《Node2Vec：Scalable Feature Learning for Networks》

**摘要：**

1. **本文的背景：**自从DeepWalk论文提出了利用Random Walk来随机生成序列化数据供给Skip-gram模型做类似词嵌入的图节点嵌入训练后，图表示学习领域的研究不断取得重大进展。而随着实验的进一步深入，研究者们也发现了DeepWalk存在一定的局限性，为此本文的作者们在DeepWalk基础上又提出了一种更为有效的方法。
2. **本文的贡献：** 提出了一种名为Node2Vec的**有偏随机游走算法**来生成图网络中的序列化数据进行图嵌入。并且，首次将图嵌入方法应用到了**连接预测**任务之中。
3. **主要创新点：**通过使用有偏的二阶随机游走策略来产生序列化数据。更详细地说，通过设置超参数p，q用有弹性、有偏的随机游走策略有效平衡了来自local（BFS）和global（DFS）的节点网络结构信息。
4. **实验结果：**证明了Node2Vec在来自不同领域的几个真实世界网络中的**多标签分类**和**连接预测**任务上都明显优于现有的最先进的技术。

**Node2Vec的优化目标：**

假设是将顶点映射为embedding向量的映射函数，而对于图中的每个顶点，我们定义 为通过采样策略采样得到的顶点的近邻顶点集合。因此，Node2Vec优化的目标就是在给定每个顶点的条件下，令其近邻顶点出现的概率最大。即：

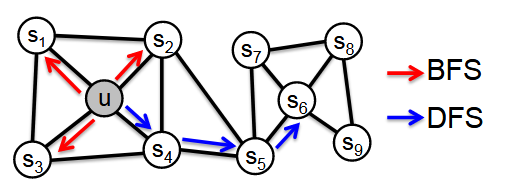
而为了方便上述公式易于处理，论文提出了两个标准假设：

1. **条件独立性假设——**假设在给定源顶点的情况下，其近邻顶点出现的概率与近邻集合中其余的顶点无关。即：
2. **特征空间对称性假设——**一个顶点在作为源顶点和作为近邻顶点时将共享同一套embedding向量。即：

因此，根据以上的两个假设条件，最终的目标函数可以简化为：

其中，配分函数（归一化分母）。而由于的计算代价过高，所以在实际训练中会采用**负采样方法**对其进行优化。

**Node2Vec的采样策略：**



如图所示，深度优先（BFS）和广度优先（DFS）策略代表着两种最极端的采样策略，其中DFS能体现网络节点的同质社群特征（homophily community），而BFS能体现网络节点的功能角色（结构）特征（structural equivalence）。但是，在真实世界中的网络往往是这两种特征的相互融合。

因此，Node2Vec通过引入两个超参数p和q来控制随机游走策略，使其能够更好地应用到不同的网络图之中。

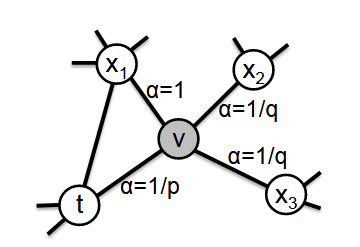
给定当前顶点，访问下一个顶点的概率为：

其中，是节点和之间的非规范化转移概率，是归一化常数。而在论文中，作者通过采用二阶随机游走对进行了定义，即：

其中，代表连接边的连接权重（无权图则默认），代表平滑搜索策略，具体公式如下：

其中，代表上一节点，代表下一节点，代表返回参数，代表进出参数，代表顶点和顶点之间的最短路径距离。

**超参数和对有偏游走策略的影响：**



如上图所示，它描述的是当从上一节点访问到当前节点时，决定下一个访问节点时每个顶点对应的。而从前面公式中可知，受超参数和的控制，具体而言：

·Return parameter，

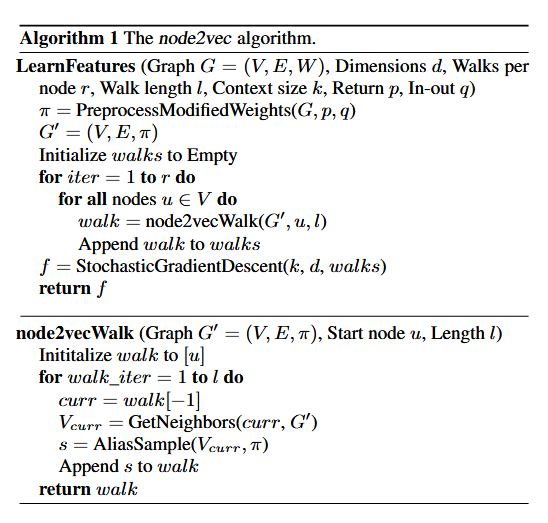
参数控制着重复访问刚刚访问过的顶点的概率。并且，注意到仅作用于的情况，而代表着下一顶点就是访问当前顶点之前刚刚访问过的顶点。因此，若较高，则访问刚刚访问过的顶点的概率会变低，反之则会变高。

