飞桨

图神经网络七日打卡营

小斯妹　百度PGL团队成员

2020.11.23

课程大纲

第一课：图学习初印象

- 图学习概述、入门路线
- 实践：环境搭建

第二课：图游走类算法

- DeepWalk, node2vec, metapath2vec
- 实践：DeepWalk, node2vec

第三课：图神经网络算法(一)

- GCN, GAT、消息传递
- 实践：GCN, GAT

第四课：图神经网络算法(二)

- 图采样、邻居聚合
- 实践：GraphSage

第五课：GNN 进阶

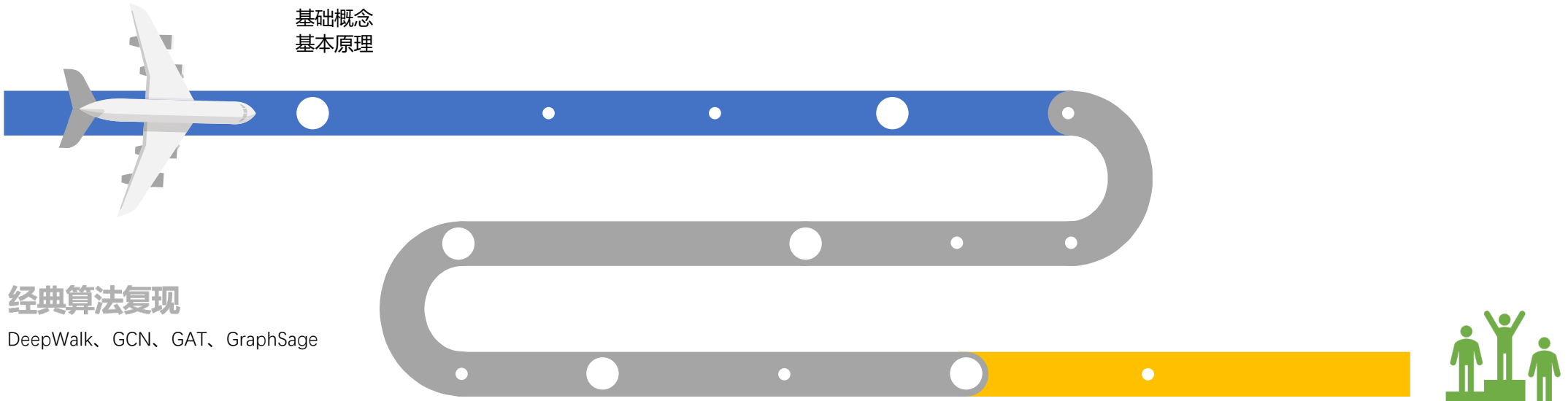
- ERNIESage, UniMP
- 实践：ERNIESage 代码讲解

后续：新冠项目实战，带你助力疫情防控

参考材料：

- 斯坦福CS224W课程： http://cs224w.stanford.edu
- 图学习库 PGL：https://github.com/PaddlePaddle/PGL

图学习入门路线


基础入门
基础概念
基本原理
经典算法复现
DeepWalk、GCN、GAT、GraphSage
前沿论文阅读
顶会论文

课程基础

- 数学基础：
  - 高等数学
  - 线性代数
  - 概率与数理统计
- 编程基础：
  - Python：numpy
  - Paddle：
    https://www.paddlepaddle.org.cn/documentation/docs/zh/beginners_guide/index_cn.html
- 机器学习基础：
  - 神经网络（FC、BPNN）

资料推荐

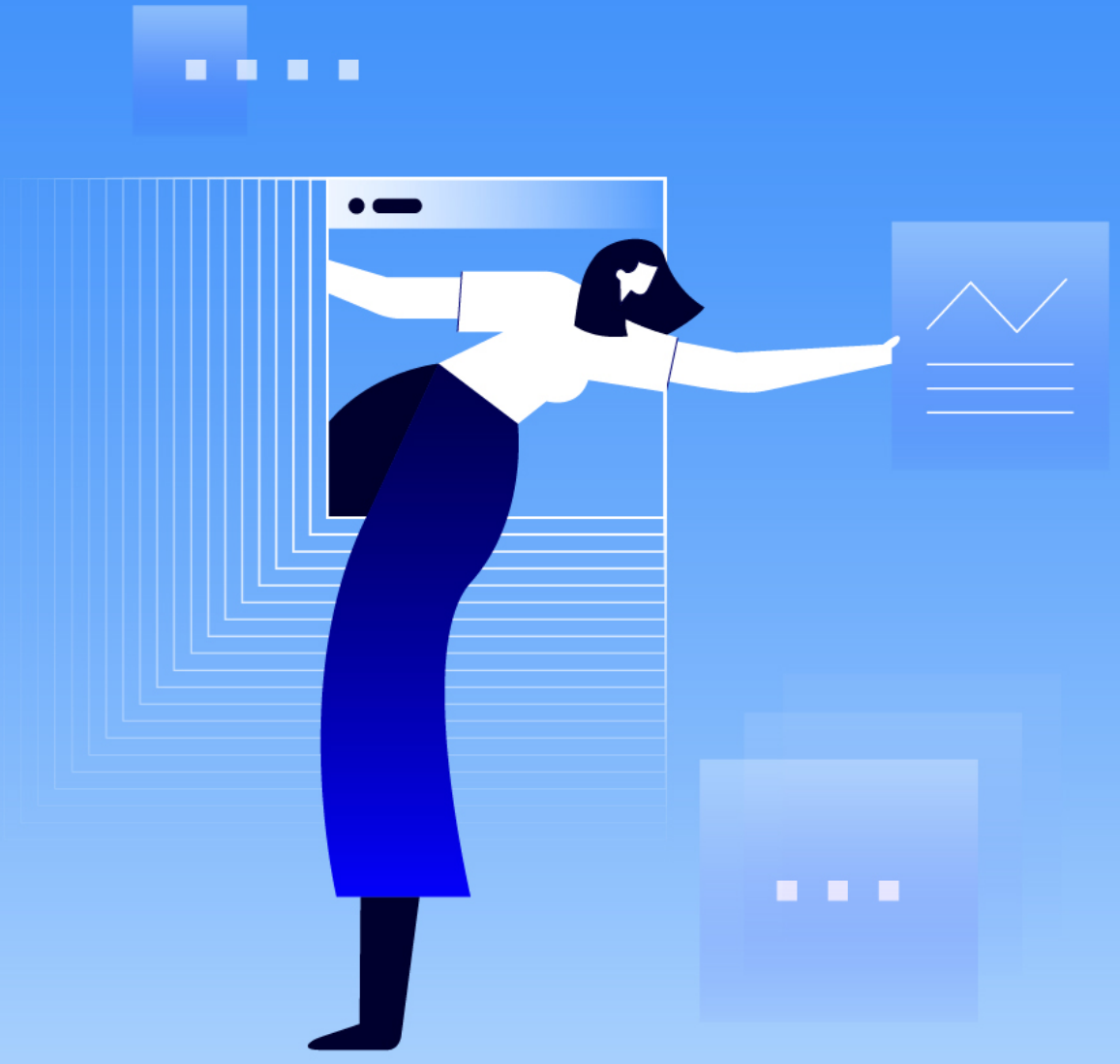
- 理论：
  - 综述
    Graph Neural Networks: A Review of Methods and Applications https://arxiv.org/pdf/1812.08434.pdf
    A Comprehensive Survey on Graph Neural Networks https://arxiv.org/abs/1901.00596
  - 视频
    - 理论课： 斯坦福 CS224W

- 动手实践：DeepWalk, node2vec, GCN, GAT, GraphSAGE（本次公开课实践内容）

- 进阶：经典论文
  - DeepWalk. " DeepWalk: Online Learning of Social Representations " https://arxiv.org/pdf/1403.6652.pdf
  - GCN. “ Semi-supervised Classification with Graph Convolutional Networks" https://arxiv.org/pdf/1609.02907.pdf
  - GAT. "Graph Attention Networks" https://arxiv.org/pdf/1710.10903.pdf
  - GraphSAGE. " Inductive Representation Learning on Large Graphs" https://arxiv.org/pdf/1706.02216.pdf
  - https://github.com/thunlp/GNNPapers

