《metapath2vec: Scalable Representation Learning for Heterogeneous Networks》

《metapath2vec:异构网络的可扩展表示学习》

**摘要：**

1. **本文的背景：**研究者们发现基于同质图的图表征学习算法，例如Deepwalk，Node2Vec等并不能很好地直接应用到点和边有多个类型的异质图上。而这主要是由于这些同质图算法并不考虑节点的类型，导致它们往往有如下两方面的缺点：1）容易偏向于出现频率高的节点类型；2）容易偏向于连接相对集中的节点。因此，本文旨在解决上述网络异构所带来的特殊挑战，从而进一步扩展图表征学习的应用范围。
2. **本文的贡献：**提出了一种专门处理异质图的图嵌入学习框架——metapath2vec，以及其改进版本metapath2vec++，有效保持了网络结构在语义上的相关性。
3. **主要创新点：**首次正式定义了Heterogeneous Network上的表示学习过程；提出了一种基于元路径的随机游走方法以适用于异构图网络。
4. **实验结果：**本文通过大量的实验证明了本文所提出的方法在各种异构网络挖掘任务中的有效性和可扩展性，如节点分类（相对于基准提高了35%—319%）和节点聚类（相对于基准提高了13%—16%）。此外，还证明了metapath2vec和metapath2vec++能自动发现异构网络中不同类型节点之间的内部语义关系，而现有工作是无法发现的。

**问题定义：**

异构图的定义：

给定一个图，对于其中每个节点和边都有对应的映射函数：，。其中，和分别表示节点和边的类型集合，且。即，图网络的节点类型与边类型的总量大于2。

异构图表示学习的定义：

给定一个异构图，其任务是学习到所有节点的嵌入表示，其中，而嵌入表示能有效捕捉到不同类型节点的结构和语义关系。

**注意：**虽然异构图上的节点类型不同，但是不同类型的节点的嵌入向量会映射到同一个特征空间中，即不同类型的节点的嵌入向量的维度都是。此外，由于网络异构性的存在，传统的基于同构网络的节点嵌入表征方法很难有效地直接应用在异构网络上。

**异质skip-gram：**

传统的skip-gram模型：

给定一个网络，其目标是使网络的局部结构概率最大化，即：

其中，代表节点的上下文节点，而这种节点的定义方式有多种，比如中利用随机游走得到的上下文节点。而代表在已知节点的情况下，上下文节点存在的概率，这种概率通常用向量内积来实现：

而为了将异构网络的结构引入到上述的skip-gram方法中，本文提出了异质skip-gram的概念，即在给定节点的情况下，最大化其异质上下文节点出现的概率，即：

其中，代表节点的第种类型的上下文节点；代表在已知节点𝑣的情况下，上下文节点存在的概率，而这种概率在本文也是用向量内积+𝑠𝑜𝑓𝑡𝑚𝑎𝑥来实现的：

此外，与传统的skip-gram中一致，为了解决每次更新时都要对所有节点进行计算的问题，异质skip-gram也引入了负采样策略。给定负样本数，则优化目标可进一步表示为：

其中，代表负采样中样本的预定义分布；激活函数；代表从负样本集中抽取的负样本。

**基于元路径的随机游走：**

同和等方法类似，也需要得到节点的游走序列，然后使用上述异质skip-gram模型进行优化。而为了在异构图上得到有效的游走序列，本文提出了Meta-Path-Based Random Walks，即基于元路径的随机游走。

具体来讲，一个元路径方案被定义为：

**注意：**元路径中相邻节点的类型是不一样的，即和属于不同类型的节点，而表示两个节点之间的连接关系。

因此，节点和之间的复合关系可进一步定义为：

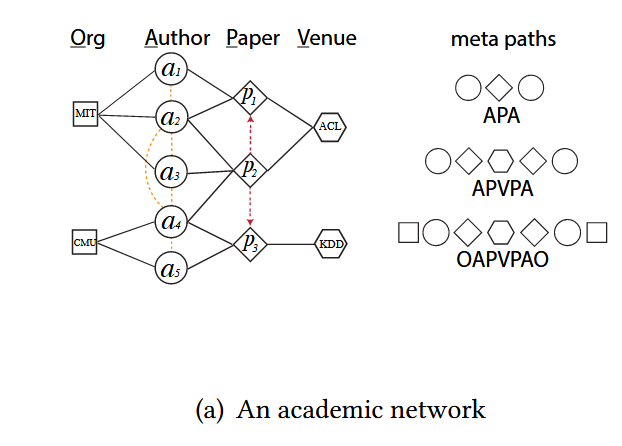
此外，与方法中一致，基于元路径的随机游走也给出了每一步的转移概率——给定一个异构网络和一个元路径方案，步骤处的转移概率定义如下：

