

Struc2vec VS GraphWave

1、数据集

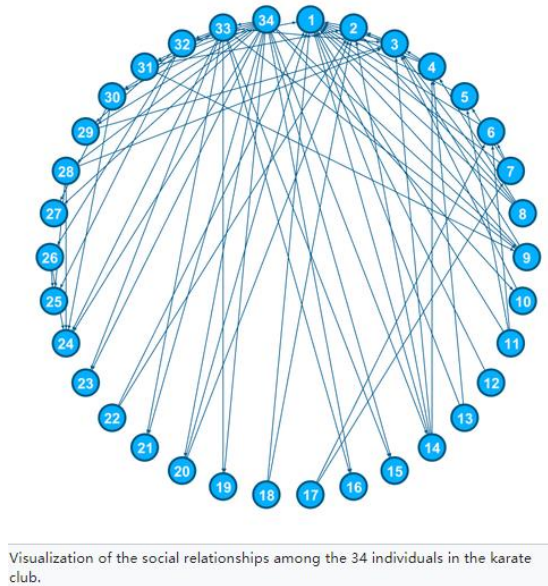


图 1 karate club 人与人连接关系图

上图 1 是 Karate club 中 34 个人相互之间的连接关系图，一共形成了 78 条边。论文 Struc2vec^[3]和 GraphWave^[4]都是为了解决长距离中 Structural roles 或者 Node roles 的问题。因为在论文 Node2vec^[2]中受限于 Random Walk 长度的限制，使得网络中或者图中距离很远的两个节点无法观测是否具有相似性，从而其角色难以判断。

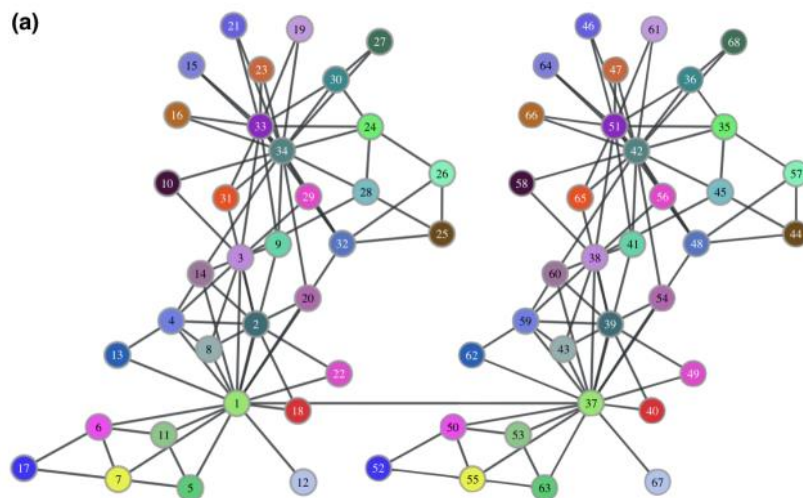


图 2 镜像的 karate club

上图 2 为 Struc2vec 论文中基于原始 Karate club 构造的镜像 karate club 图，图中

一共有 68 个节点，157 条边。

2、Struc2vec 实验结果复现

从上图 2 中可以看出其中 1-37、42-34、33-51 三对节点在图 2 中的扮演的角色是一致的。因此论文 Struc2vec 需要对由这 68 个节点组成的图或者网络进行 Embedding，使得在 Embedding Space 中 1-37、42-34、33-51 三对节点的距离应该很近。论文中先给出了 Deep Walk^[1]和 Node2vec 的实验结果如下图 3 所示。