- 前沿研究方向：
  - 针对 GNN 而言：图采样技术、邻居聚合、深度图卷积、图预训练

飞桨



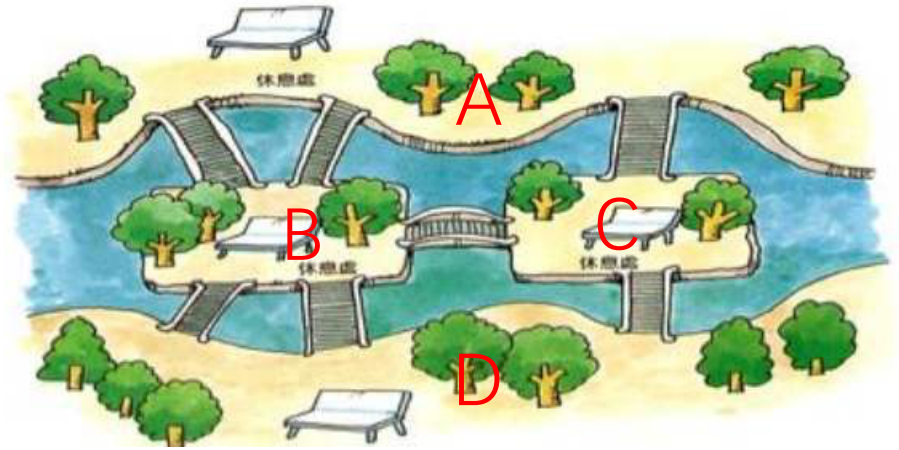
第一课 图学习初印象

小斯妹 百度PGL团队成员

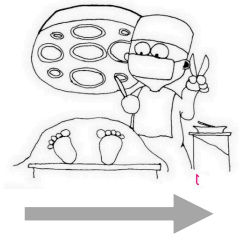
2020.11.23

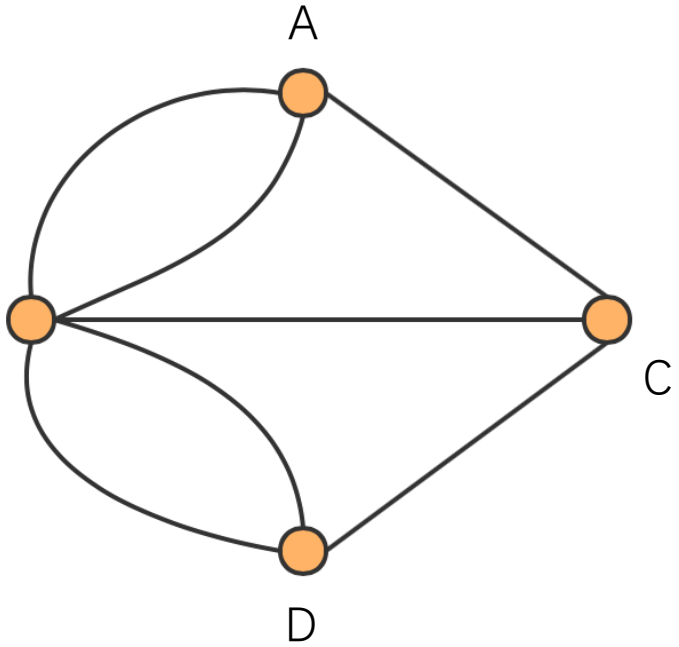
什么是图？

图论的开始



七桥问题





图(Graph)就是节点(Vertices/Nodes)以及边(Edge)