假设当前节点为，那么下一个节点的选择策略为：

1. 如果与之间存在边，且的节点类型与元路径方案中的定义一致，那么这些节点被选中的概率是均等的，也就是可选节点个数。
2. 如果与之间存在边，但的节点类型与元路径方案中的定义不一致，那么则不会选择该节点，即概率为0。
3. 如果与之间不存在边，则不会选择该节点。

而除了上述三点原则外，元路径还需遵循对称原则，也就是路径中第一个节点和最后一个节点的类型应该一致，即：

于是，基于元路径的随机游走策略就能确保不同类型节点之间的语义关系能够被恰当地整合到skip-gram中。以下图为例：



假设在一个学术网络中，表示作者节点，表示论文，表示作者所属组织，表示期刊会议。而根据上述元路径的定义，可以有以下几种简单的元路径：

1. ：一篇论文中两个作者的合著关系。
2. ：两个作者在同一期刊会议上发表了论文。
3. ：两个分别所属组织的作者在同一期刊会议上发表了论文。

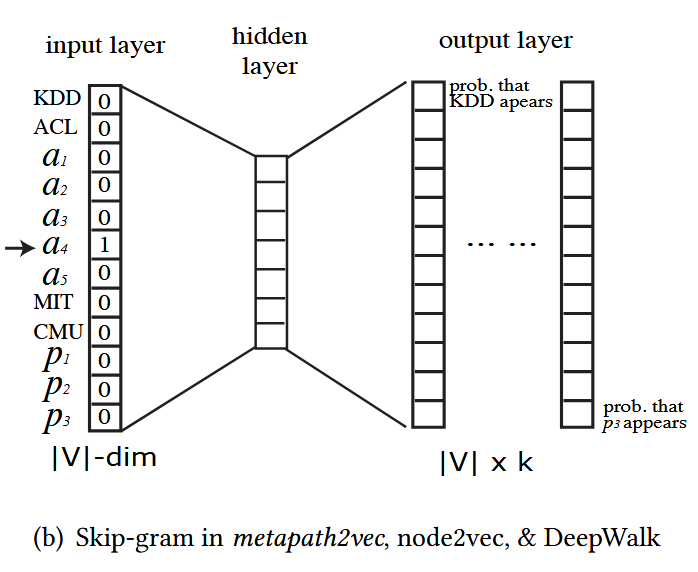
在传统的随机游走过程中，从节点过渡到节点上的walker的下一步可以是它周围所有类型的节点，即、、、、和。然而，在元路径方案“”下，下一步选择的节点则会偏向于论文节点（严格遵循此路径的语义）。

**metapath2vec++：**

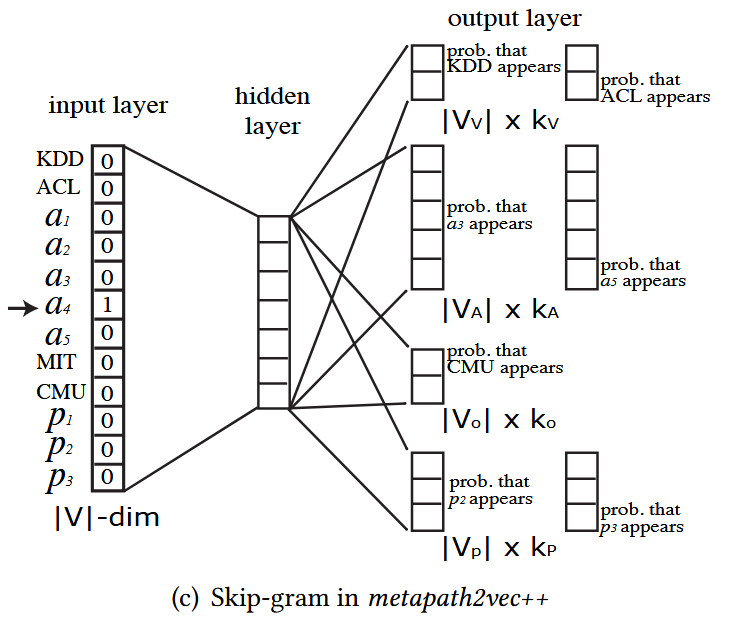
实际上，上述异质skip-gram模型和基于元路径的随机游走方案就构成了本文所提出的方法，而所谓的方案则是在的基础上对优化目标做出了进一步的改进，具体细节如下：

