Network Representation Learning: A Survey

网络表示学习是近年来提出的一种新的网络分析的学习方法，通过保留网络拓扑结构、顶点内容和其他信息，在新的向量空间里方便地处理原始网络，以便进一步的分析。

论文回顾了当前在数据挖掘和机器学习领域中网络表示学习的文献，根据底层的学习机制、打算保存的网络信息和算法设计和方法，提出了新的分类方法来分类和总结目前最先进的网络表示学习技术；总结了用于网络表示学习验证的评估协议，包括已发布的基准数据集（benchmark datasets），评估方法和开源算法。

1. 通常用离散邻接矩阵来表示一个网络，它只捕获临近的节点之间的关系。这种方式不能体现更复杂、更高阶的结构关系。
2. 网络表示学习面临的挑战：**结构保持Structure-preserving**（同时保存局部和全局的结构）、**内容保持Content-preserving**（如何利用好vertex attribute）、**数据稀疏性Data sparsity**（structrure-level relatedness（不是相邻节点之间的联系难以发现）， content-level vertex similarity(很多vetex attribute通常missing)）、**可扩展性Scalability**（设计有效学习节点表示NRL算法的必要性）
3. 信息网络定义：$G = (V, E, X, Y)$

TABLE 1
A summary of common notations

G	The given information network
V	Set of vertices in the given information network
E	Set of edges in the given information network
$\|V\|$	Number of vertices
$\|E\|$	Number of edges
m	Number of vertex attributes
d	Dimension of learned vertex representations
$X \in \mathbb{R}^{\|V\| \times m}$	The vertex attribute matrix
$\mathcal{Y}$	Set of vertex labels
$\|\mathcal{Y}\|$	Number of vertex labels
$Y \in \mathbb{R}^{\|V\| \times \|\mathcal{Y}\|}$	The vertex label matrix

一阶临近、二阶临近、高阶临近、结构临近（structural role proximity）、社区内临近（intra-community proximity）（社区内的节点-节点连接紧密，社区外稀疏）

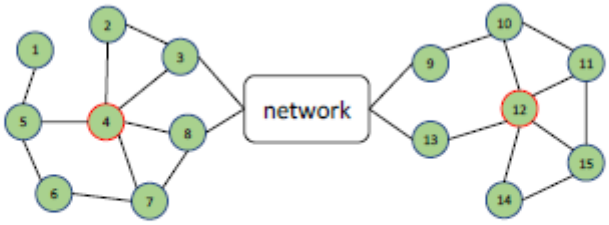
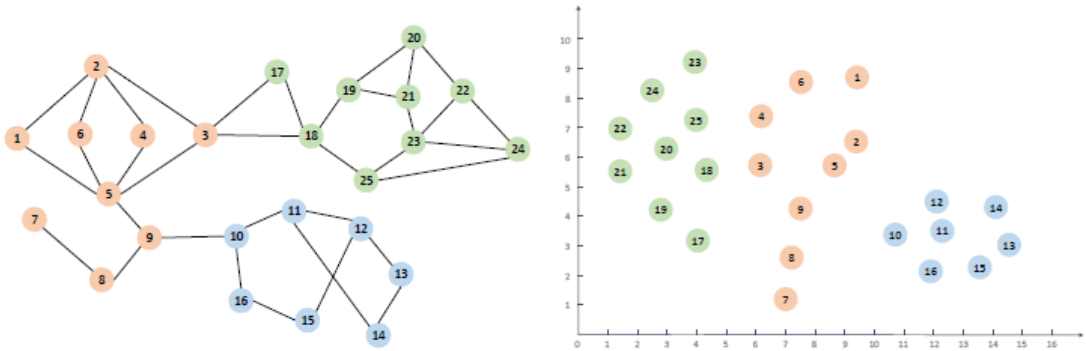


Fig. 1. An illustrative example of structural role proximity. Vertex 4 and vertex 12 have similar structural roles, but are located far away from each other.

网络表示学习的任务就是学习函数$f: v->r_v \in R^d$, r_v是节点v的学习向量表示，d是学习表示的维度。需要满足：（1）低维：$d << |V|$（2）informative：节点表示要保存节点临近、节点属性、节点label（3）连续性。

