# 图嵌入

## def

图嵌入是一种将图数据（通常为高维稠密的矩阵）映射为低维稠密向量的过程，能够很好地解决图数据难以高效输入机器学习算法的问题。

嵌入的过程中有一种共识：向量空间中保持连接的节点彼此靠近。

总的来说大致可以将图上的嵌入分为两种：节点嵌入和图嵌入。当需要对节点进行分类，节点相似度预测，节点分布可视化，一般采用节点的嵌入；当需要在图级别上进行预测或者预测整个图结构，我们需要将整个图表示为一个向量。

## why

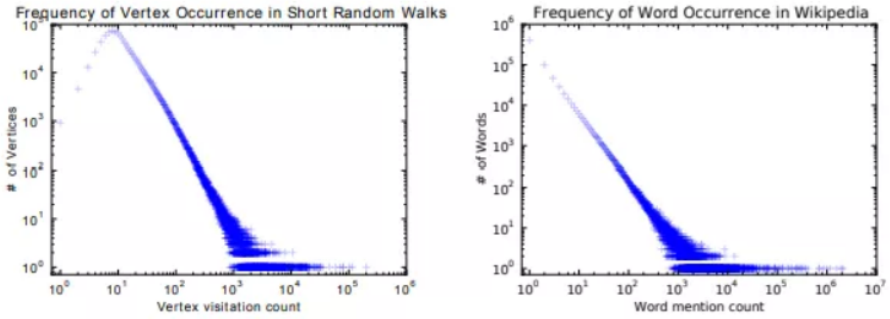
- 在graph上直接进行机器学习具有一定局限性
- 图嵌入能够压缩数据
- 向量计算比直接在图上操作更简单快捷

图嵌入满足需求：

- 属性选择：表示拓扑，节点连接和节点邻域
- 可扩展性：嵌入能够处理大型图，并且高效
- 嵌入维度的把控：维度越大保留信息越多，时间空集复杂度越大；维度越小保留信息也少，时间空集复杂度越小

## method

节点嵌入借鉴了word2vec的方法，因为图中的节点和与语料库中的单词都满足幂律分布。




(a) YouTube Social Graph (b) Wikipedia Article Text

Figure 2: The distribution of vertices appearing in short random walks (2a) follows a power-law, much like the distribution of words in natural language (2b).

（幂律分布：在预料库中出现的频率越高的单词数量越少，多数都是频率低的词；在网络中，在短随机游走序列出现的频率越高的节点数量越少，多数都是偏孤立的节点）

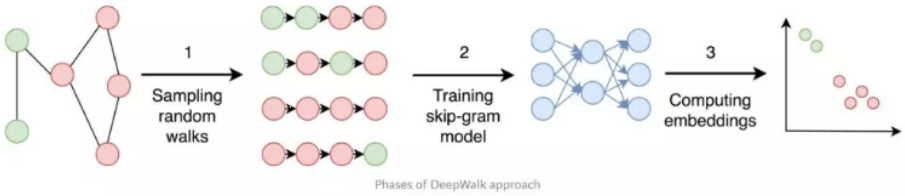
### 节点嵌入方法

1. DeepWalk

   DeepWalk通过随机游走学习出一个网络的表示。随机游走起始于选定节点，然后从当前节点移至随机邻居，并执行一定的步数。

   - 采样：通过随机游走对图上的节点进行采样,在给定的时间内得到一个节点构成的序列，论文研究表明从每个节点执行32到64次随机遍历就足够表示节点的结构关系

- 训练skip-gram：机游走与word2vec方法中的句子相当。文本中skip-gram的输入是一个句子，在这里输入为自随机游走采样得到了一个序列，进一步通过最大化预测相邻节点的概率进行预测。通常预测大约20个邻居节点-左侧10个节点，右侧10个节点；
- 计算嵌入