已知，中的优化目标为：

其中：



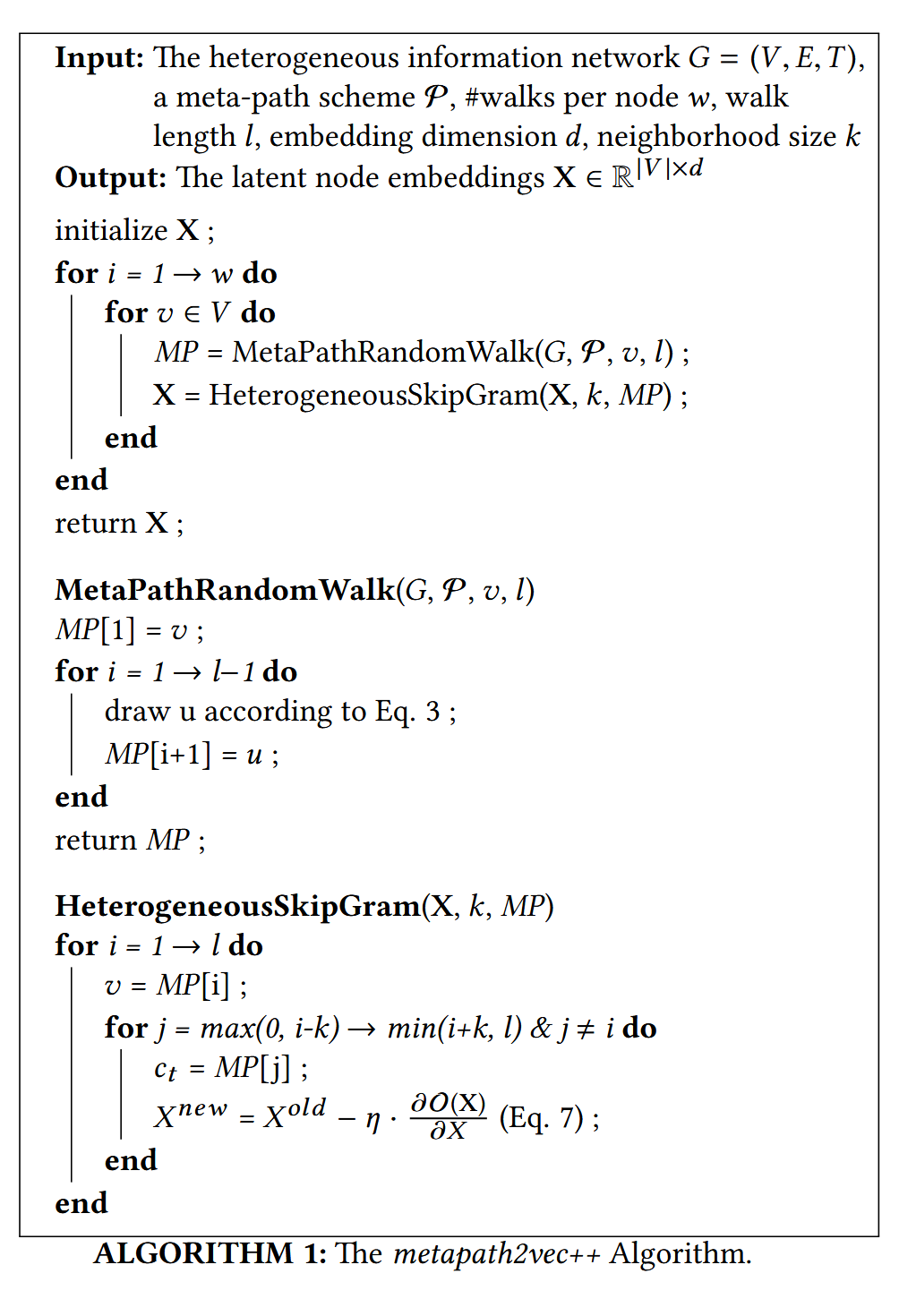
根据要求，我们需要构造节点的邻居函数（上下文邻居）。而从元路径游走策略可知，中是根据元路径方案来确定的。然后，在进行计算时，我们会考虑所有类型的节点，而这种方式忽略了中节点的类型信息，如上图所示。即，在进行负采样时，不会区分节点类型，所有类型的节点都有可能被采样到，即上面式子中的，而作者认为这是不合理的。因此，改进后的结果如下：



这样，就使得我们在计算节点向量间的内积并进行时，会根据不同类型的节点的上下文进行归一化，如上图所示。因此，的优化目标就可以进一步定义为：

从这里我们也可以看出被提出的根本原因：之前所有方法都是基于同质图，同质图中所有节点类型都一致，因此进行负采样时不需要额外考虑节点类型，而异质图中节点类型不一致，因此自然而然地就会想到负采样时区分节点类型。