·In-out parameter，

参数控制着游走是向外还是向内的，若，则随机游走倾向于访问和接近的顶点（更偏向于BFS）。若，则随机游走倾向于访问远离的顶点（更偏向于DFS）。

**总之，通过调节和，我们可以平滑地调整采样策略，以探索图网络的不同特性。**

**Node2Vec算法实现：**

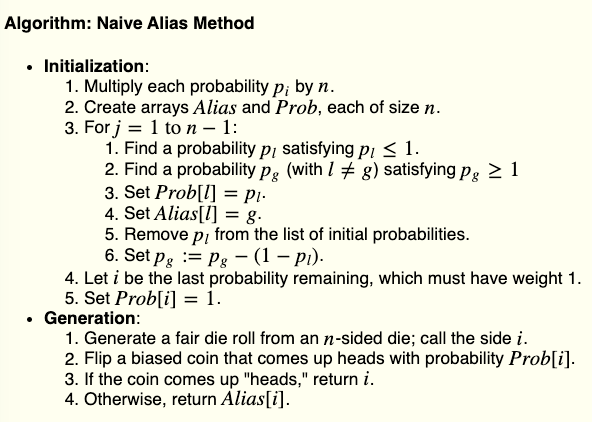


Node2Vec的算法伪代码如上图所示，其中表示生成序列的起始节点，代表上一节点，代表当前节点，代表下一节点，代表随机游走序列的节点个数（长度），代表上一节点的邻居节点，代表从中采样的节点个数，代表嵌入向量的维度，和代表游走策略的超参数，代表每个节点进行随机游走的次数。

整体的特征学习算法主要分为3步：

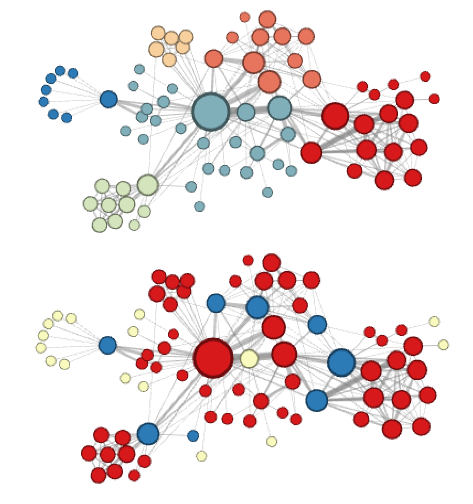
1. 生成随机游走的采样策略；
2. 对每个节点生成个随机游走序列；
3. 采用Skip-Gram训练得到节点的嵌入表示。

而在Node2VecWalk算法实现中，作者还用到了**AliasSample采样优化算法**，它的算法实现伪代码如下图所示。简单来说，Alias-Sample通过利用内存空间（预处理，）来换取运行时间，使得采样的时间复杂度（效率）能够达到，大大节省了采样所需的时间。并且，它在大量反复抽样的情况下优势更为突出，能有效将离散分布的抽样转化为均匀分布的抽样。