$$G = (V, E)$$

注：一般情况下，我们认为图(graph)和网络(network)两种术语可交替使用。通常我们称为图。

实际的例子

图是一种统一描述复杂事物的语言。

社交网络

节点：人
边：人与人之间的各种联系，如父母关系、朋友关系、同事关系等等。

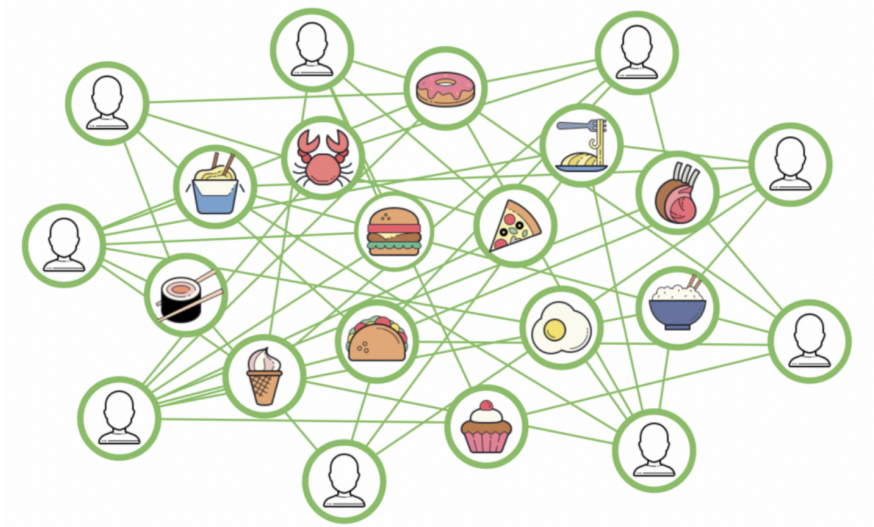
互联网

节点：网页
边：　网页与网页之间的超链接关系

实际的例子

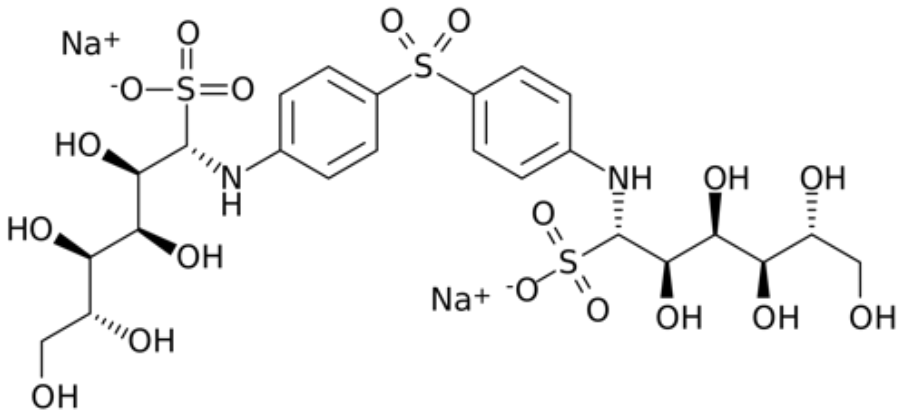
图是一种统一描述复杂事物的语言。

还有什么事物是可以用图来表示的呢？



推荐系统

节点：用户和商品
边：　用户、商品之间的购买、点击等关系



化学分子

节点：原子
边：　原子之间的相互作用力，也称为化学键

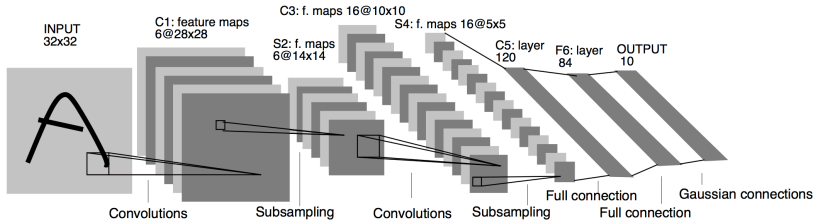
何为图学习？

数据对象：图

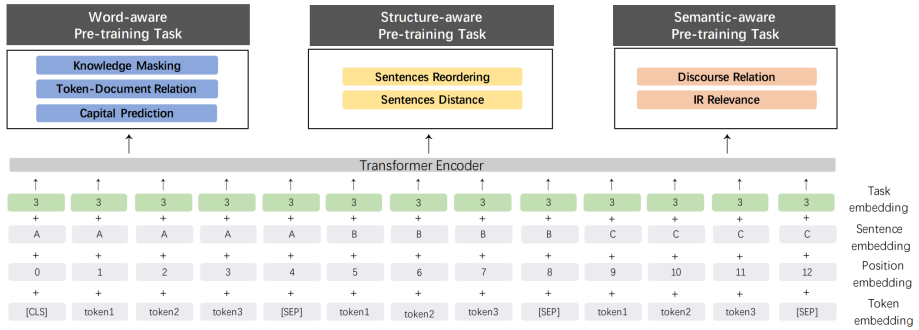
- 语音、图像、文本具有整齐**规则**的数据结构

语音

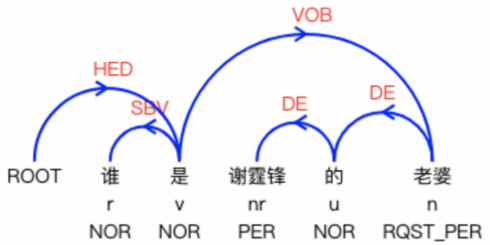
图像



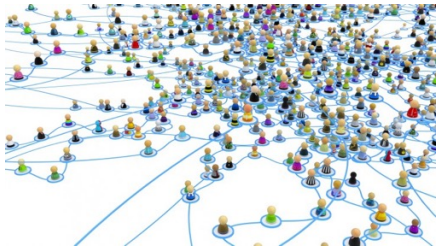
文本



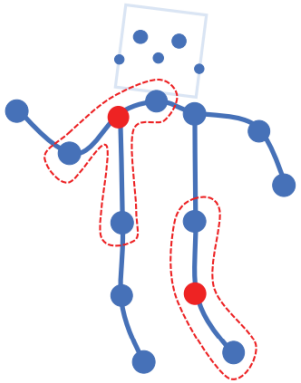
- 现实中的图是**不规则的**，难以直接建模



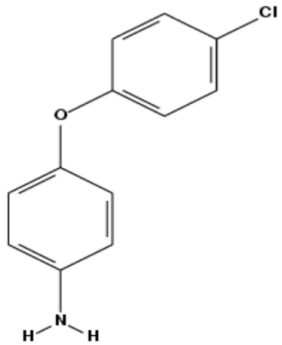
语法树



社交网络



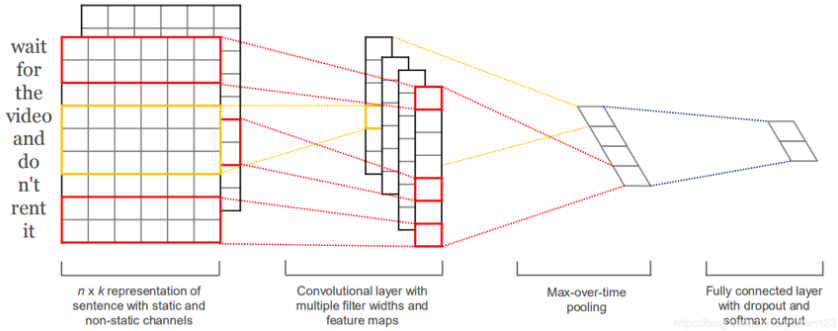
人体骨骼



化学分子

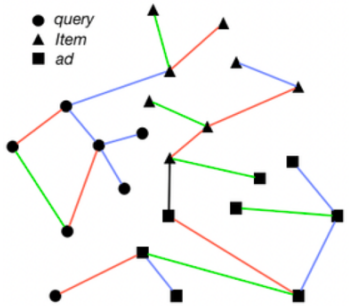
图学习的优势

一般深度学习：难以处理不规则数据。



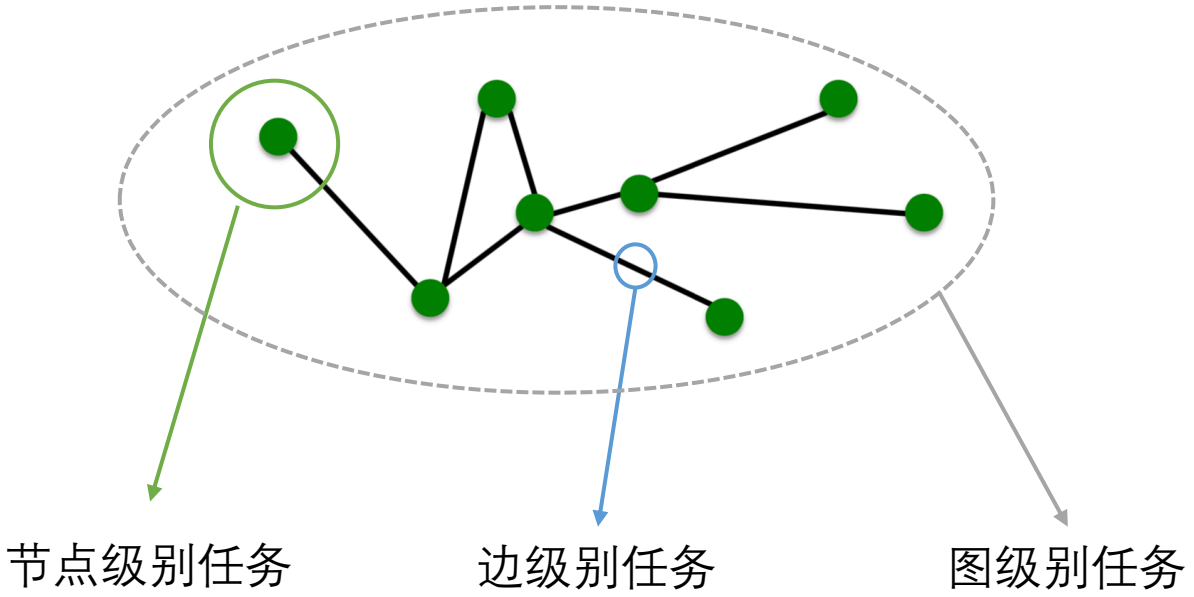
图学习：可以方便地处理不规则数据(图)，充分利用图结构信息。

地图

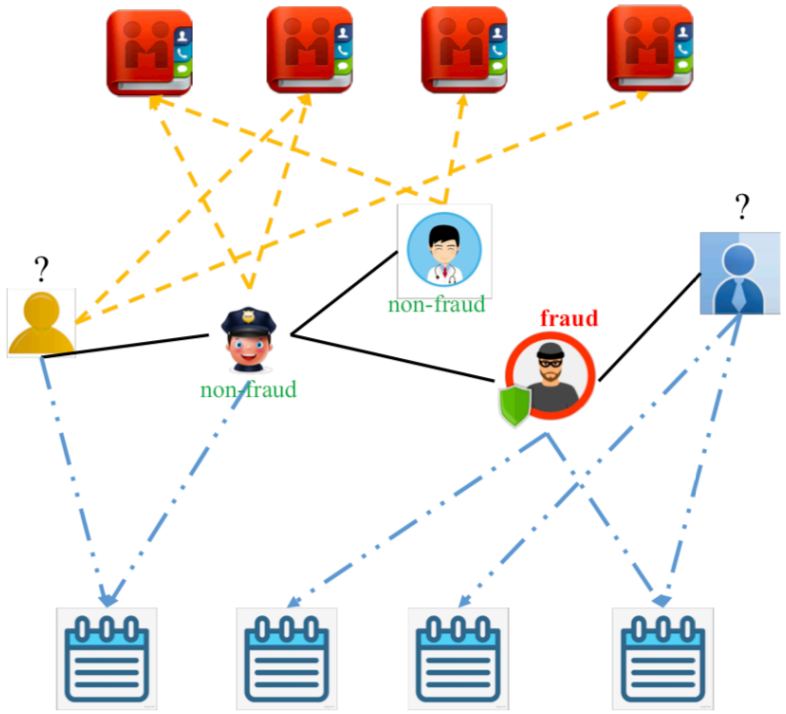