**metapath2vec++算法伪代码：**



**实验结果及分析：**

数据集：

两个数据集——Aminer、DBIS，每个数据集中节点的种类为3。

Baselines：

DeepWalk、node2vec、LINE、PTE、Spectral Clustering、Graph Factorization

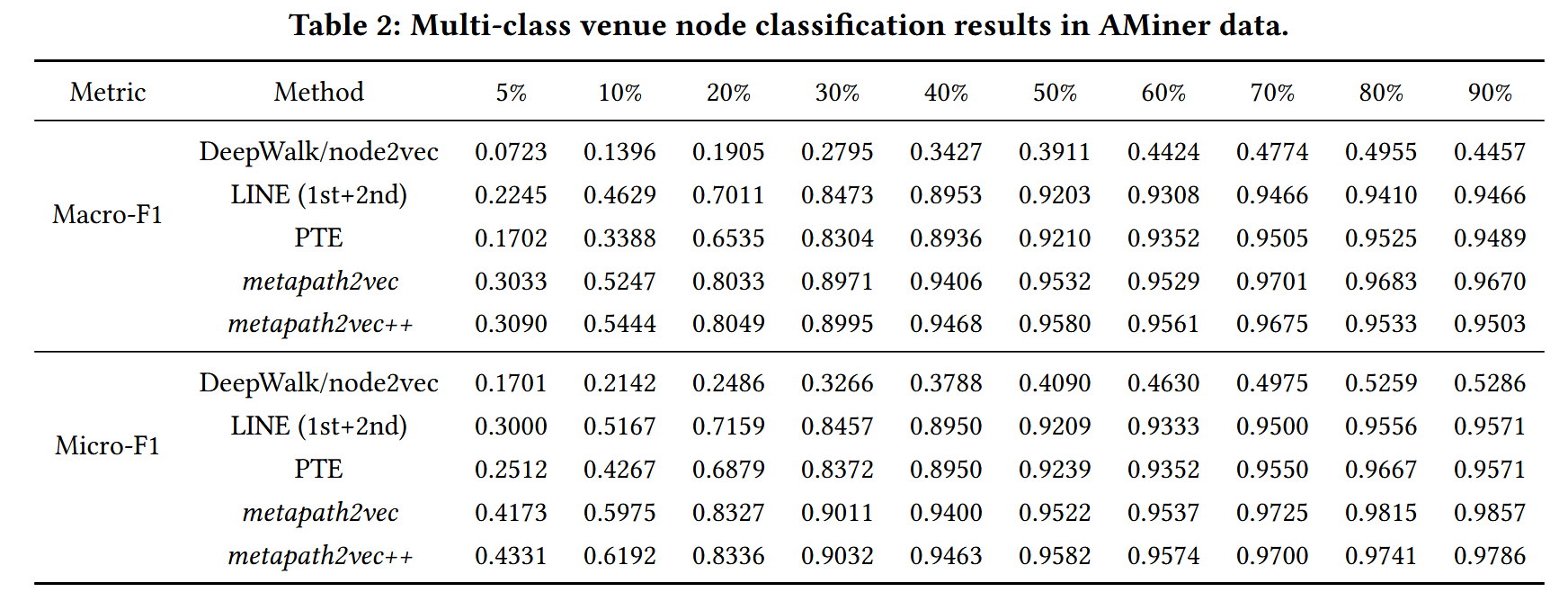
多个任务：

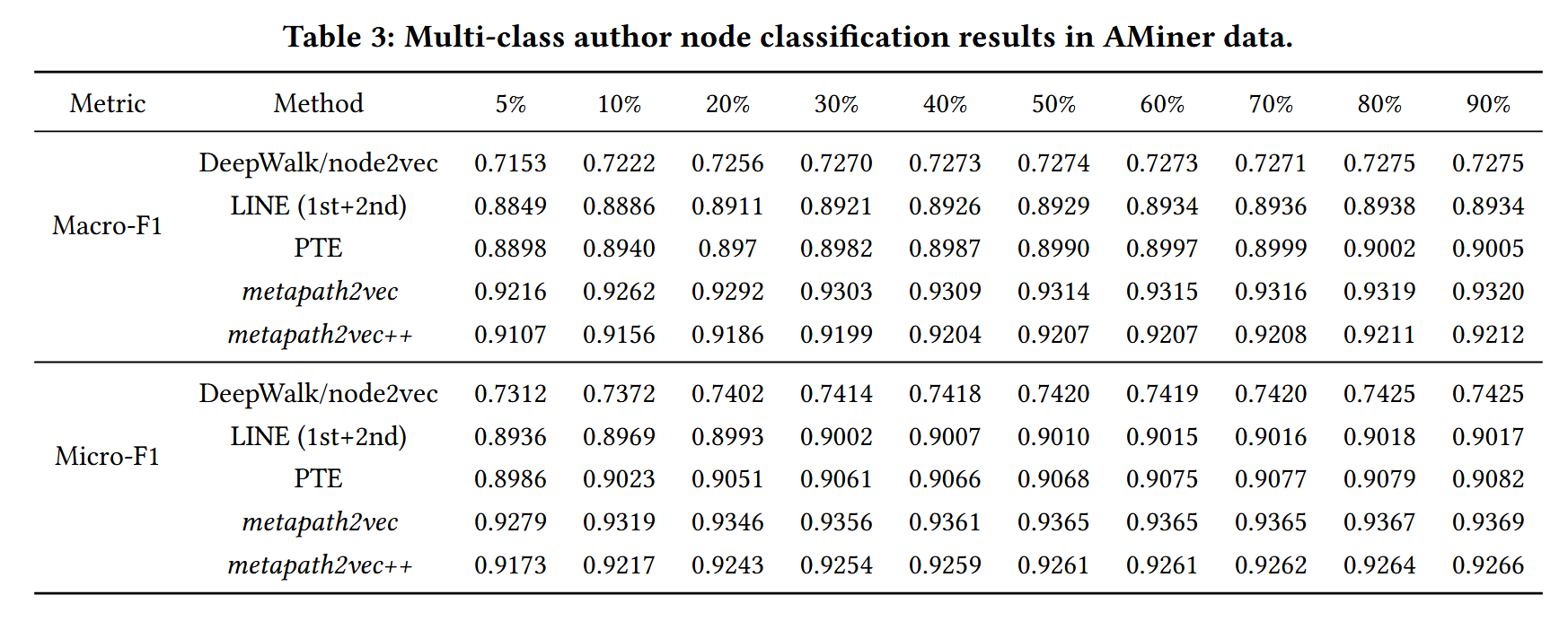
节点分类（Multi-Class Classification）、节点聚类（Node Clustering）、相似性（Case Study：Similarity Search）、可视化（Case Study：Visualization）、参数实验（Scalability）

实验参数：

1. 每个节点的随机游走次数：1000；
2. 每次随机游走的长度：100；
3. 嵌入向量维度：128；
4. 邻域大小：7；
5. 负采样大小：5；
6. 学术网络中涉及的元路径方案有和。