概念图：



(a) Input: Information Network (b) Output: Vertex Representations

具有临近的节点在节点表示图中距离近。

4. NPL方法的分类：（三种主要的信息源：网络结构，节点属性，节点标签）

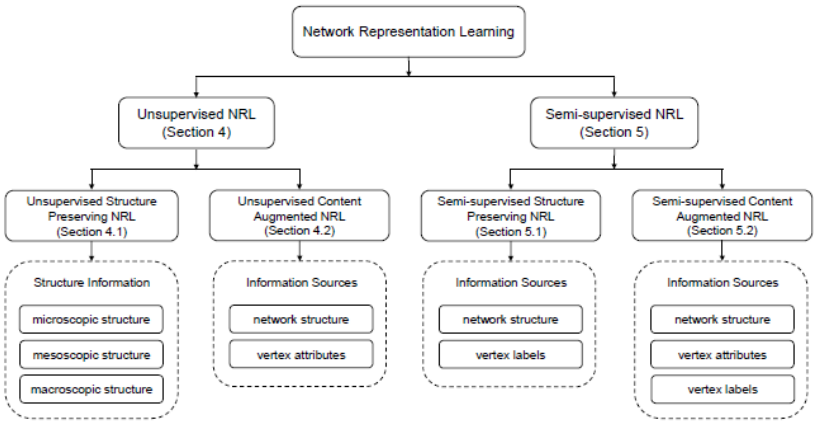


Fig. 3. The proposed taxonomy to summarize network representation learning techniques. We categorize network representation learning into two groups, *unsupervised network representation learning* and *semi-supervised network representation learning*, depending on whether vertex labels are available for learning. For each group, we further categorize methods into two subgroups, depending on whether the representation learning is based on network topology structure only, or augmented with information from node content.

从方法论角度分类：

TABLE 3
A categorization of NRL algorithms from methodology perspectives

Methodology	Algorithms	Advantage	Disadvantage
Matrix Factorization	Social Dim. [31], [32], GraRep [26], HOPE [35], GraphWave [39], M-NMF [28], TADW [7], HSCA [20], MMDW [46], DMF [8], LANE [30]	capture global structure	high time and memory cost
Random Walk	DeepWalk [6], node2vec [34], APP [36], DDRW [45], GENE [48], TriDNR [50], UPP-SNE [43], struct2vec [38], SNS [40], PPNE [44], SemiNE [49]	relatively efficient	only capture local structure
Edge Modeling	LINE [1], TLINE [47], LDE [51], pRBM [29], GraphGAN [37]	efficient	only capture local structure
Deep Learning	DNGR [9], SDNE [19]	capture non-linearity	high time cost
Hybrid	DP [41], HARP [42], Planetoid [52]	capture global structure	

5. 无监督网络表示学习：

    1. 无监督结构保存网络表示学习

    微观结构保存NRL：**DeepWalk**（泛化skip-gram，通过所给的randomwalk序列预测上下文节点，学习节点v_i的表示，randomwalk序列描述邻居结构，deepwalk表示了在embedding空间中有着类似邻居的节点，所以保留了二阶和高阶的临近）；**LINE**（large-scale information network embedding）通过对一阶、二阶临近建模学习节点表示；**GraRep**（和deepwalk的idea一样，扩展skip-gram模型来捕获高阶的临近，特别的，对每个节点，grarep定义其k阶邻居作为上下文节点，来学习k步的节点表示）；**DNGR**（Deep Neural Networks for Graph Representations）；**SDNE**（Structural Deep Network Embedding）：利用深度自编码模型来捕获网络结构的非线性；**node2vec**：BFS、DFS；**HOPE**（High-order Proximity Preserved Embedding）；**APP**（Asymmetric Proximity Preserving graph embedding）：捕获非对称临近，蒙特卡洛；**GraphGAN**：通过对对抗学习框架的连接行为建模来学习节点表示。

TABLE 4
A summary of microscopic structure preserving NRL algorithms

Algorithms	First-order Proximity	Second-order Proximity	High-order Proximity
DeepWalk [6]		✓	✓
LINE [1]	✓	✓	
GraRep [26]		✓	✓
DNGR [9]		✓	✓
SDNE [19]	✓	✓	
node2vec [34]		✓	✓
HOPE [35]		✓	✓
APP [36]		✓	✓
GraphGAN [37]	✓		

Structural Role Proximity Preserving NRL（结构临近）：**struct2vec**：将顶点的结构角色相似度编码成一个多层图，在多层图执行deep-walk学习顶点表示。**GraphWave**：使用spectral graph wavelet diffusion patterns，将节点邻居结构嵌入一个低维空间并保存structural role proximity。 **Structural and Neighborhood Similarity preserving network embedding (SNS)**：用structural role proximity提高了基于random walk的方法。SNS把每个节点表示为graphlet degree vector。

Intra-community Proximity Preserving NRL：**学习潜在社会维度**：Community Detection--旨在发现一组组内联系比组间联系更紧密的community。clustering techniques：modularity maximization，spectral clustering，edge clustering。**M-NMF**（Modularized Nonnegative Matrix Factorization）：增加了拥有更宽社区结构的二阶、高阶邻近来学习更多信息节点嵌入（区别于学习潜在社会维度仅考虑社区结构）。总结：这两种方法都依赖矩阵分解，难以扩展到大规模。