商品广告

图能做什么？——图学习的应用



节点级别任务——金融诈骗检测



节点分类任务

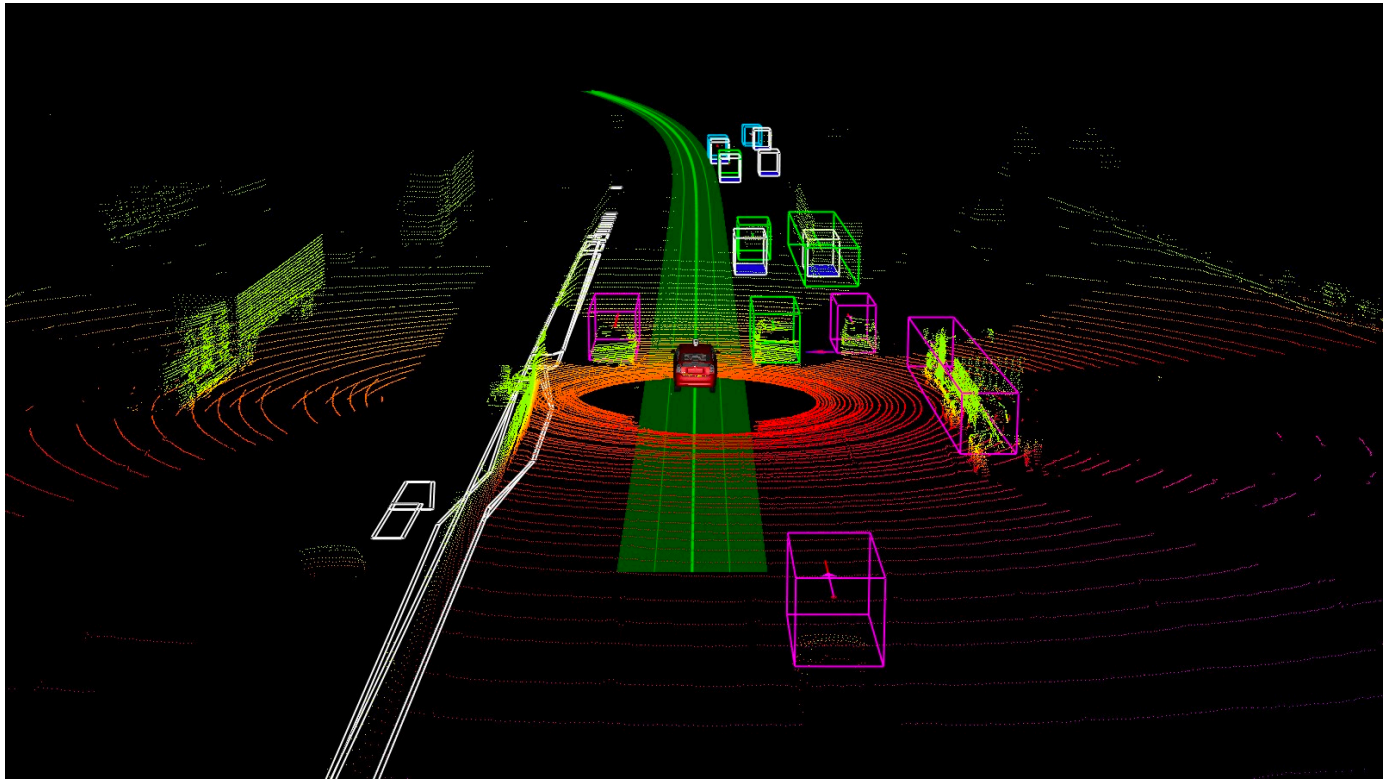
图模型

预测潜在的金融诈骗分子

A Semi-supervised Graph Attentive Network for Financial Fraud Detection

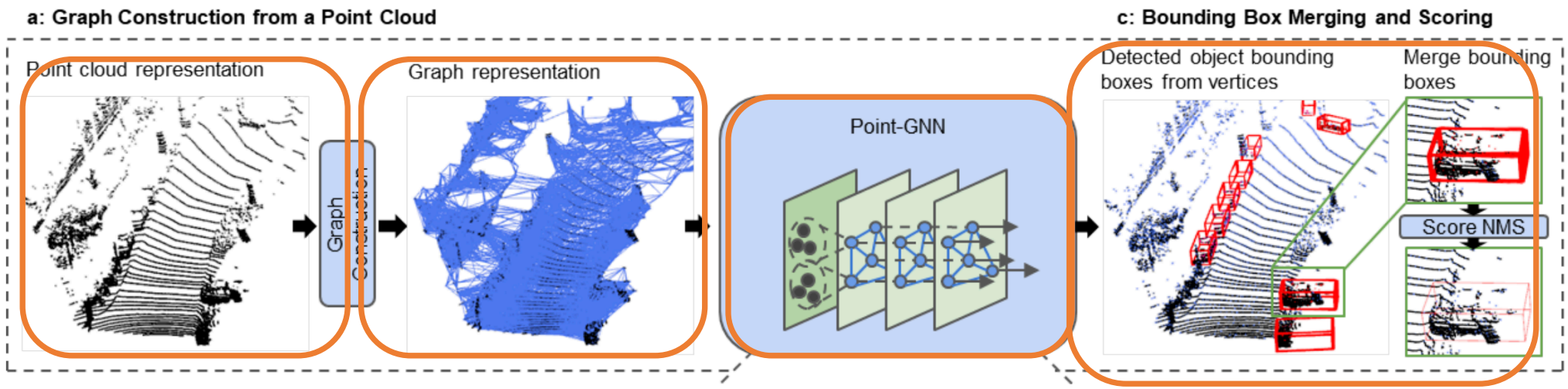
节点级别任务——目标检测

目标检测是自动驾驶中非常重要的一门技术。



节点级别任务——目标检测

从3D点云中利用图模型进行目标检测



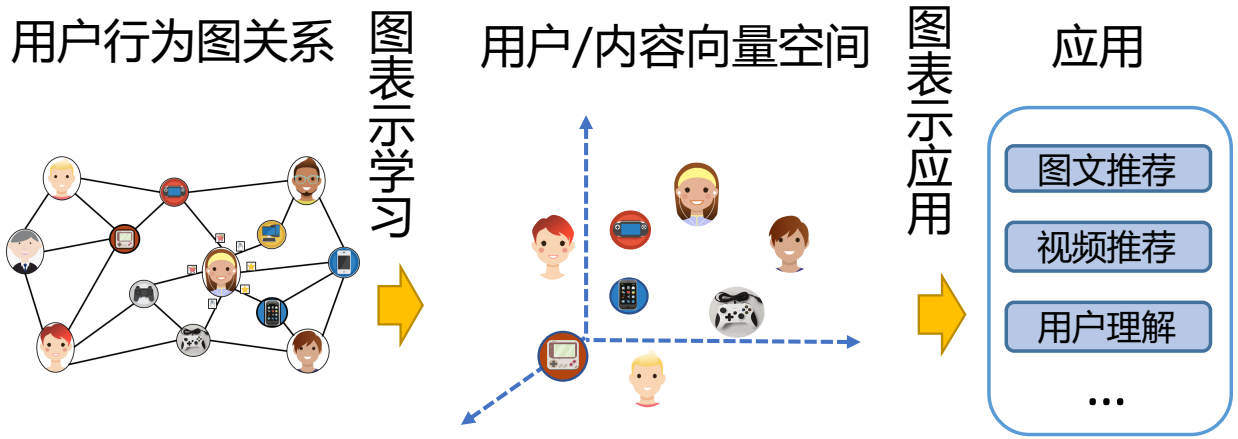
点云数据　　建图　　图模型　　目标检测

Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud

边级别任务——推荐系统

新闻推荐

- **推荐系统是经典的Link Prediction任务**



图级别任务——气味识别

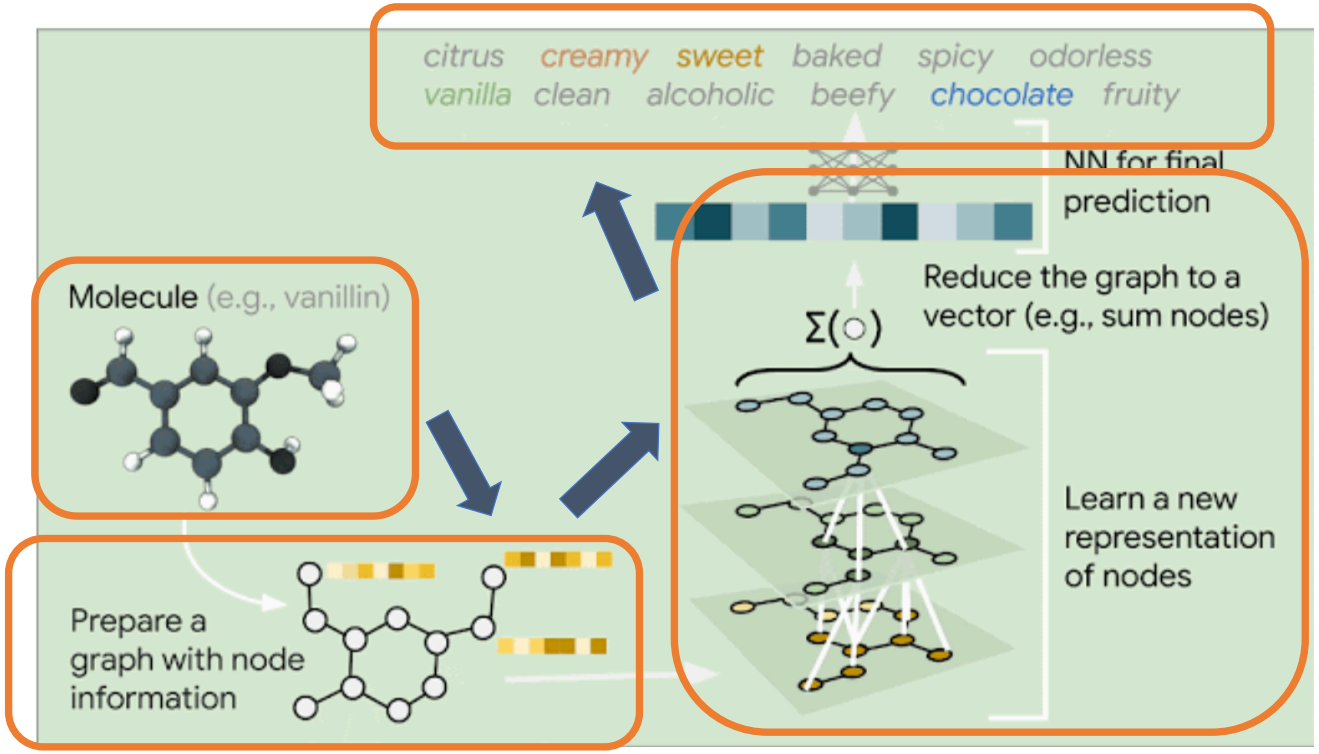
月季

玫瑰

图级别任务——气味识别

分子结构

建图

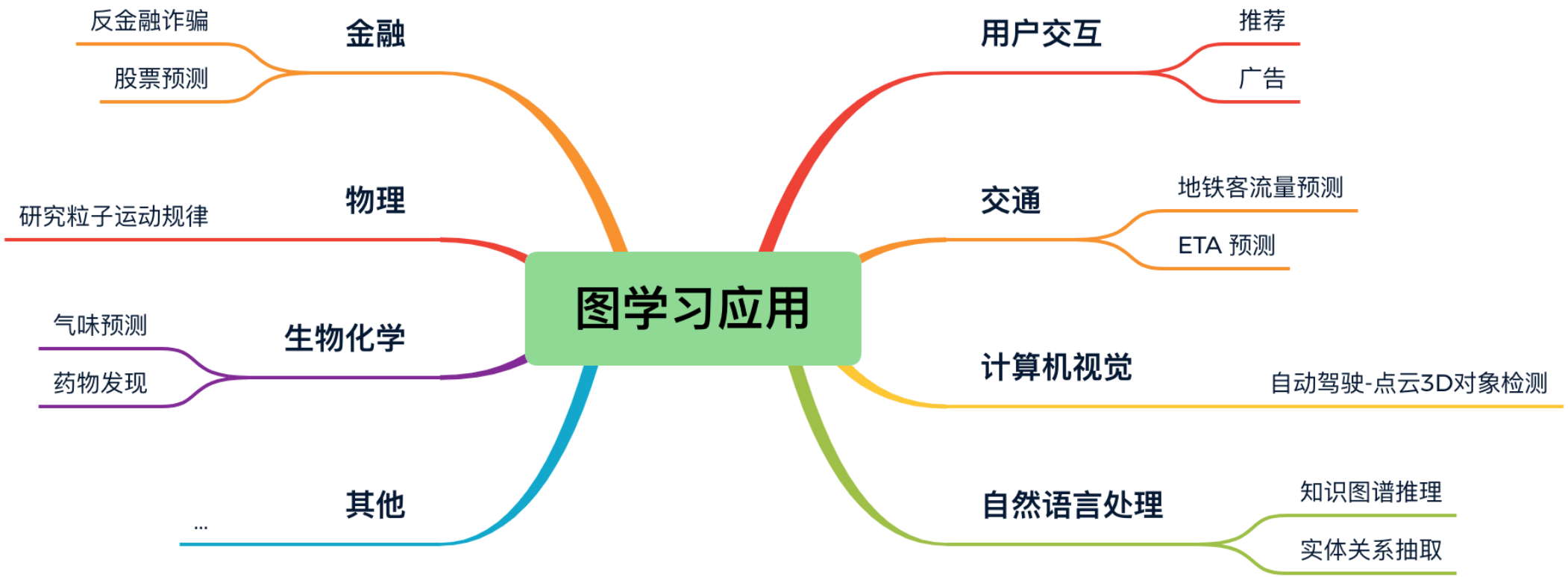


气味分类

图模型

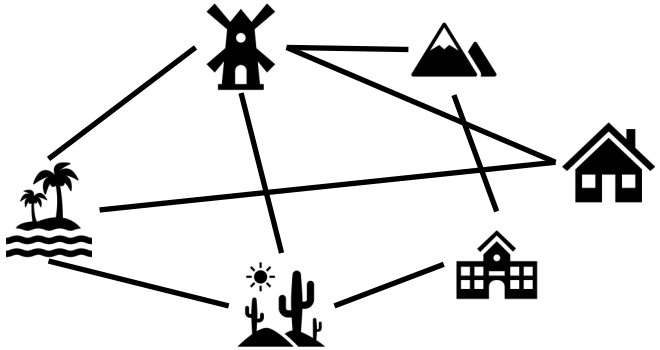
https://ai.googleblog.com/2019/10/learning-to-smell-using-deep-learning.html

图学习应用

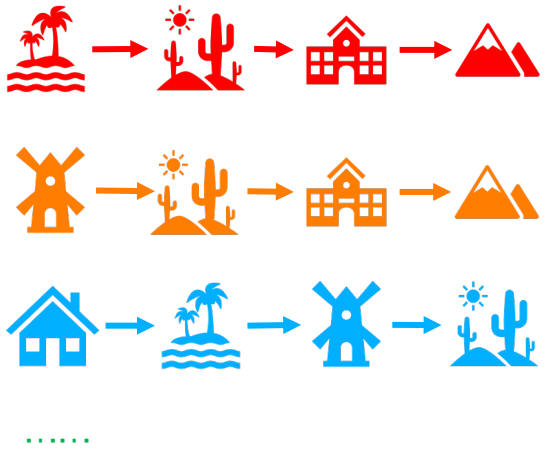

图学习应用
金融
反金融诈骗
股票预测
物理
研究粒子运动规律
生物化学
气味预测
药物发现
其他
...
用户交互
推荐
广告
交通
地铁客流量预测
ETA 预测
计算机视觉
自动驾驶-点云3D对象检测
自然语言处理
知识图谱推理
实体关系抽取

图学习是如何做的？

图游走类算法



假设序列最大长度为4.