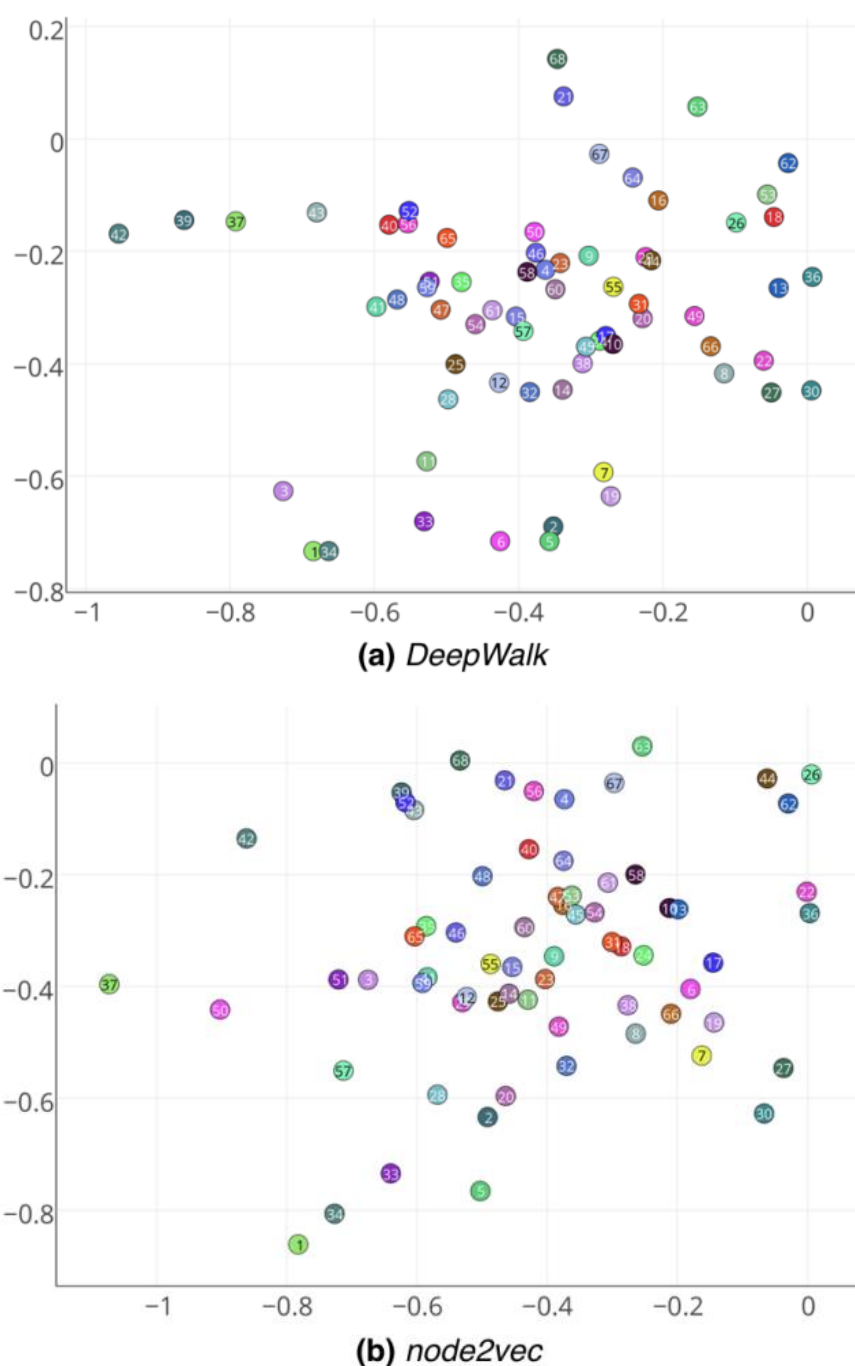


图 3 Deep Walk 和 Node2vec 的实验结果

从上图 3 中可以看出 Deep Walk 和 Node2vec 并没有很好的 Embedding，因为 1-37、42-34、33-51 这三对节点距离比较远。然后论文给出了 Struc2vec 方法的实验结果，如下图 4 所示。