节点分类实验结果：

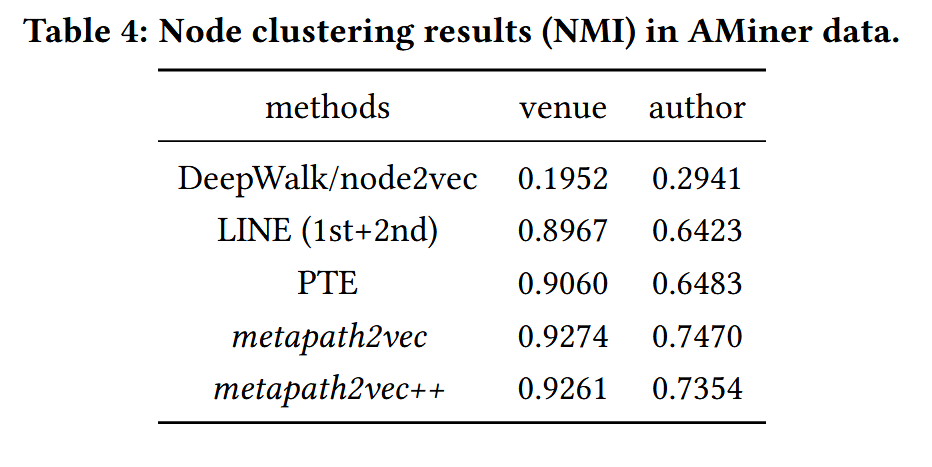




如图所示，可以看到metapath2vec和metapath2vec++相比基准算法表现要更好。

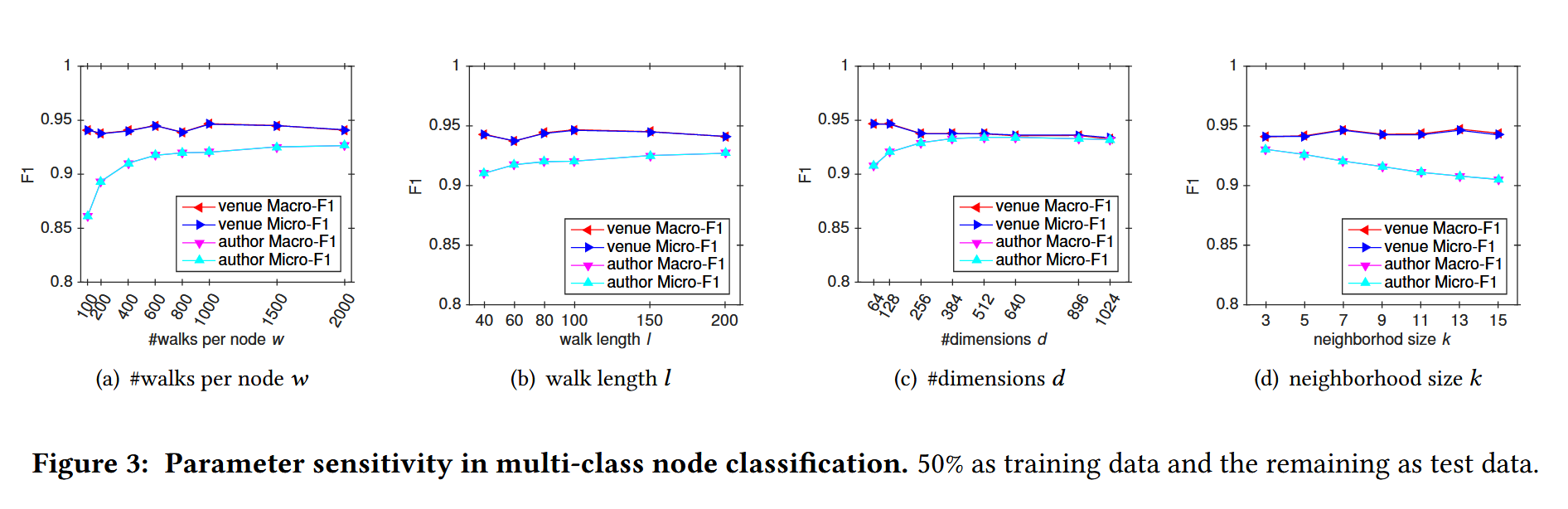
节点聚类实验结果：