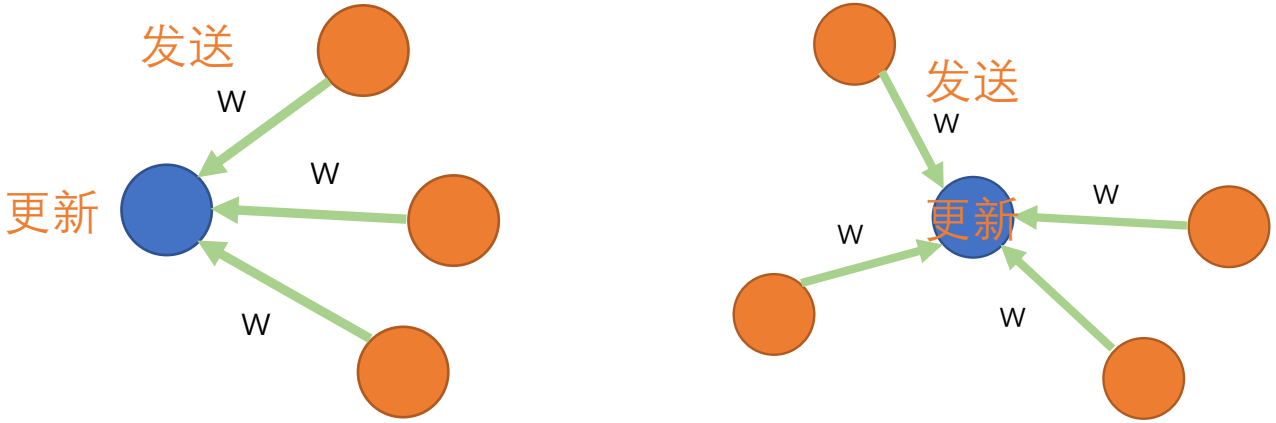
图表示学习

下游任务：
节点分类等

下节课内容

图学习是如何做的？

图神经网络算法——消息传递



图学习算法分类

知识图谱也是典型的图。

图学习框架——PGL

Github: https://github.com/PaddlePaddle/PGL

API文档： https://pgl.readthedocs.io/en/latest/

Paddle Graph Learning

易用

支持异构图的Meta Path和Message Passing双机制

高效

基于LoD Tensor的并行消息聚合算法

规模

内置分布式图引擎，支持十亿节点、百亿边的巨图训练

丰富

涵盖Graph Neural Network和Graph Representation等前沿模型

PGL——易用性

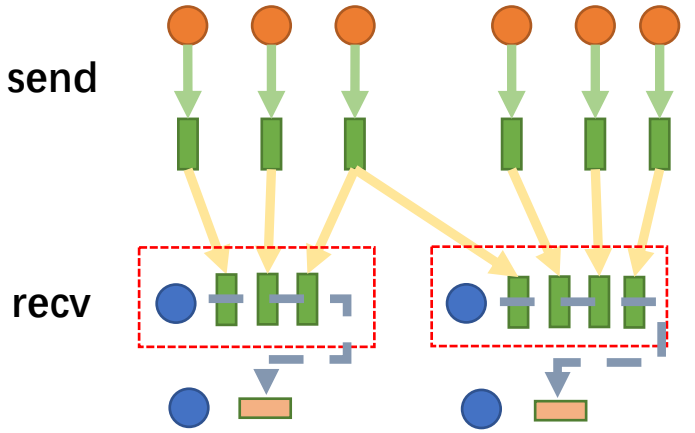
1. 建图方便

```
# construct a graph
num_nodes = 4
edge_list = [(0,2), (2,1), (1,0), (3,2)]
g = graph.Graph(num_nodes=num_nodes, # 节点数
                edges=edge_list,     # 边列表
                node_feat=None,      # 节点特征，可不写
                edge_feat=None)      # 边特征，可不写
```

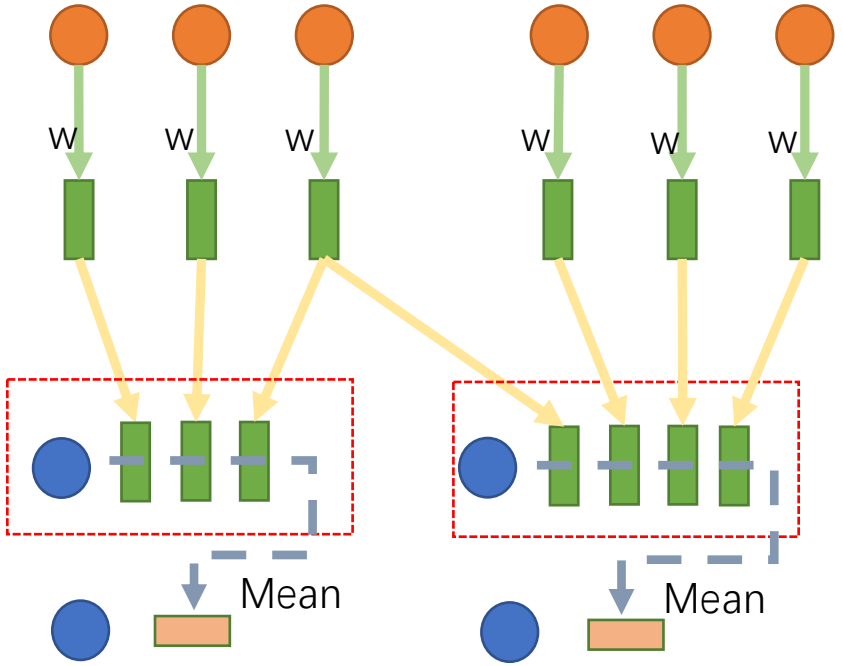
2. 采用Message Passing机制时，只需要Send/Recv实现图网络

```
# define message function
def send_func(src_feat, dst_feat, edge_feat):
    # In this tutorial, we return the feature vector of the source nodes.
    return src_feat['h']
```

```
# define reduce function
def recv_func(feat):
    # We sum the feature vector of the source nodes.
    return fluid.layers.sequence_pool(feat, pool_type='sum')
```



PGL——高效性

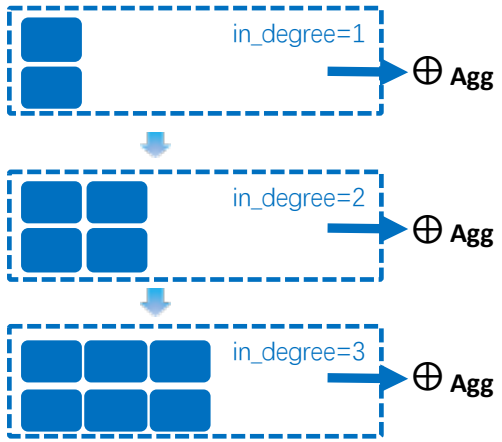


难点　节点获取的消息不定长

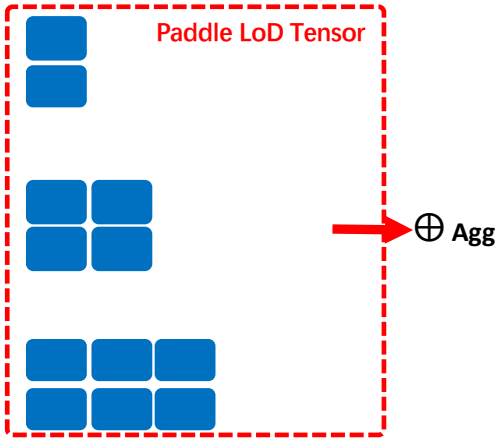
适配复杂的用户聚合函数

高效：业界首个并行消息聚合

主流框架的串行消息聚合



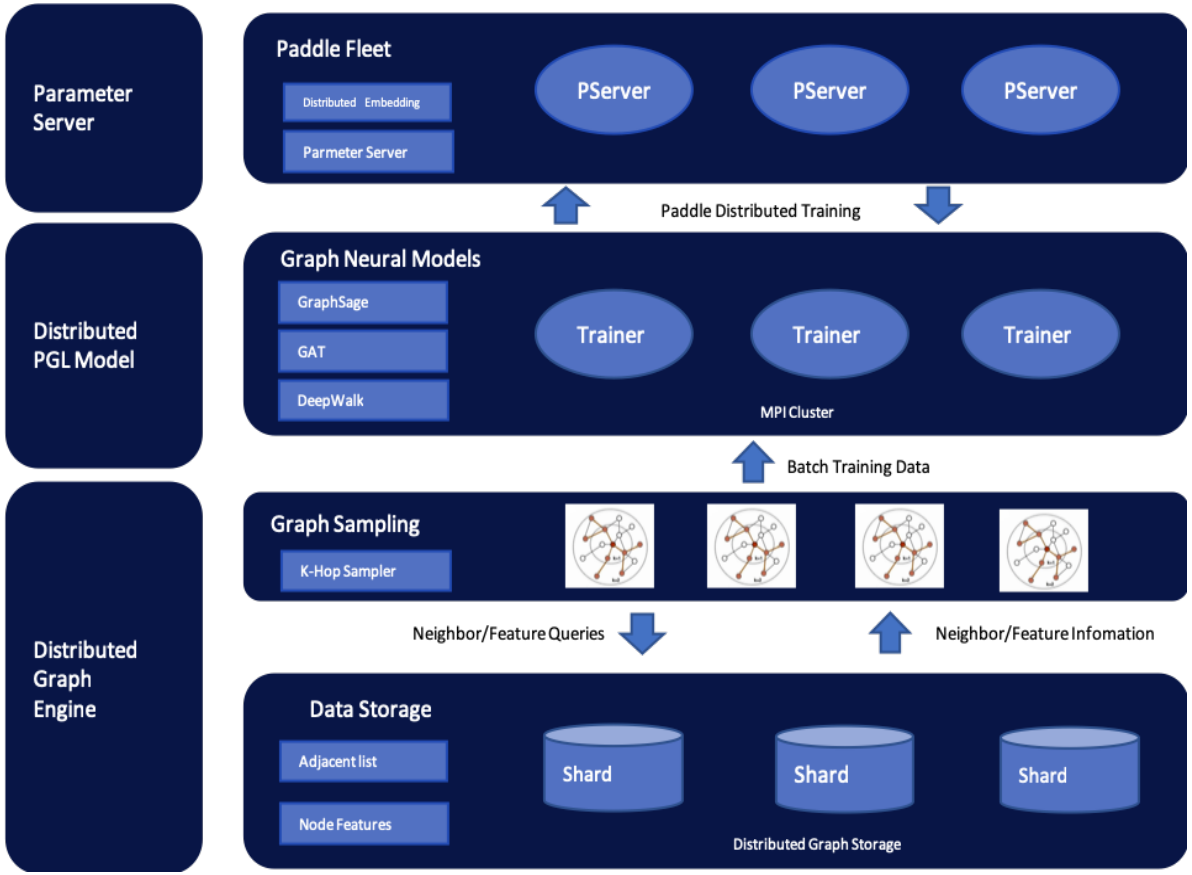
PGL并行消息聚合



PGL	复杂用户定义	LodTensor机制聚合	**并行计算**
	简单模型	Scatter-gather机制	并行计算
主流框架	复杂用户定义	Degree bucketing机制	**串行计算**