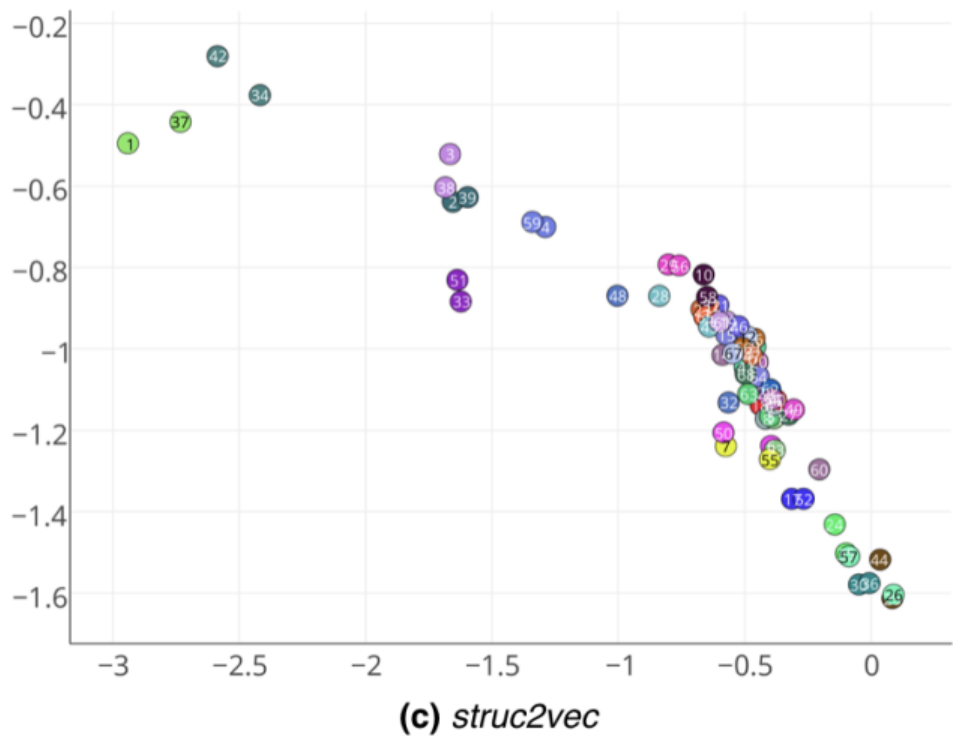


图 4 Struc2vec 的原始实验结果

从上图 4 中可以看出 Struc2vec 对网络中的 Nodes roles 做了更好的 Embedding。因为 1-37、42-34、33-51 这三对节点挨着很近。因此我对该实验结果进行了复现，如下图 5 所示。