**说明：**采用NMI指标（归一化互信息）来衡量，值越大越好。而聚类结果是在得到节点表征后运行均值算法得到的。

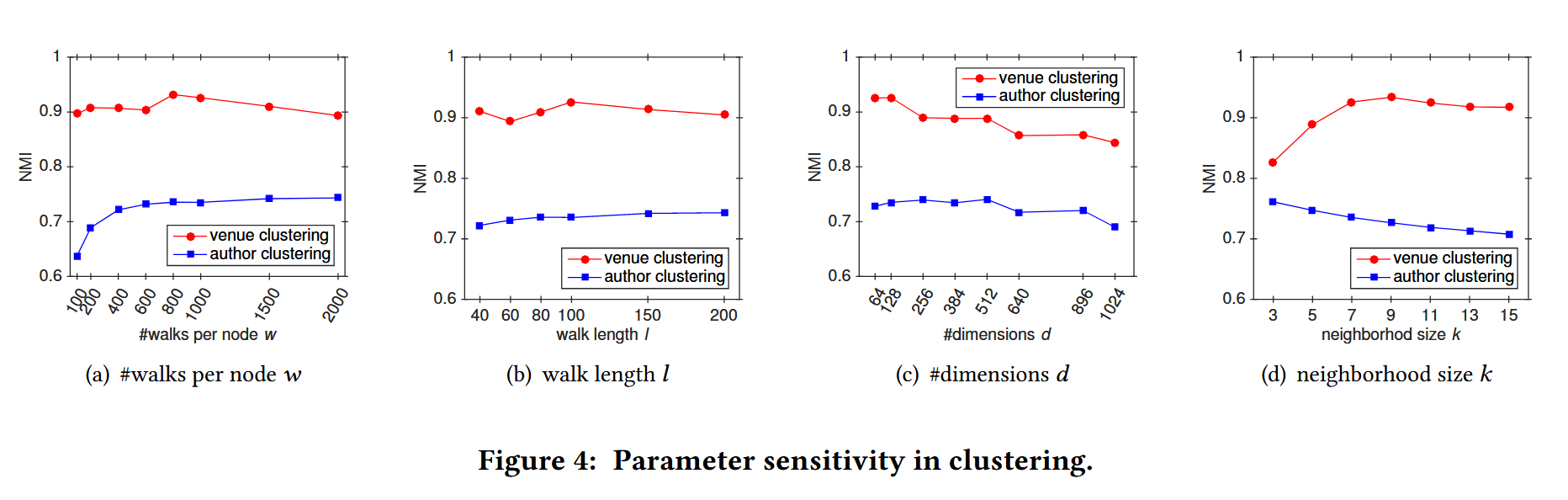


如图所示，可以看到metapath2vec和metapath2vec++相比基准算法表现要更好。

参数敏感性分析：



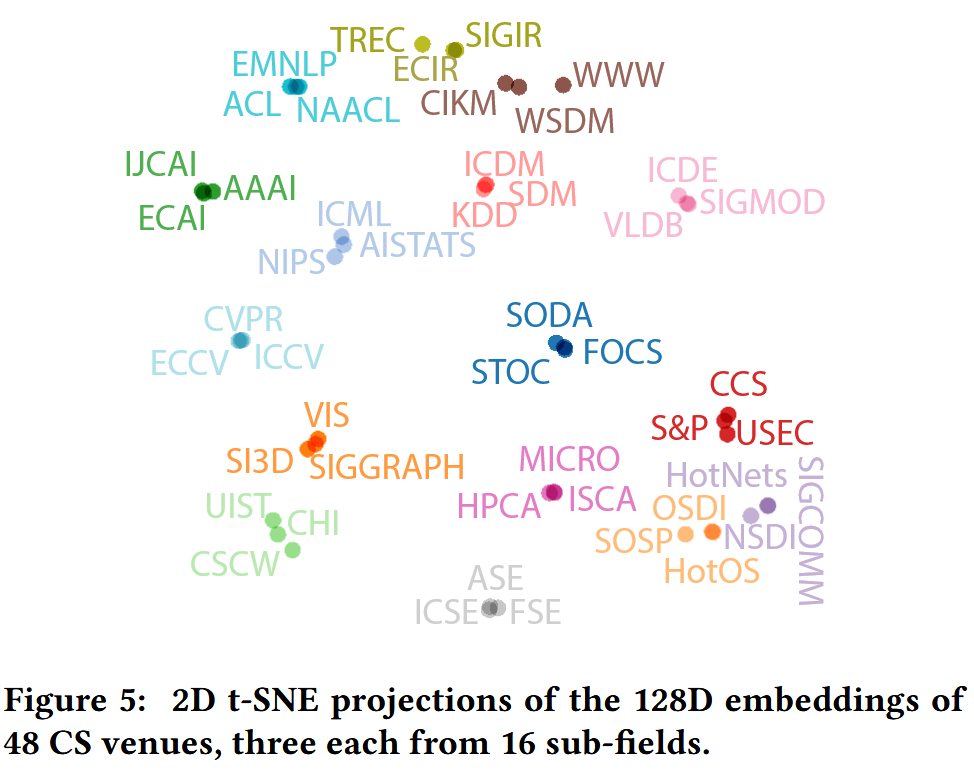