PGL——大规模

规模：PGL分布式解决方案——十亿节点、百亿边的巨图训练



- 满足超大规模节点Embedding需求
- Graph Embedding
- Graph Neural Network
- 解决数据Feed与训练速度差瓶颈
- 满足超大规模图分片存储需求

PGL——丰富性

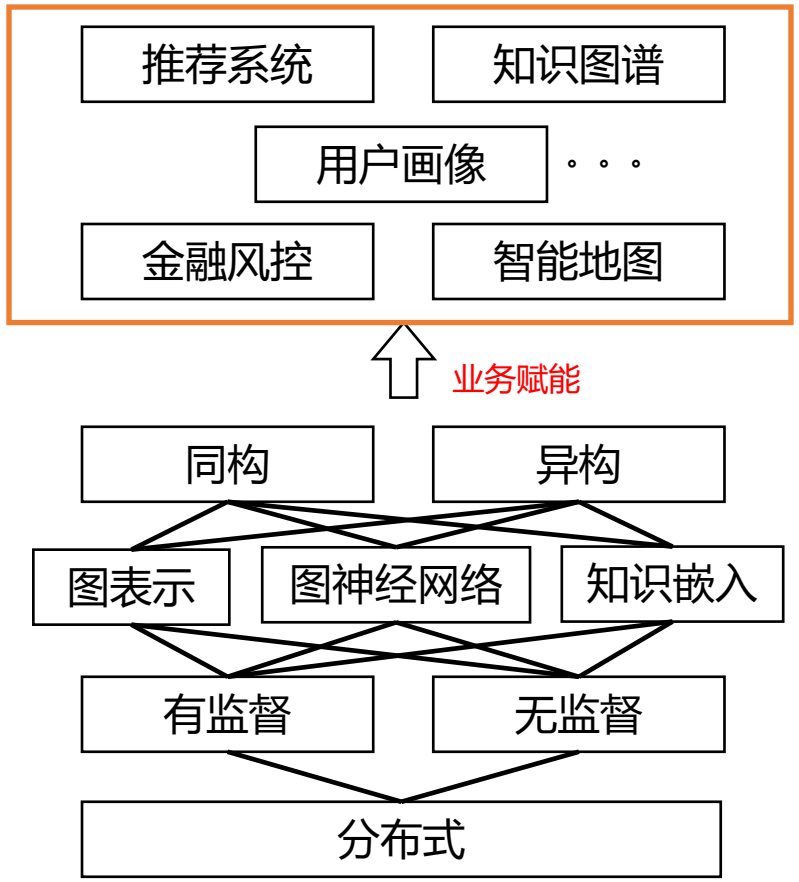
丰富：预置多种主流的图学习模型

编号	模型	特点
1	DeepWalk	DFS随机游走的表示学习
2	node2vec	结合DFS及BFS的表示学习
3	LINE	基于一阶、二阶邻居的表示学习
4	struc2vec	基于结构相似的表示学习
5	metapath2vec	基于元路径的表示学习
6	Mulit-metapath2vec++	自研：混合元路径的表示学习
7	GES	加入节点特征的图表示学习方法
8	GATNE	异构图的图表示学习
9	GAT	基于Attention的图卷积网络
10	GCN	图卷积网络
11	GIN	图同构网络
12	SGC	简化的图卷积网络
13	GaAN	引入门控 Attention 机制的图卷积网络
14	SAGPool	基于 self-attention 的图分类模型
15	STGCN	时空图卷积网络
16	GraphSAGE	基于邻居采样的大规模图卷积网络

编号	模型	特点
17	DeeperGCN	深层图卷积网络
18	GCNII	深层图卷积网络
19	APPNP	结合 GCN 和 PageRank 的传播模型
20	MixAggGNN	自研：基于混合聚合函数的图神经网络
21	GNN-Index	自研：结合语义索引技术的图神经网络
22	ERNIESage	自研：结合预训练语义模型的图神经网络
23	UniMP	自研：统一标签传递和图神经网络的统一模型
24	Distribute_DeepWalk	分布式随机游走
25	Distribute_GraphSAGE	分布式GrapSAGE
26	Distribute Pinsage	分布式Pinsage
27	Distribute_GIN	分布式GIN
28	Distribute_metapath2vec	分布式 metapath2vec
29	TransE	Translating Embedding
30	TransR	基于Relation空间的Translating Embedding
31	RotatE	基于Relational Rotation的Translating Embedding
32	BigBird	Sparse Transformer

PGL——百度落地应用举例

业务范围



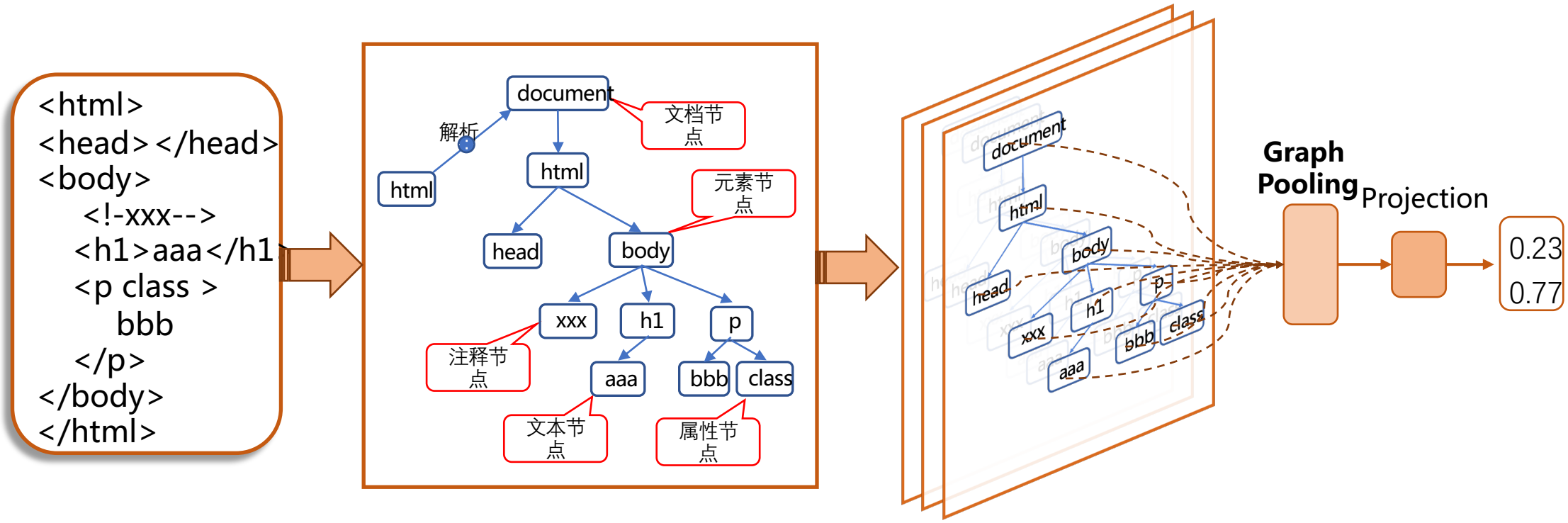
PGL——基于GAT的网页反作弊项目

网页作弊
- 内容作弊
- 框架作弊

是否作弊？

图分类

PGL——基于GAT的网页反作弊项目



PGL自研模型

1. ERNIESage: 落地百度内部多个应用，在COLING 2020 Text Graph-14 竞赛中获得第一；

Results

#	User	Entries	Date of Last Entry	Team Name	MAP ▲
1	webbley	26	09/21/20	Baidu PGL	0.6033 (1)
2	alvysinger	31	06/25/20		0.5843 (2)
3	aisys	9	07/10/20		0.5233 (3)

2. UniMP: 10月份在图神经网络OGB榜单三大半监督节点分类数据集刷新 SOTA；

Leaderboard for ogbn-products

The classification accuracy on the test and validation sets. The higher, the

Package: >=1.1.1

Rank	Method	Test Accuracy	Validation Accuracy	Contact
1	UniMP	0.8256 ± 0.0031	0.9308 ± 0.0017	Yunsheng Shi (PGL team)

第五节课会进行详细解释

Leaderboard for ogbn-arxiv

The classification accuracy on the test and validation sets. The higher, t

Package: >=1.1.1

Rank	Method	Test Accuracy	Validation Accuracy	Contact
1	UniMP_large	0.7379 ± 0.0014	0.7475 ± 0.0008	Yunsheng Shi (PGL team)