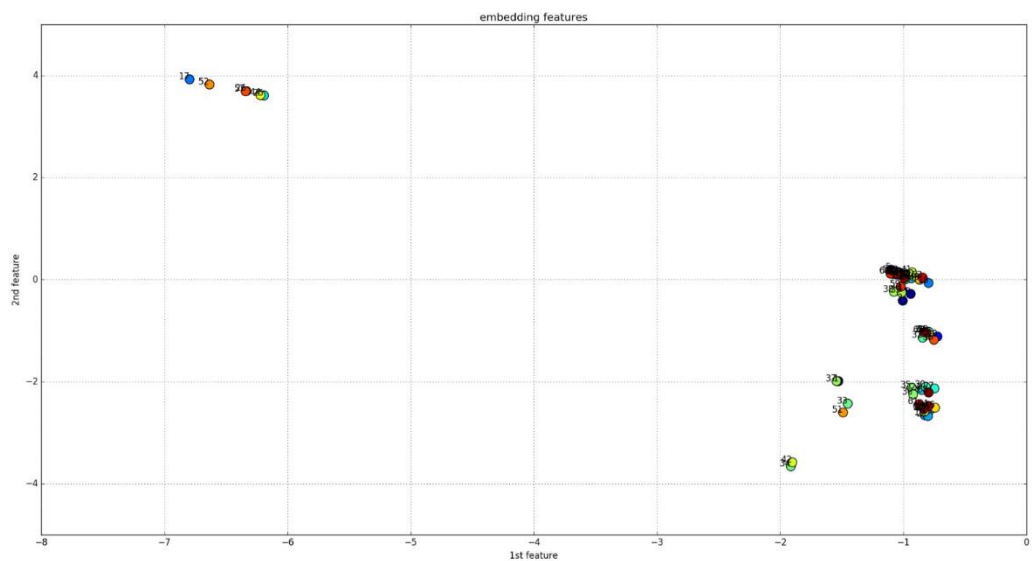


图 5 Struc2vec 的复现实验结果

将上图进行放大后，如下图 6 所示。发现 1-37，34-42，51-33 这三对节点都挨着很近，参照原始数据集图 2，发现确实 Struc2vec 是有效的。

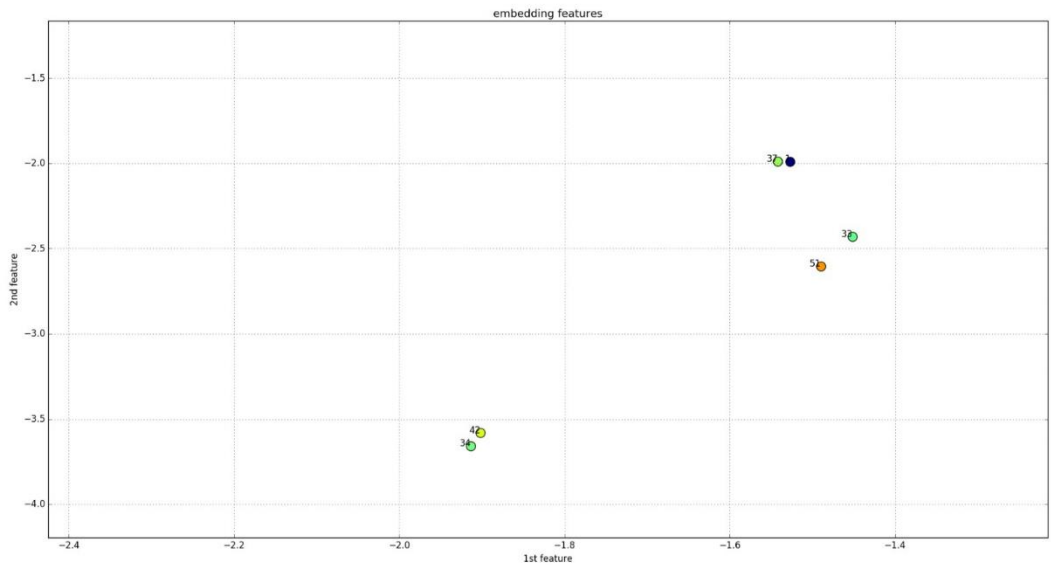


图 6 Struc2vec 的复现实验结果放大

因为原论文中在 Skip-gram 模型中采用的是 Hierachical-softmax，其计算速度比较快,但是牺牲了准确度。因此我把 Hierachical-softmax 改成了 Negative Sampling（取 K=15）以提升其准确度，其实验效果如下图 7 所示。发现确实有提升，因为在 Embedding Space 中 1-37，34-42，51-33 这三对节点距离更近了。

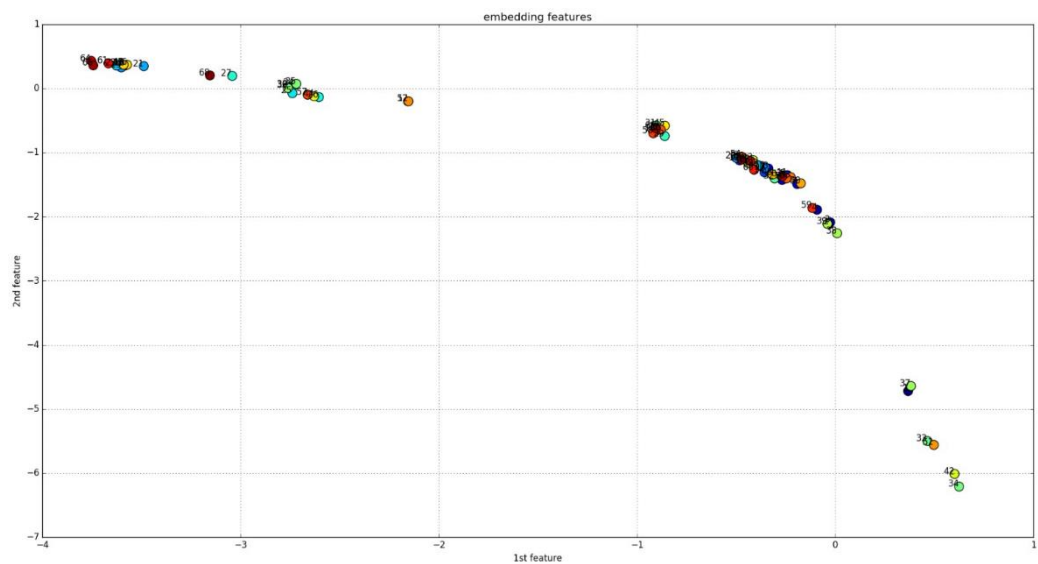


图 7 Negative Sampling 实验结果

将图 7 进行放大后，如下图 8 所示。发现 1-37，34-42，51-33 这三对节点呈现出很明显的间隔，表示 Negative Sampling 改善了一些精度，但是个人觉得好像也不是什么很大的改善。最好的结果应该是三对节点分别完全重合（见下文