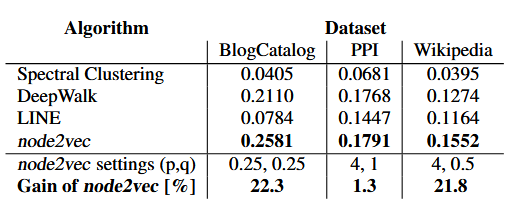
**实验部分**

1. **案例研究：《悲惨世界》网络**

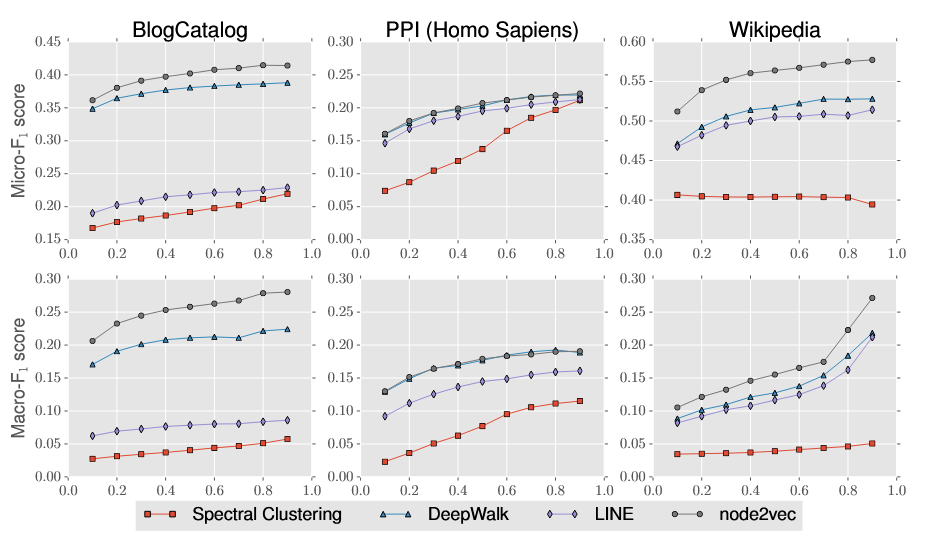


如图所示，这是雨果的《悲惨世界》的人物社交关系图。其中，顶部网络的节点颜色展示了节点之间的**同质社群特征**（对应超参数p大q小）；底部网络的节点颜色代表了节点之间的**功能角色特征**（对应超参数p小q大）。可见，不同的超参数p和q能发掘网络中不同的特征信息。

1. **多标签分类任务**

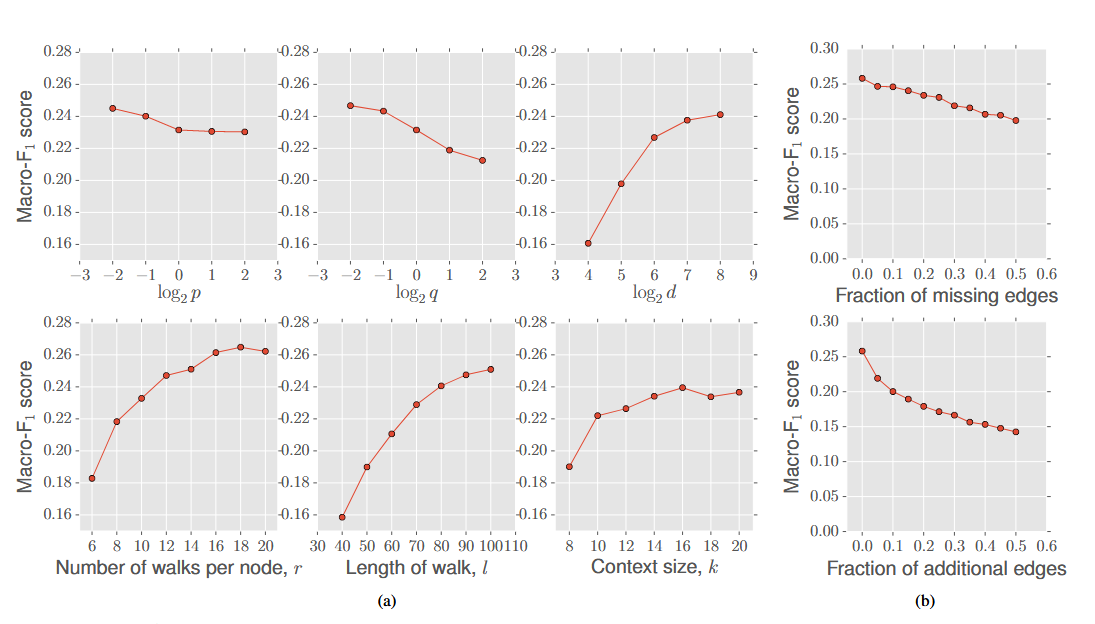


如图所示，在Macro-F1得分上，**Node2Vec在不同的数据集上都取得了相比其他方法的最好成绩**。并且，在部分数据集上，Node2Vec甚至能达到20%的提升。



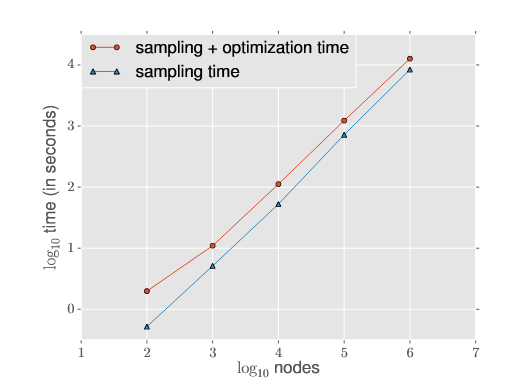
如图所示，不同训练的标记数据量对各算法的性能评估。其中，轴表示标记数据的比例，而顶部和底部中的轴分别表示Micro-F1和Macro-F1分数。而从中我们可以发现，**DeepWalk和Node2Vec在 PPI 上提供了相当的性能，但在所有其他网络中，Node2Vec表现最佳。此外，标记数据越多，算法效果越好。**