Leaderboard for ogbn-proteins

The ROC-AUC score on the test and validation sets. The higher, the

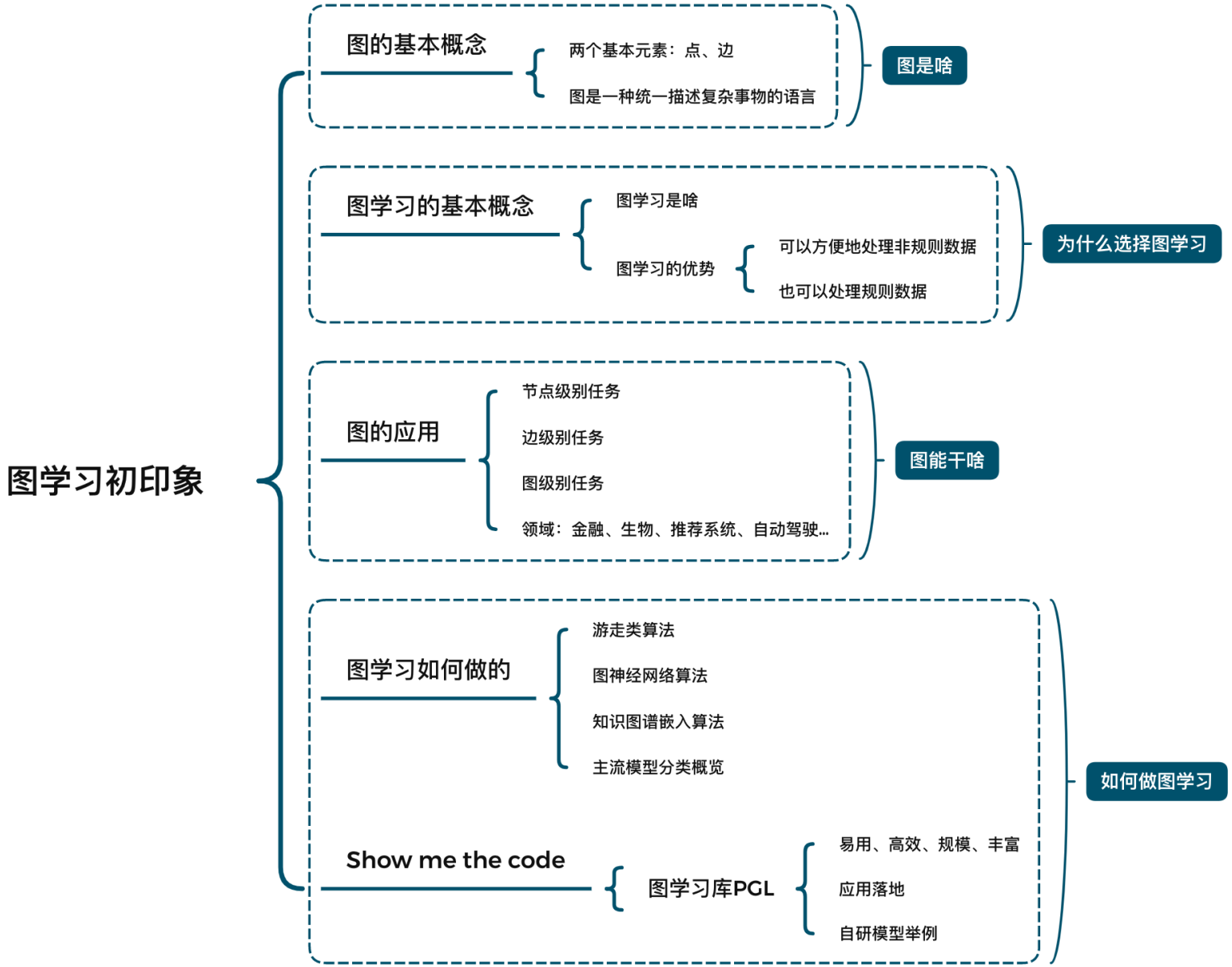
Package: >=1.1.1

Rank	Method	Test ROC-AUC	Validation ROC-AUC	Contact
1	UniMP	0.8642 ± 0.0008	0.9175 ± 0.0006	Yunsheng Shi (PGL team)

PGL运行实践

AIStudio Notebook 案例

总结


图学习初印象
图的基本概念
两个基本元素：点、边
图是一种统一描述复杂事物的语言
图是啥
图学习的基本概念
图学习是啥
图学习的优势
可以方便地处理非规则数据
也可以处理规则数据
为什么选择图学习
图的应用
节点级别任务
边级别任务
图级别任务
领域：金融、生物、推荐系统、自动驾驶...
图能干啥
图学习如何做的
游走类算法
图神经网络算法
知识图谱嵌入算法
主流模型分类概览
Show me the code
图学习库PGL
易用、高效、规模、丰富
应用落地
自研模型举例
如何做图学习

课后作业——跑通 PGL 的GCN 和 DeepWalk

Paddle 安装指引：详情可查看 PaddlePaddle 官网

```
pip install paddlepaddle==1.8.5 # -i https://pypi.tuna.tsinghua.edu.cn/simple/
pip install pgl
git clone --depth=1 https://github.com/PaddlePaddle/PGL
cd PGL/examples/gcn
python train.py
```

注意：

1. 如果在 pip 安装时遇到网络问题：
   - 使用清华源
   - https://pypi.org/ 下载 whl包或者源码包安装
2. 如果中途安装中断：
   - 重新安装即可
3. 独立的 python 环境：conda

详细要求：在 AIStudio 课程后查看~

```
# 本地：推荐使用conda创建独立python环境
conda create -n Paddle python=3.6 # 注意，3.8版本目前不支持~
conda activate Paddle
```

飞桨

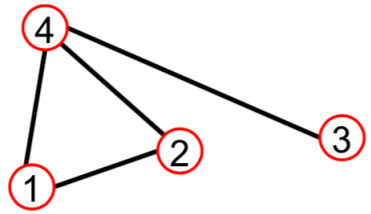
第一课补充 简单的图基础

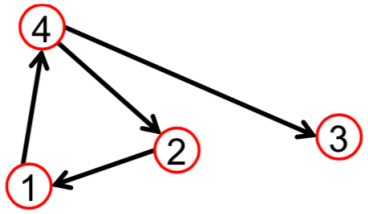
小斯妹 百度PGL团队成员

2020.11.23

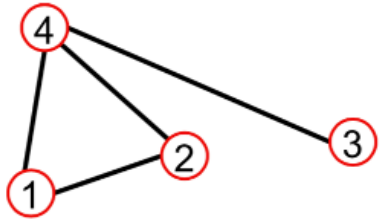
图的分类

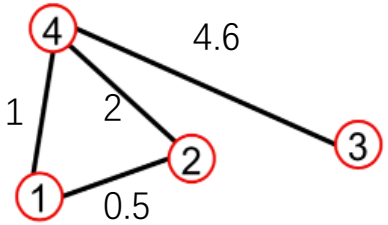
无向图 vs 有向图





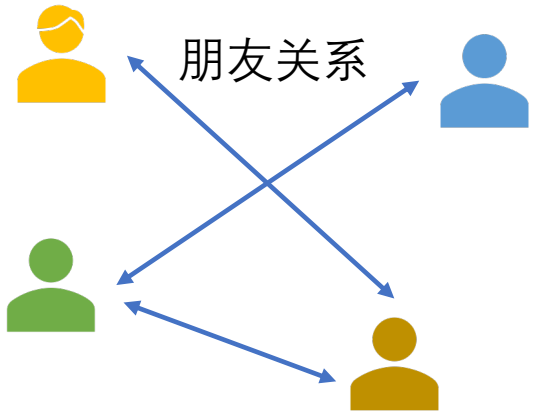
无权图 vs 有权图



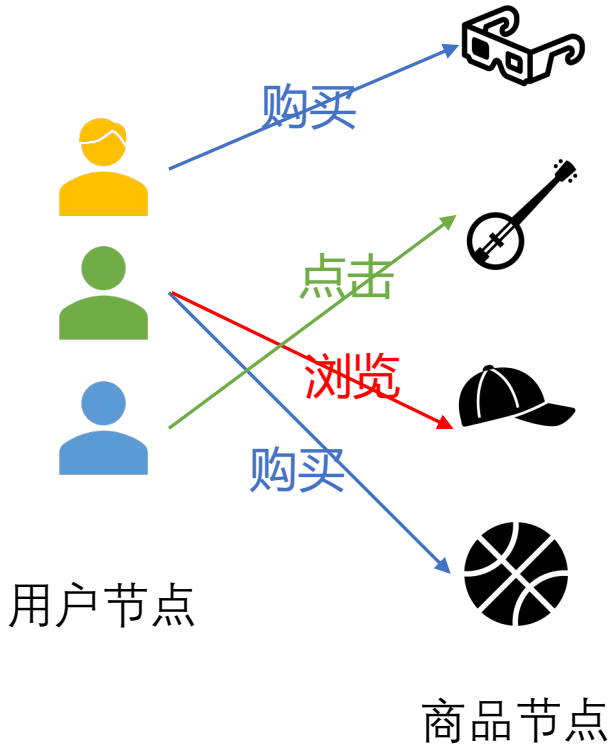


同构图 vs 异构图

同构图、异构图举例



同构图



异构图

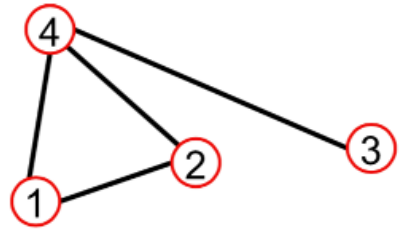
图的度和邻居

$G = (V, E)$

节点4

度(degree)　　邻居(neighbor)