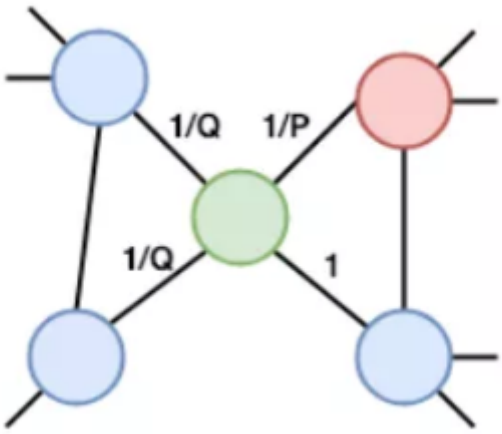
Phases of DeepWalk approach

DeepWalk通过随机游走获取节点的局部上下文信息但不能保留节点的局部关系。

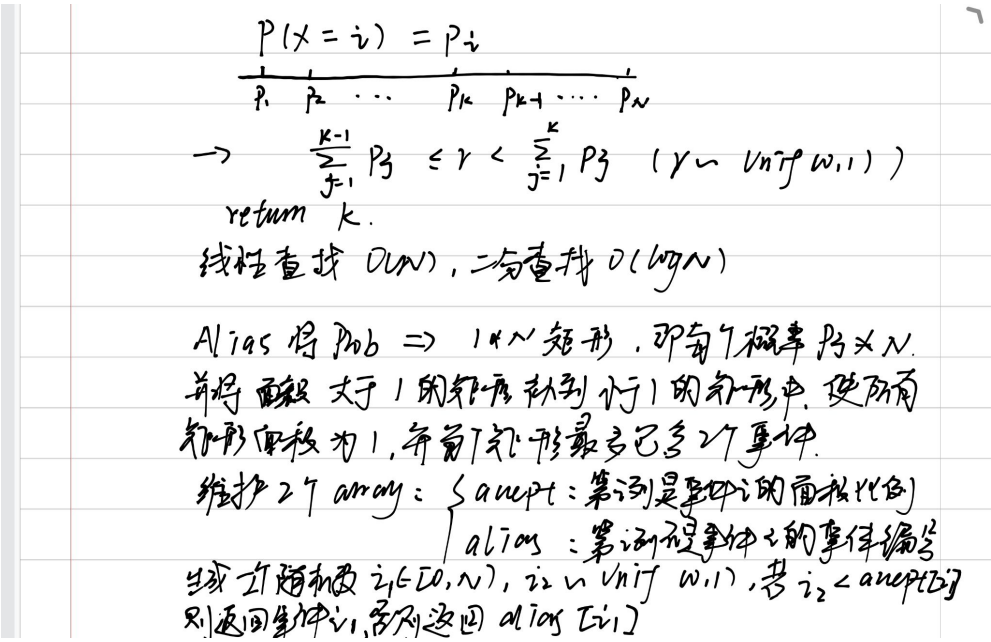
2. Node2vec

Node2vec是DeepWalk的改进版，定义了一个bias random walk的策略生产序列，仍然用skip-gram去训练。

该算法引入了参数P和Q，参数Q关注随机游走中未发现部分的可能性，即控制着游走是向外还是向内，参数P控制了随机游走返回到前一个节点的概率。也就是说，参数P控制节点局部关系的表示，参数Q控制较大邻域的关系。




Node2vec在顶点路径采样时用了时间复杂度为O(1)的alias采样。



3. SDNE

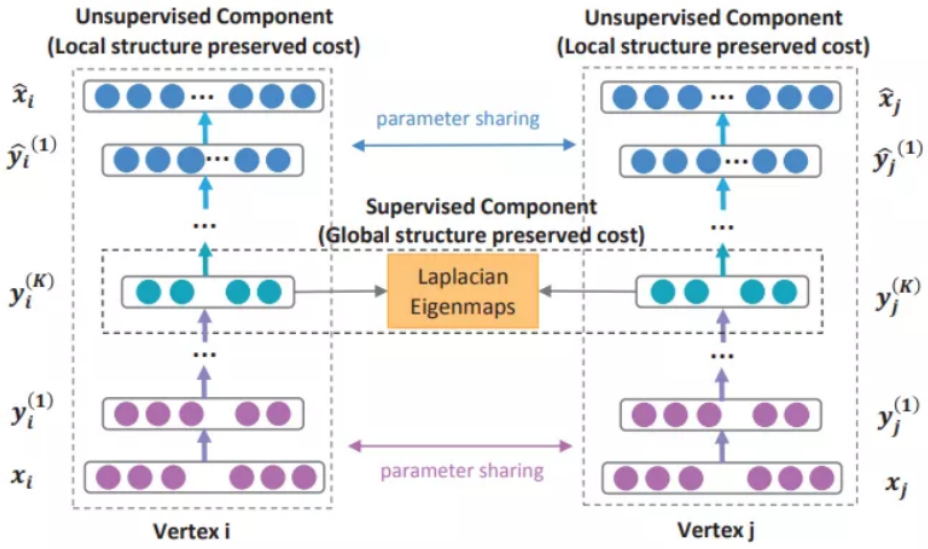
SDNE没有采用随机游走的方法而是使用自动编码器来同时优化一阶和二阶相似度，学习得到的向量表示保留局部和全局结构，并且对稀疏网络具有鲁棒性。

一阶相似度表征了边连接的成对节点之间的局部相似性。如果网络中的两个节点相连，则认为它们是相似的。

二阶相似度表示节点邻域结构的相似性，它能够了表征全局的网络结构。如果两个节点共享许多邻居，则它们趋于相似。

SDNE模型包括两部分：

1. 无监督的自动编码器，寻找可以重构其邻域节点的嵌入。
2. 有监督的拉普拉斯特征映射，当相似顶点在嵌入空间彼此映射很远时，该特征映射就会受到惩罚。




> 摘自 https://mp.weixin.qq.com/s?__biz=MzUyNzcyNzE0Mg==&mid=2247484484&idx=1&sn=5eec159195a8df164575d6498cb953d0&chksm=fa7a6b0dcd0de21b80bb4283f39c404d7b37069866a56048d3552ae702fe7140cc14a22e8d29&scene=21#wechat_redirect