如图所示，可以看到在节点分类任务中，遍历次数在1000时得分最高，遍历深度在100时得分最高，嵌入维度在128时效果最好，而邻域大小需要视情况而定。



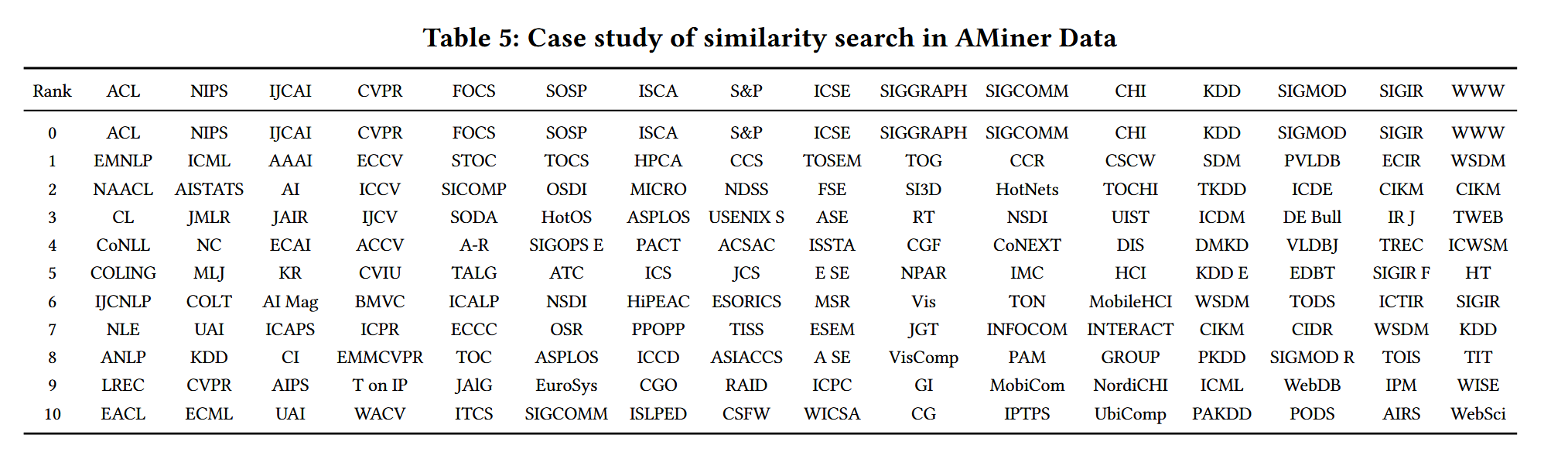
如图所示，可以看到在节点聚类任务中，遍历次数在800时效果最好，遍历深度在100时效果最高，嵌入维度在128时效果最好，而邻域大小仍需要视情况而定。