GraphWave 方法)。

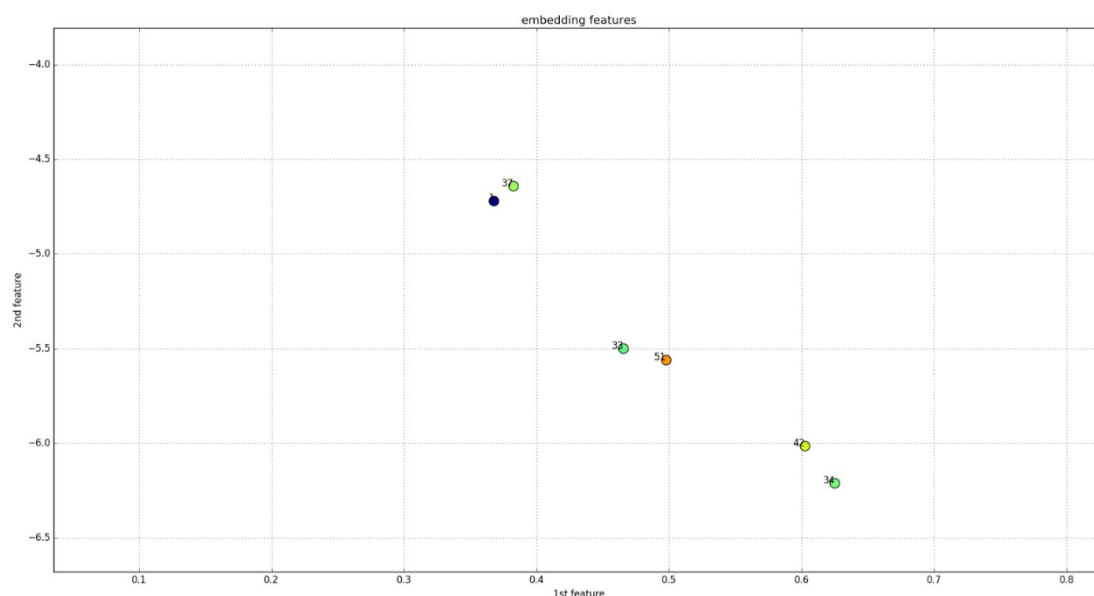


图 8 Negative Sampling 实验结果放大

经过对论文 Struc2vec 的研读和实验复现，自己在乐哥的指导下对其中的一些改进的地方进行了思考，主要有以下三点：

- (1) 利用图的基本属性来手工构造，增加多样性。
- (2) 在 Struc2vec 里头那个 fastDTW 的计算方式不知道有没有更好的。

(3) Skip-gram 里头原始论文用的 Hierarchical-softmax 替换为用 Negative sampling 效果更好（我这边做了实验，见上图 7 和 8）。

其中关于第一点，在和乐哥讨论后，收获蛮大，也发现了自己的差距——“读书不得间”，对细节和本质的地方把握不够。乐哥提出：Struc2vec 本质是基于节点来构造图，增加了多样性。我们是否可以用节点做为根节点，然后将图展开成树，从而判断树的相似性来判断节点的相似性？或者用节点做为根节点，然后将图展开成子图，从而转换为矩阵，利用矩阵分解等来判断矩阵的相似性从而判断节点相似性。

3、GraphWave 实验结果复现

在研读 Struc2vec 中，我发现了斯坦福大学被 KDD2018 收录的论文 GraphWave，该方法效果比 Struc2vec 更好，如论文给出的结果下图 9 所示。其主要思想是通过将小波处理为概率分布并使用经验特征函数表征分布，从而开拓了谱图小波的新用途——Nodes Embedding。在粗略阅读论文后发现其省略了需要手工设计特征增加多样性的过程，就是上述提出的第一点，目前还在深入研读论文中（该论文数学层面理解有点难）。

