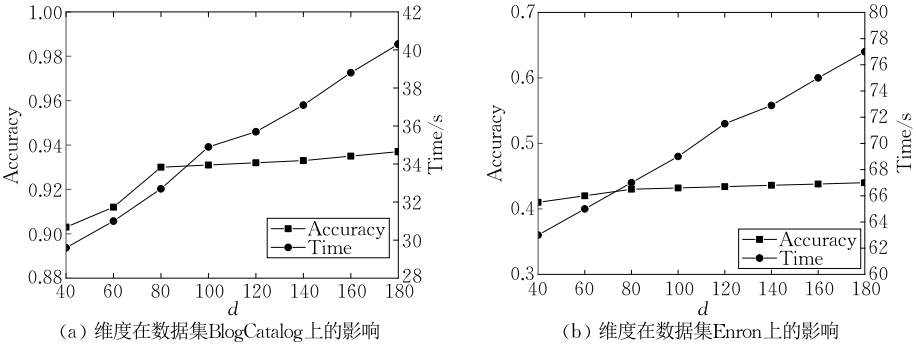
本文通过改变典型子图的大小来验证预匹配过程中获得的种子数量对算法性能的影响，由于典型子图的节点数量 $K=\beta\sqrt[2]{n}$，其中 n 为原始图节点的数量，β 为可调参数，图 7 展示了 β 的不同取值在 PPI 和 BlogCatalg 两个数据集上的实验结果. 当 $\beta\leq 4$ 时，图对齐的准确率会随着 β 的增大急剧提升，但当 $\beta>4$ 时，准确率的提升明显放缓，而种子节点匹配所耗费的时间会随着 β 的增大而线性上升. 以上结果表明，当典型子图过小时，会导致预匹配所得的种子节点数量较少，从而导致后续基于融合图的匹配结果准确度较低. 而当预匹配获取到一定数量的种子节点后，增加更多的种子节点也不会明显提高基于融合图的匹配准确率，且典型子图过大会导致预匹配阶段消耗的时间大大增加，为了兼顾准确性和计算效率，本文设置 $\beta=8$.

为了选择合适的维度值，本节将图表示学习方法输出的节点向量维度设置从 40 增加到 180，在 BlogCatalg 和 Enron 两个数据集上进行实验，评估向量维度对图对齐方法准确率和整体效率的影响，实验结果如图 8 所示. 准确率随着维度 d 的增加会缓慢提升，但所消耗的计算时间也会随之增加，且增长速度远高于准确率的提升. 这主要是由于本文所提的图对齐方法利用表示向量之间的相似性实现节点对齐这一步骤会随着维度增加而需要更长的搜索时间开销. 综合考虑算法准确率和计算效率，本文设置表示向量的维度 $d=80$.



(a) 参数β在数据集PPI上的影响　　(b) 参数β在数据集BlogCatalg上的影响

图 7　种子节点数量对准确率和效率的影响



(a) 维度在数据集BlogCatalog上的影响　　(b) 维度在数据集Enron上的影响

图 8　表示向量维度 d 对准确率和效率影响

5.4 计算效率对比

表 3 详细描述了不同方法在实验数据集上的运行时间，本文所提的图对齐方法在计算效率上明显优于其他方法，与传统图对齐方法 BigAlign、FINAL 和 IsoRank 相对比，本文所提方法 LGA 的运行时间在所有数据集上均处于绝对优势. 与同样使用表示学习的图对齐方法 REGAL 相比，REGAL 以在结构差异较大的图对齐任务中具有一定的鲁棒性为代价，避免了基于随机游走采样的计算开销，因此效率取得一定的提升，但与本文的方法 LGA 相对比依然有较高的计算成本，特别是在大规模图数据集上，其计算时间开销会显著增加，REGAL 方法在 Youtube 上运行时间将近 30 个小时，在拥有 3 千万条边的大型数据集 Pokec 上无法在三天之内完成计算. 与使用图卷积网络的 WAlign 方法相比，WAlign 在小规模数据集上有较高的时间效率优势，但随着节点数量的增长，其计算损失函数与参数更新会耗费大量的时间和内存，在接近 8 万个节点数据集上 Slashdot 上运行时间大于三天. 图 9 为各方法运行时间与节点数量之间的关系曲线图，随着节点数量的增长，FINAL、IsoRank 和 BigAlign 的运行时间会急剧增加. 而 REGAL、WAlign、LGA 三种方法在节点数量大于 35k 后时间消耗也开始大幅度增加，在大规模数据集上不具备可扩展性. 与这些方法不同的时，本文方法 LGA 随着网络节点数量的增加其运行时间始终保持平缓增长，在不同规模的数据集上均能高效地实现图对齐. 特别是本文方法在待匹配图之间的结构差异较大时仍能够达到良好的节点匹配准确率，并且在运行效率方面相比其他方法有较大的优势.