可视化结果：

下图是会议节点降维的结果，相同类型节点相同颜色。

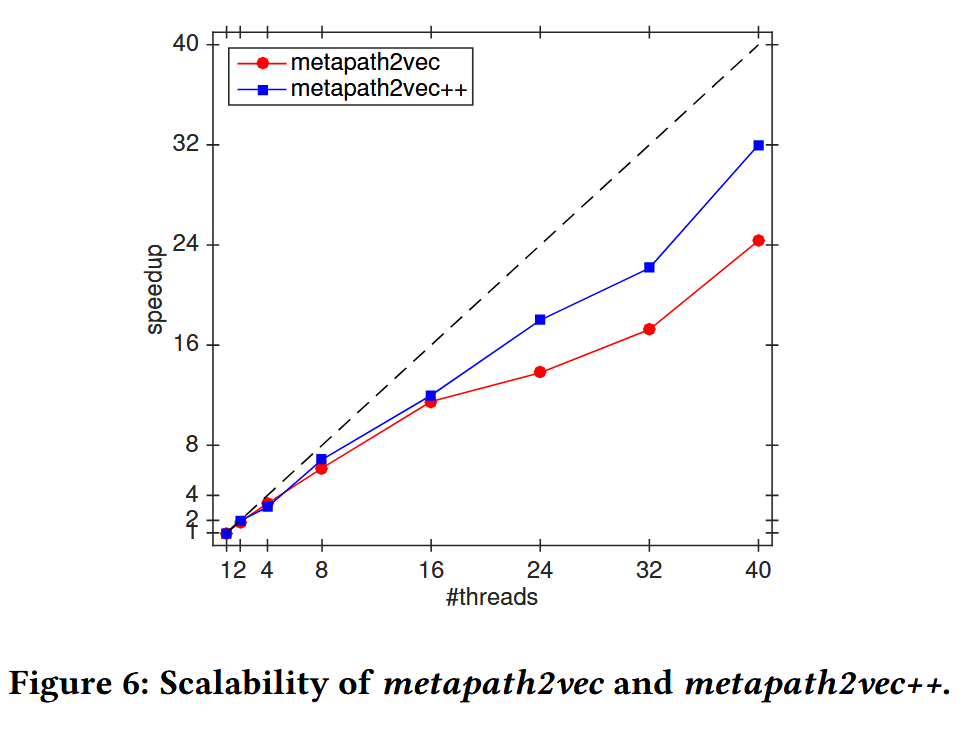


会议节点相似性排名（越相关越靠上）：



可扩展性：

下图是显示当可用资源（线程）变化，模型性能指标的提升效果。而我们可以看到模型的可扩展性不错，基本呈线性。



**结论：**

本文提出了异构图嵌入模型和。具体而言，首先定义了元路径方案及其对应的随机游走策略，该策略能够捕获不同类型节点及关系的结构和语义相关性。然后，再将随机游走序列输入到本文提出的异质skip-gram中以得到所有类型节点的嵌入表示。而需要注意的是，中skip-gram的负采样策略与同质skip-gram一致，考虑了所有节点；而中skip-gram在进行负采样时只考虑与上下文节点同类型的节点。此外，大量实验还表明，和学习的潜在特征表示能够改进各种异构网络挖掘任务，如相似性搜索、节点分类和聚类等。

**对本文的感悟：**

这篇论文首次将random walk + skip-gram的框架拓展到了异构图的研究领域，并且还有力证明了基于异构图的随机游走算法能够表达不同类型节点之间的语义和结构关联。总之，本文不仅拓展了深度学习研究图的类型，还进一步将深度学习领域的模型引入到了更广泛的图学习任务之中，也为后续类似HAN、GTN模型的提出创造了坚实的理论基础。