node2vec: Scalable Feature Learning for Networks - 2016/08 KDD

动机：现有采样方法例如 IsoMap 涉及矩阵操作，十分耗时、扩展性差，且其基于流形的假设较强（这点在 2.2 中做了说明）；另外随着深度学习发展，一些类似于 DeepWalk 等基于路径游走的建模方法被提出，但是这里采样策略多样化，不同采样对应不同的特征表达，都是刚性假设，无法随任务不同做出调整，不能适应灵活的目标（node2vec 在原文中提出，主要克服的是这个困难）。

基本思路：提出了一种基于二阶马尔科夫（指当前点下一步游走方向不仅仅依赖于当前点，还依赖于上一次经过的顶点）的有偏采样游走策略，来综合 BFS/DFS 探索 Graph 上的两种性质：同质性（homophily，比如在同一个社区中）、结构等同性（structural equivalence，不强调连通，注重顶点周围结构相同，比如不同社区的各自核心）。最后通过游走好的路径，借助自然语言处理中的 Skip-gram 模型来处理（假设：相似的词语出现在相似的邻居周围），从而获得每个顶点的 k 维连续特征。

Alias Method：通过 O(n)的预处理，使得之后可以对一个多项分布通过 O(1)的效率进行采样，参考：http://blog.csdn.net/mandycool/article/details/8182672

二阶马尔科夫性的说明：（当前点下一跳的概率分布，涉及当前点和上一跳顶点）

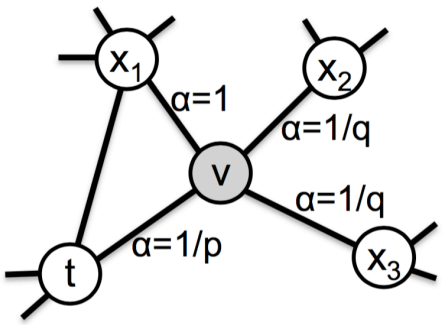

Figure 2: Illustration of the random walk procedure in *node2vec*. The walk just transitioned from t to v and is now evaluating its next step out of node v. Edge labels indicate search biases α.

其中对于一条路径 t->v，那么下一跳从 v 向其他节点的概率，依赖于当前顶点 v，以及顶点 t 到 v 的邻居的最短路径。在实现中，考虑使用了一个以 edge 为索引的数组，来保存各种情况下的下一跳的概率分布。整体空间复杂度相较于一阶马尔科夫需要多乘一个平均度数 a，整体空间复杂度为 a^2|V|。

算法流程：

Algorithm 1 The *node2vec* algorithm.

LearnFeatures (Graph $G = (V, E, W)$, Dimensions d, Walks per node r, Walk length l, Context size k, Return p, In-out q)
  π = PreprocessModifiedWeights(G, p, q)
  $G' = (V, E, \pi)$
  Initialize $walks$ to Empty
  for $iter = 1$ **to** r **do**
    for all nodes $u \in V$ **do**
      $walk$ = node2vecWalk(G', u, l)
      Append $walk$ to $walks$
  f = StochasticGradientDescent(k, d, $walks$)
  return f

node2vecWalk (Graph $G' = (V, E, \pi)$, Start node u, Length l)
  Inititalize $walk$ to $[u]$
  for $walk\_iter = 1$ **to** l **do**
    $curr = walk[-1]$
    V_{curr} = GetNeighbors($curr$, G')
    s = AliasSample(V_{curr}, π)
    Append s to $walk$
  return $walk$

参考：http://snap.stanford.edu/node2vec/#code

实验效果：

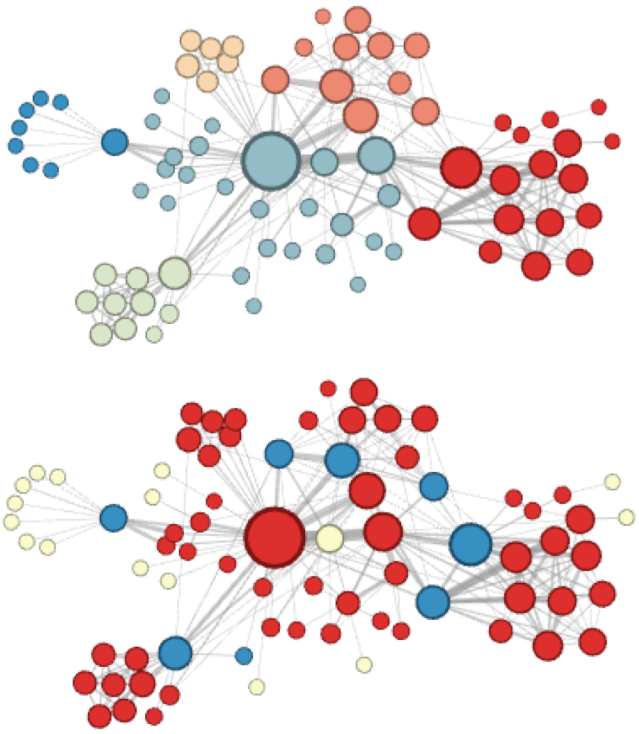

Figure 3: Complementary visualizations of Les Misérables coappearance network generated by *node2vec* with label colors reflecting homophily (top) and structural equivalence (bottom).

上方，同质性（社区）的展示，p=1 & q=0.5，同色点在社会中基本处于同一个社区范围；

下方，结构等同性（社会角色），p=1 & q = 2.0，例如蓝色点，基本都是不同社区之间的桥梁、黄色点，大部分都是交流比较少的顶点（但也存在少量错误）。

注：文中实验表明参数敏感性还是比较高的，所以对于不同任务也会需要调参.

宋军帅

2017/03/13