表 3 不同方法的运行时间比较 （单位：s）

名称	BigAlign	FINAL	IsoRank	REGAL	WAlign	LGA
PPI	43	108	62	14	**3**	13
BlogCatalg	452	678	543	218	**24**	34
Enron	2823	16155	9537	165	1223	**73**
Slahdot	21443	59370	34301	1068	时间＞3 天	**104**
Youtube		内存溢出		107602	时间＞3 天	**11031**
Pokec		内存溢出		时间＞3 天	内存溢出	**27036**

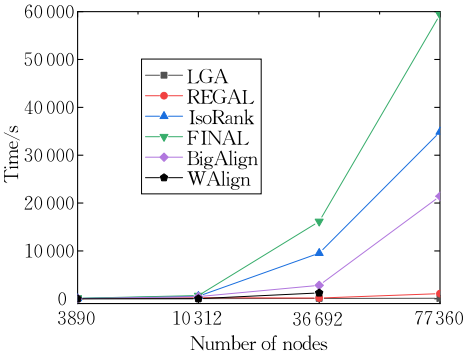


图 9 各算法计算效率随节点规模的变化曲线

5.5 消融实验

本节利用消融实验来验证本文提出的基于社区结构进行高阶信息传播（LGA-Comm）这一机制在图表示学习时的有效性，在表示学习阶段设置直接迭代进行高阶信息传播（LGA-Full）和不进行信息传播（LGA-Adj），在 PPI、BlogCatalog、Enron 和 Youtube 四个数据集上进行图对齐任务并进行对比，实验过程中将移除边的比例设置为 10%，基于不同传播方式取得的匹配准确率如图 10 所示. 从图中可以看出，基于社区结构进行高阶信息传播这一机制在进行图对齐任务时能达到与直接迭代信息传播几乎相同的效果，并且得到的匹配准确率明显高于不进行信息传播.

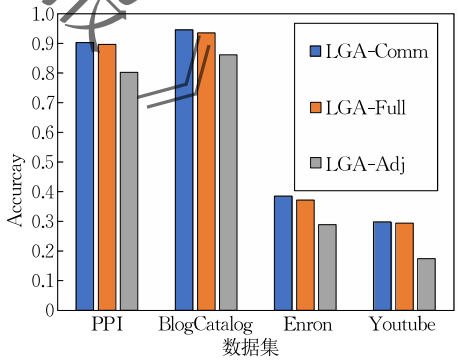


图 10 消融实验结果

本节对基于社区结构进行高阶信息传播和直接迭代信息传播的效率也进行了比较，结果如表 4 所示，基于社区结构进行信息传播耗时不到直接迭代计算的十分之一，具有较大的效率优势. 相比不进行信息传播，节点表示向量只能学习到图网络的局部结构信息，使得节点向量在空间中推理能力下降，其最终的准确率低于另外两种方法.

表 4　不同方式高阶信息传播耗时比较（单位：s）

名称	LGA-Comm	LGA-Adj
PPI	0.87	**0.07**
BlogCatalog	4.92	**0.24**
Enron	5.18	**0.41**
Youtube	59.48	**4.28**

因此，本实验表明图数据的全局结构信息能明显提高图对齐任务的准确率，并且本文提出的基于社区结构进行高阶信息传播在保证有效性的情况下能够大幅度降低学习图全局结构信息所需耗费的时间.

6　结　论

为了解决大规模图对齐任务中的初始种子节点发现难和图表示学习算法效率与有效性难以兼顾的问题，本文提出了一种基于稀疏矩阵分解和个性化随机游走相结合的图表示学习方法，能够对大规模网络进行高效学习，得到的表示向量能够捕获图的全局结构信息，很好地适用于图对齐任务.在此基础上提出一种基于种子节点融合的大规模图对齐算法LGA，与以往利用种子节点监督图表示学习的范式不同，本文所提方法直接在融合图上进行无监督的图表示学习，直接将待匹配的图数据映射到统一的向量空间中，避免了由于表示空间转换导致的匹配准确率较低的问题.最后在多种类型和不同规模的数据上进行了广泛的实验验证，结果表明本文所提方法的能够取得远高于其他方法的匹配准确率和计算效率，能够应用于大规模图数据的高效对齐.