Macroscopic Structure Preserving NRL：旨在保存全局网络特性。**DP**（degree penalty principle）：无标度特性（scale free）：大多数节点连接稀疏，少数连接紧密。学习无标度特性保存节点表示：spectral embedding、deepwalk。**HARP**（hierarchical represenation learning for networks）：捕获全局pattern。deepwalk、line学习节点表示。

2. 无监督内容增强的网络表示学习：顶点属性为测量顶点之间的内容相似性提供了直接证据，通过合并网络结构和顶点属性可以加强网络表示学习。

   TADW（Text-Associated DeepWalk）、HSCA（Homophily, Structure, and Content Augmented Net-work Representation Learning）、pRBM（Paired Restricted Boltzmann Machine）、UPP-SNE(User Profile Preserving Social Network Embedding)、PPNE（Property Preserving Network Embedding）：优化目标：结构驱动目标、属性驱动目标。

6. 半监督网络表示学习

   1. 半监督网络表示学习（利用顶点标签）

      DDRW（Discriminative Deep Random Walk）、 Max-Margin DeepWalk (MMDW)、Transductive LINE (TLINE)、 Group Enhanced Network Embedding (GENE)、 Semi-supervised Network Embedding (SemiNE)

   2. 半监督内容增强NRL：

      Tri-Party Deep Network Representation (TriDNR)：从网络结构、顶点内容、顶点标签三个信息源中学习顶点表示。 Linked Document Embedding (LDE)、 Discriminative Matrix Factorization (DMF)、 Predictive Labels And Neighbors with EmbeddingsTransductively Or Inductively from Data (Planetoid)、 Label informed Attribute Network Embedding (LANE)

TABLE 5
A summary of semi-supervised NRL algorithms

Discriminative Learning Strategy	Algorithm	Loss function	Advantage	Disadvantage
fitting a classifier	DDRW [45]	hinge loss	a) directly optimize classification loss; b) perform better in sparsely labeled scenarios	prone to overfitting
	MMDW [46]	hinge loss		
	TLINE [47]	hinge loss		
	DMF [8]	square loss		
	SemiNE [49]	logistic loss		
modeling vertex label relation	GENE [48]	likelihood loss	a) better capture intra-class proximity; b) generalization to other tasks	require more labeled data
	TriDNR [50]	likelihood loss		
	LDE [51]	likelihood loss		
	Planetoid [52]	likelihood loss		
joint vertex label embedding	LANE [30]	correlation loss		

7. 应用

   1. 顶点分类：大部分网络顶点没有标签因为标签成本比较大，顶点分类需要充分利用顶点之间的关系。网络表示学习遵循着基于网络结构自动学习顶点特征的原则。更好的顶点分类可以提供更好的分类准确度。
   2. 链路预测：基于已观察连接和特性，推断出实体对之间的新的关系或者正在发生的相互作用。（推荐朋友、预测蛋白之间的相互作用）
   3. 聚合：网络聚类指网络顶点划分为一组群集，在同一个cluster里的节点紧密连接。最大化cluster内的连接，最小化cluster间的连接。
   4. 可视化：传统的可视化方法在大规模的网络中面临挑战。可视化的第一步：降低网络规模，找到一个低维的网络表示，保持网络原本的结构，相似的节点在低维度空间相邻。
   5. 推荐：社交网络：结构、内容、顶点标签信息、地理和时空信息、POI(对于兴趣的观点)
   6. 知识图：数据库系统中的一种新型数据结构，对数以亿计的实体的结构信息和他们丰富的关系进行编码。

8. 评估协议

   1. 基准数据集：社会网络（顶点标签由用户兴趣组定义，但用户属性不可用）、语言网络（单词共现网络，通过词的类比和文档的分类来评价从网络中学习到的词的嵌入）、引用网络

（由作者-作者引用关系或论文-论文引用关系构成的定向信息网络）、协作网络、网页网络、生物网络、通信网络、交通网络

2. 评估方法：网络重建、顶点分类、顶点聚合、链路预测、可视化
3. 实验结果：文献中常用的两种评估方法：顶点分类、顶点聚合（Accuracy、Normaliszed Mutual Infomation）

9. 未来的研究方向：

理论：缺乏对算法特性的理论分析

动态：现有的NRL的研究主要考虑静态网络，但是现实场景中网络不总是静态的。动态网络的一些让静态网络的嵌入无法工作的特性：（1）顶点内容特征随着时间漂移；（2）增加新顶点和新边要求学习和更新顶点表示更加有效率；（3）网络的规模不是固定的。

可扩展性：主要挑战，采用随机梯度下降优化的基于randomwalk和边缘建模的方法比通过特征分解分解和迭代优化解决的基于矩阵分解的方法效率更高。基于矩阵分解的方法在合并顶点属性和发现社区结构方面表现出了巨大潜力，但需要解决其在大规模网络中应用的可扩展性问题。