1. **调参对比实验和扰动分析**



如图所示，不同参数的大小都会影响最终的F1得分，而通过随机剔除连接和随机增加连接的实验（如图(b)所示），我们也可以发现Node2Vec算法的**鲁棒性较好**，能有效对抗扰动（缺失和噪声）。

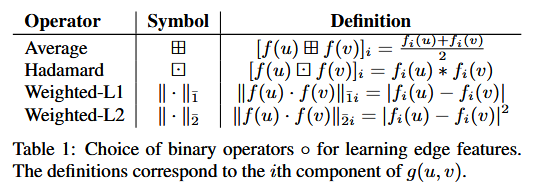
1. **可扩展性**



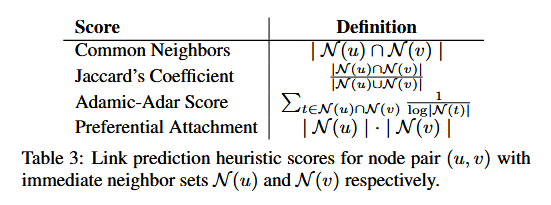
如图所示，随着网络中节点个数的增长，算法的运行时间是和节点数呈线性相关的。因此，Node2Vec的算法**可扩展性较好**。

1. **连接预测**

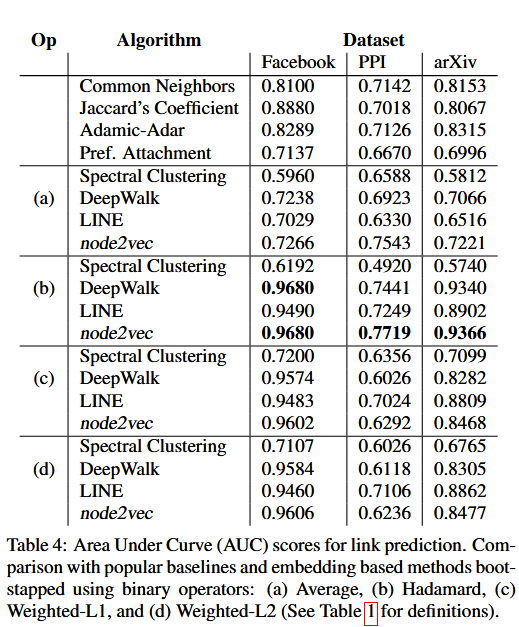
**说明：**在本篇论文中，作者通过利用定义的二元运算符来学习节点之间的边特征。并且，还将这些方法与一些常见的启发式方法进行了比较。具体的二元运算符如下表所示：



而常见的启发式方法如下表所示：



而实验的最终结果如下：



可见，采用Hadamard方式的Node2Vec和DeepWalk算法均取得了最好的结果。

**总结：**

DeepWalk存在的缺点：

1. 用完全随机游走来训练节点嵌入向量（刚性策略效果并不好）。
2. 仅能反映**相邻节点**的**社群**相似信息。
3. 无法反映节点的**功能角色**相似信息。

Node2Vec图节点嵌入算法：

* Node2Vec解决了**图嵌入**问题，将图中的每个节点映射为一个向量 (嵌入)。
* 向量 (嵌入) 包含了节点的语义信息 (相邻社群和功能角色)。
* 语义相似的节点，向量 (嵌入) 的距离也近。
* 向量 (嵌入) 可用于后续的分类、聚类、Link Prediction、推荐等任务。
* 在DeepWalk完全随机游走的基础上，Node2Vec增加了p、q参数，实现了有偏随机游走。而不同的p、q组合，对应了不同的探索范围和节点语义。
* DFS深度优先搜索，相邻的节点，向量 (嵌入) 距离相近。
* BFS广度优先搜索，相同功能角色的节点，向量 (嵌入) 距离相近。
* 而DeepWalk可以看成是Node2Vec在的特例。

总之，相比其他算法而言（例如DeepWalk和LINE），Node2Vec中的搜索策略既灵活又可控，我们可以通过调节参数p和q来探索网络邻域，并且这些搜索参数具有更直观的解释性。此外，从实际的角度来看，Node2Vec还是可扩展的，且对扰动也具有良好的鲁棒性。

**对于本文的感悟：**

总的来说，Node2Vec是在DeepWalk的基础上发展出来的，并且成功地解决了DeepWalk所存在的一些缺陷，而这些有效的改进离不开作者们敏锐的洞察力和深入浅出的思考。从整体来看，Node2Vec仅仅是对DeepWalk做出了一点优化，但优化的效果却好的出人意料。或许论文的创新点就是这样，不断地在前人的基础上发掘问题，解决问题，乃至更近一步、产生质变。