本文方法完全利用图的拓扑结构信息来实现图对齐，具有较广的应用范围.然而，部分图数据包含了丰富的属性信息，对于提高图对齐的准确性具有重要作用，为此，在今后工作中将继续改进本文算法以便结合网络的属性信息进一步提高图对齐的准确率.

参　考　文　献

[1] Marin A, Wellman B. Social network analysis: An introduction. The SAGE Handbook of Social Network Analysis, SAGE, 2011: 11-25

[2] Perc M. Growth and structure of Slovenia's scientific collaboration network. Journal of Informetrics, 2010, 4(4): 475-482

[3] Alm E, Arkin A P. Biological networks. Current Opinion in Structural Biology, 2003, 13(2): 193-202

[4] Liu X, Murata T, Kim K S, et al. A general view for network embedding as matrix factorization//Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM'19). New York, USA, 2019: 375-383

[5] Yin Y, Wei Z. Scalable graph embeddings via sparse transpose proximities//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19). New York, USA, 2019: 1429-1437

[6] Man T, Shen H, Liu S, et al. Predict anchor links across social networks via an embedding approach//Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16). Palo Alto, USA, 2016: 1823-1829

[7] Yu Liang, Gao Lin, Sun Peng-Gang. Research on algorithms for complexes and functional modules prediction in protein-protein interaction networks. Chinese Journal of Computers, 2011, 34(7): 1239-1251(in Chinese)
(鱼亮，高琳，孙鹏岗. 蛋白质网络中复合体和功能模块预测算法研究. 计算机学报，2011，34(7)：1239-1251)

[8] Wu An-Biao, Yuan Ye, Qiao Bai-You, et al. The influence maximization problem based on large-scale temporal graph. Chinese Journal of Computers, 2019, 42(12): 2647-2664(in Chinese)
(吴安彪，袁野，乔百友等. 大规模时序图影响力最大化的算法研究. 计算机学报，2019，42(12)：2647-2664)

[9] Zhang Fu, Yang Lin-Yan, Li Jian-Wei, et al. An overview of entity alignment methods. Chinese Journal of Computers, 2022, 45(6): 1195-1225(in Chinese)
(张富，杨琳艳，李健伟等. 实体对齐研究综述. 计算机学报，2022，45(6)：1195-1225)

[10] Liu L, Cheung W K, Li X, et al. Aligning users across social networks using network embedding//Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16). Palo Alto, USA, 2016: 1774-1780

[11] Klau G W. A new graph-based method for pairwise global network alignment. BMC Bioinformatics, 2009, 10(1): 1-9

[12] Koutra D, Tong H, Lubensky D. Big-Align: Fast bipartite graph alignment//Proceedings of the 2013 IEEE 13th International Conference on Data Mining. Dallas, USA, 2013: 389-398

[13] Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proceedings of the National Academy of Sciences, 2008, 105(35): 12763-12768

[14] Zhang S, Tong H. FINAL: Fast attributed network alignment //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). New York, USA, 2016: 1345-1354

[15] Heimann M, Shen H, Safavi T, et al. REGAL: Representation learning-based graph alignment//Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18). New York, USA, 2018: 117-126

[16] Chu X, Fan X, Yao D, et al. Cross-network embedding for multi-network alignment//Proceedings of the World Wide Web Conference (WWW'19). New York, USA, 2019: 273-284

[17] Trung H T, Van Vinh T, Tam N T, et al. Adaptive network alignment with unsupervised and multi-order convolutional networks//Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE). Dallas, USA, 2020: 85-96

[18] Gao J, Huang X, Li J. Unsupervised graph alignment with Wasserstein distance discriminator//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD'21). New York, USA, 2021: 426-435

[19] Wang C, Zhao Z, Wang Y, et al. DeepMatching: A structural seed identification framework for social network alignment//Proceedings of the 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna, Austria, 2018: 600-610

[20] Zhou F, Liu L, Zhang K, et al. DeepLink: A deep learning approach for user identity linkage//Proceedings of the IEEE Conference on Computer Communications (IEEE INFOCOM 2018). Honolulu, USA, 2018: 1313-1321

[21] Derr T, Karimi H, Liu X, et al. Deep adversarial network alignment//Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM'21). New York, USA, 2021: 352-361

[22] Nguyen T T, Pham M T, Nguyen T T, et al. Structural representation learning for network alignment with self-supervised anchor links. Expert Systems with Applications, 2021, 165: 113857