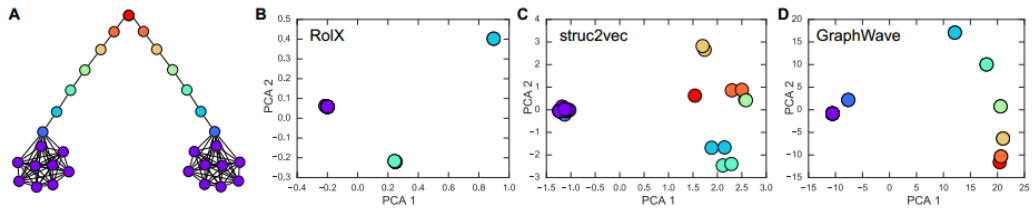


图 9 GraphWave 原始论文实验结果

因此，我就大胆的在镜像 karate club 数据集中用 GraphWave 方法进行 Embedding 结果也是让我惊讶了，确实是很好！如下图 10 所示，存在很多重合的节点。

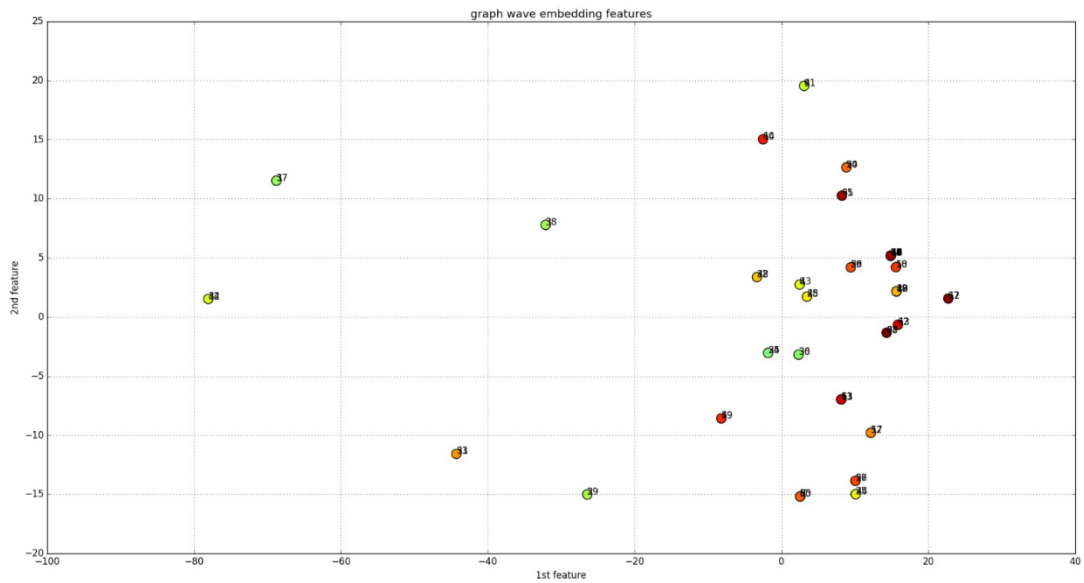


图 10 GraphWave 在镜像 karate club 数据集上的实验结果

提取 1-37, 34-42, 51-33 这三对节点的具体 Embedding 数值如下表 1 所示，简直完全重合，而且不只这三对节点。

表 1

x1	x2	id
-68.78969827	11.52408111	1
-68.78969827	11.52408111	37
-78.04661263	1.522077312	42
-78.04661263	1.522077312	34
-44.27573662	-11.58139654	51
-44.27573662	-11.58139654	33

4、结论

需要对 KDD2018 上的 GraphWave 进行深入研究，因此其效果目前在 Nodes roles 确认上最好。

5、参考文献

- [1] DeepWalk Online Learning of Social Representations(KDD2014)
- [2] Node2Vec Scalable Feature Learning for Networks(KDD2016)
- [3] Struc2vec Learning Node Representations from Structural Identity(KDD2017)
- [4] Learning Structural Node Embeddings via Diffusion Wavelets(KDD2018)