无向图

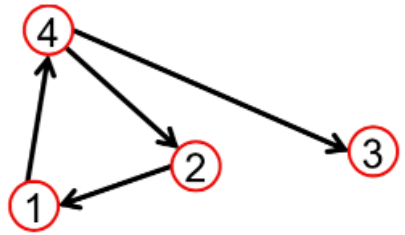


V={1, 2, 3, 4}
E={12, 21, 14, 41, 24, 42, 34, 43}

degree = 3

neighbors = 1, 2, 3

有向图

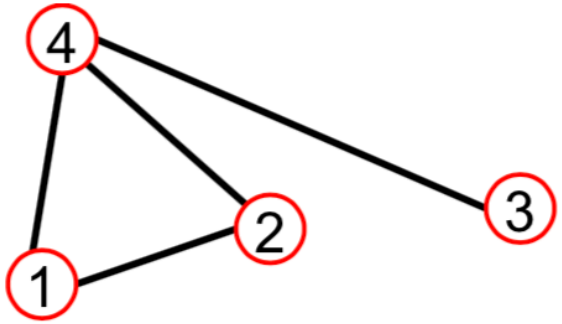


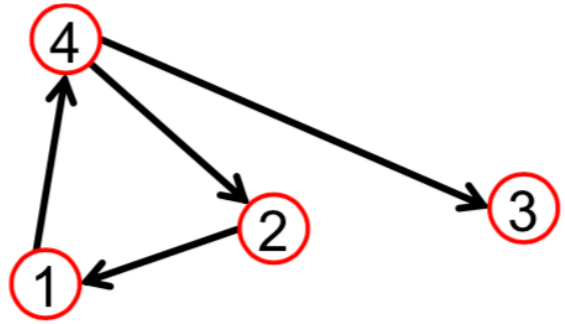
V={1, 2, 3, 4}
E={14, 21, 42, 43}

degree = 3
indegree = 1
outdegree = 2

predecessor = 1
successor = 2,3

图的表示——邻接矩阵





如果节点 i 和节点 j 之间有边： $A_{ij} = 1$

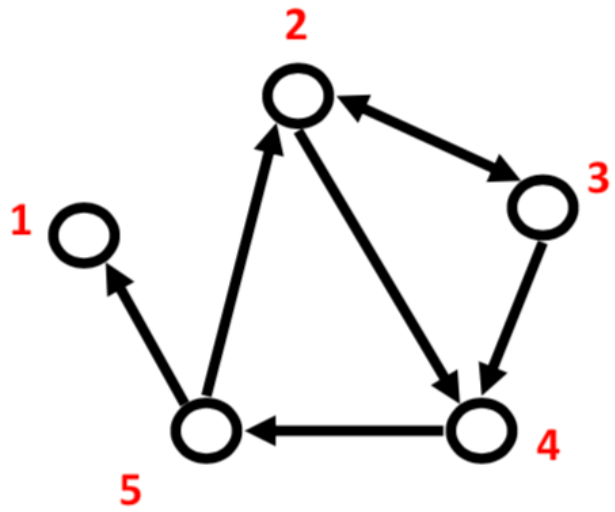
否则： $A_{ij} = 0$

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

无向图的邻接矩阵是对称矩阵。

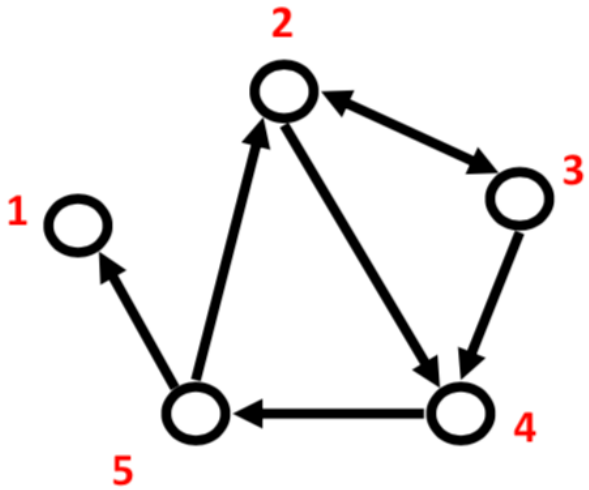
图的表示——邻接表



- 对于稀疏(sparse)大(large)图而言非常友好。
- 允许我们快速获得给定节点的所有邻居。

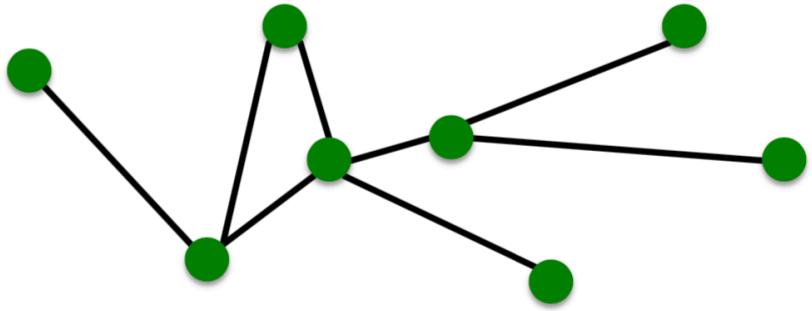
1:
2: 3, 4
3: 2, 4
4: 5
5: 1, 2

图的表示——边集



(2,3)
(2,4)
(3,2)
(3,4)
(4,5)
(5,1)
(5,2)

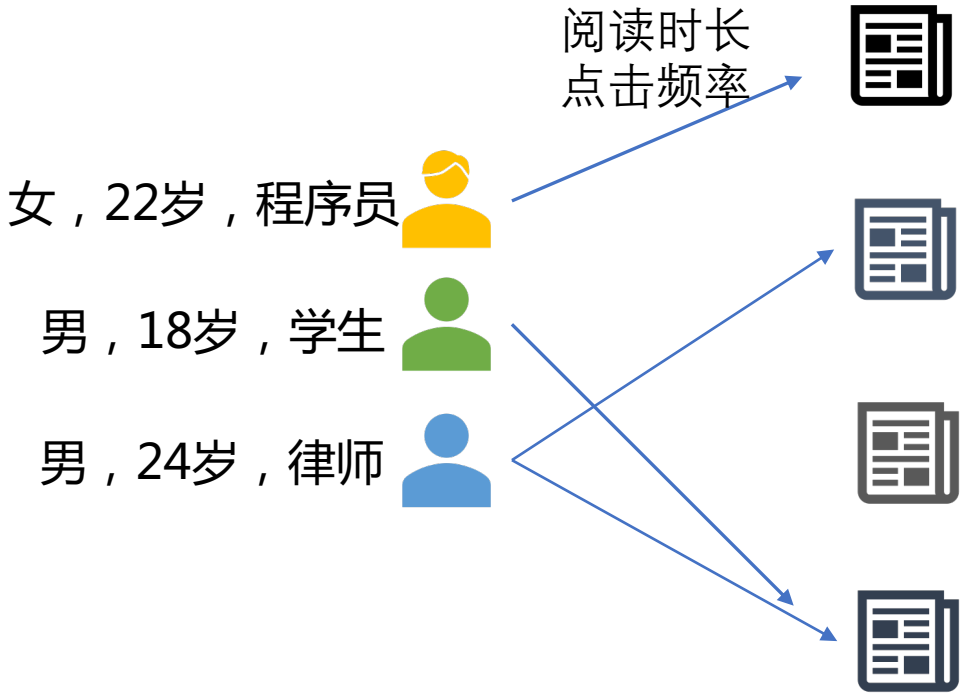
结构特征、节点特征、边特征

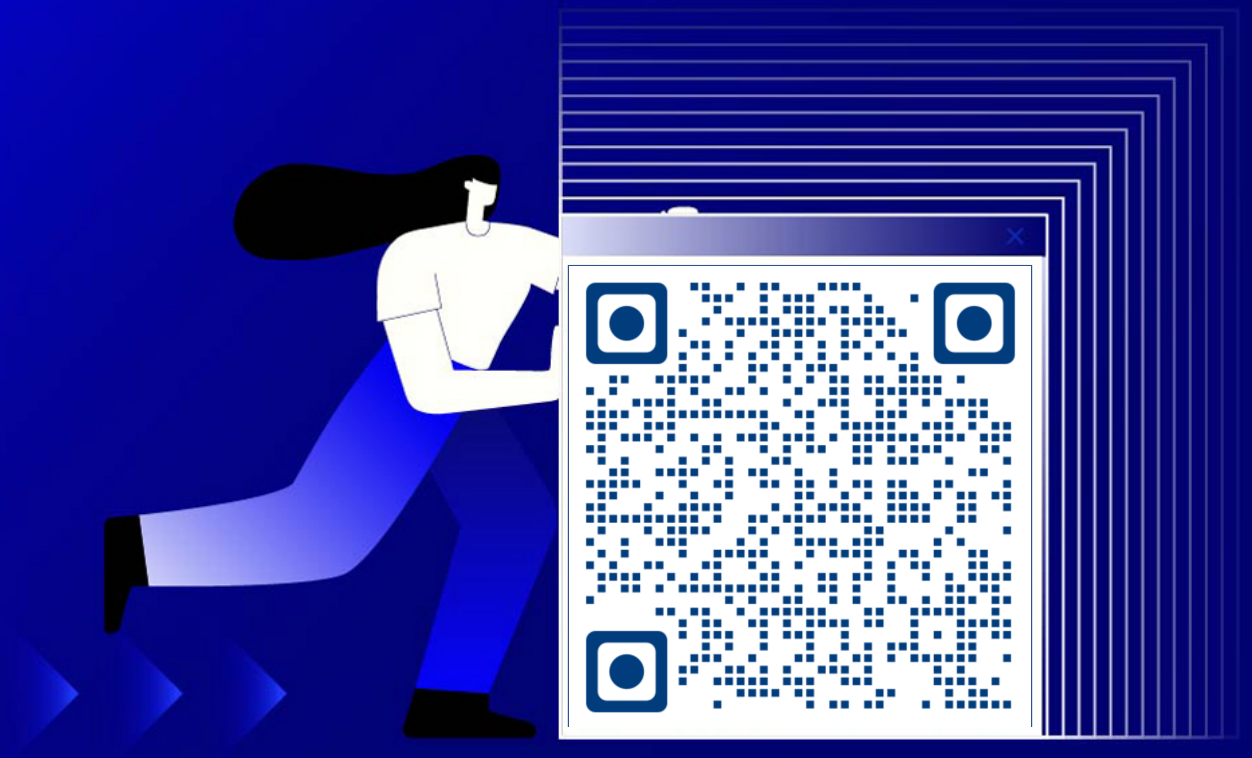


$G = (V, E)$

点的特征e，边的特征v

$G = (V, E, e, v)$





PGL github

谢谢观看