[23] Zheng C, Pan L, Wu P. JORA: Weakly supervised user identity linkage via jointly learning to represent and align. IEEE Transactions on Neural Networks and Learning Systems, 2022, (1): 1-12

[24] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14). New York, USA, 2014: 701-710

[25] Hirose O. A Bayesian formulation of coherent point drift. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(7): 2269-2286

[26] Wang Y, Peng Q, Wang W, et al. Network alignment enhanced via modeling heterogeneity of anchor nodes. Knowledge-Based Systems, 2021, 250: 109116

[27] Zhu H, Zhou M C, Alkins R. Group role assignment via a Kuhn-Munkres algorithm-based solution. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2011, 42(3): 739-750

[28] Chiang W L, Liu X, Si S, et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19). New York, USA, 2019: 257-266

[29] Yang R, Shi J, Xiao X, et al. Homogeneous network embedding for massive graphs via reweighted personalized PageRank. Proceedings of the VLDB Endowment, 2020, 13(5): 670-683

[30] Musco C, Musco C. Randomized block Krylov methods for stronger and faster approximate singular value decomposition. Advances in Neural Information Processing Systems, 2015, 2015: 1396-1404

[31] Li Y, Wang Y, Zhang T, et al. Learning network embedding with community structural information//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 2937-2943

[32] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826

[33] Newman M E J. Detecting community structure in networks. The European Physical Journal B, 2004, 38(2): 321-330

[34] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008 (10): P10008

[35] Benson A R, Gleich D F, Leskovec J. Higher-order organization of complex networks. Science, 2016, 353(6295): 163-166

[36] Steinhaeuser K, Chawla N V. Community detection in a large real-world social network//Social Computing, Behavioral Modeling, and Prediction. Boston, USA: Springer, 2008: 168-175

[37] Sah P, Singh L O, Clauset A, et al. Exploring community structure in biological networks with random graphs. BMC Bioinformatics, 2014, 15(1): 1-14

[38] Yoon M, Jung J, Kang U. TPA: Fast, scalable, and accurate method for approximate random walk with restart on billion scale graphs//Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE'18). Paris, France, 2018: 1132-1143

[39] https://github.com/THUDM/ProNE/tree/master/data

[40] http://snap.stanford.edu/data/email-Enron.html

[41] http://snap.stanford.edu/data/soc-Slashdot0811.html

[42] http://snap.stanford.edu/data/com-Youtube.html

[43] http://snap.stanford.edu/data/soc-Pokec.html

[44] Pedarsani P, Figueiredo D R, Grossglauser M. A Bayesian method for matching two similar graphs without seeds//Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello, USA, 2013: 1598-1607

[45] Zhang S, Tong H, Tang J, et al. Incomplete network alignment: Problem definitions and fast solutions. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(4): 1-26

[46] Wang C, Wang Y, Zhao Z, et al. Credible seed identification for large-scale structural network alignment. Data Mining and Knowledge Discovery, 2020, 34(6): 1744-1776

WANG Chen-Xu, Ph. D., associate professor. His current research interests include graph data mining, artificial intelligence and network security.

ZHOU Jun-Ming, M. S. His current research interests include graph alignment and network representation learning.

JIANG Pei-Jing, M. S. candidate. His current research interests include graph alignment and network representation learning.

Background

The topic studied in this paper is graph alignment in the field of graph data mining. Graph alignment has long been studied because of its important applications in various fields such as social network analysis, bioinformatics, and computer vision. At present, researchers focus on employing graph representation learning models for graph alignment with development of artificial intelligence.

In this paper, we propose a large-scale unsupervised graph alignment framework based on topological structure representation learning. We first select a typical subgraph from each of the graphs as a candidate set of seed nodes. The local topological information is used to retrieve a set of highly reliable seed node pairs. Then, we use the seed nodes to fuse the matching graphs, and propose an efficient unsupervised representation learning algorithm to map the fused graph into a unified vector space. Finally, large-scale graph alignment is realized based on the learned node vectors. Compared with existing methods, the proposed approach uses the least time in large-scale graph alignment tasks and achieves the best performance of alignment accuracy. Moreover, the structural differences of graphs have limited impacts on the performance of the proposed method.

The research presented in this paper is supported in part by the National Natural Science Foundation of China (No. 62272379), the Natural Science Basic Research Plan in Shaanxi Province (No. 2021JM-018), the National Key R&D Program of China (No. 2021YFB1715600), and the Fundamental Research Funds for the Central Universities (No. 1191320006).