阿里云

2019阿里云峰会·上海
开发者大会
DEVELOPER CONFERENCE

AliGraph：大规模图神经网络平台

李永（九丰）

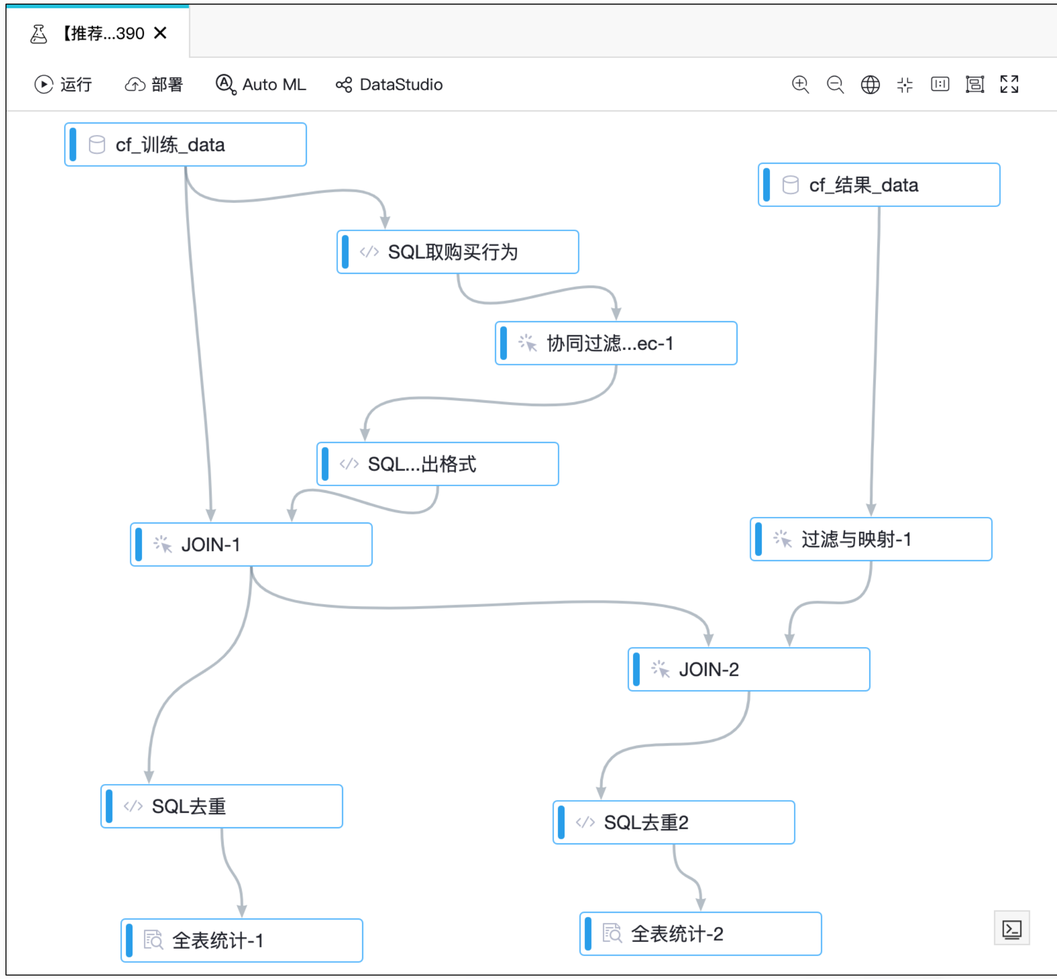
阿里巴巴资深技术专家
计算平台事业部

PAI机器学习平台介绍

https://data.aliyun.com/product/learn

简单易用 | 功能强大

PAI机器学习平台介绍



PAI机器学习平台特性

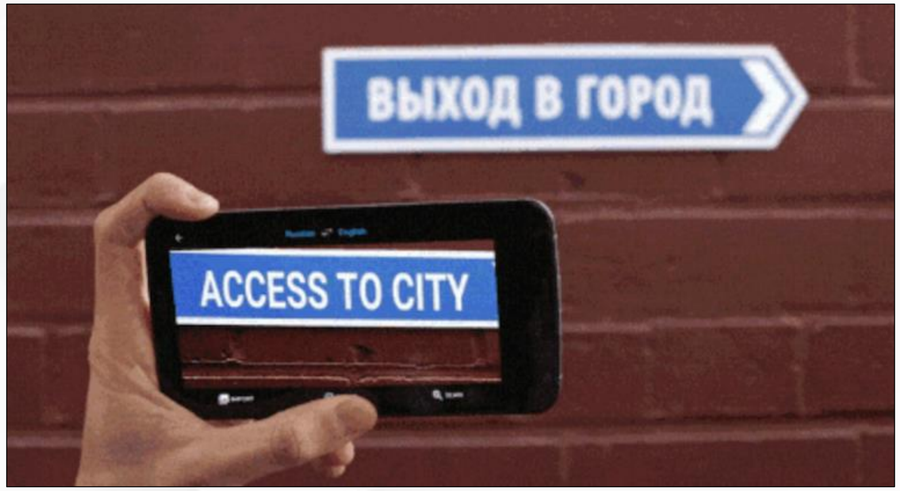
一站式
机器学习平台
数据处理，特征加工
模型训练，在线预测

STOA
机器学习算法
经典机器学习
阿里自研算法
兼顾离线&实时

精心优化的
深度学习引擎
阿里内部大规模任务打磨，
PAI-Tensorflow引擎

弹性
模型预测服务
简化模型部署负担
支持扩缩容，降低成本

深度神经网络面临的两个问题



模型可解释性

难以理解模型特征以及决策逻辑

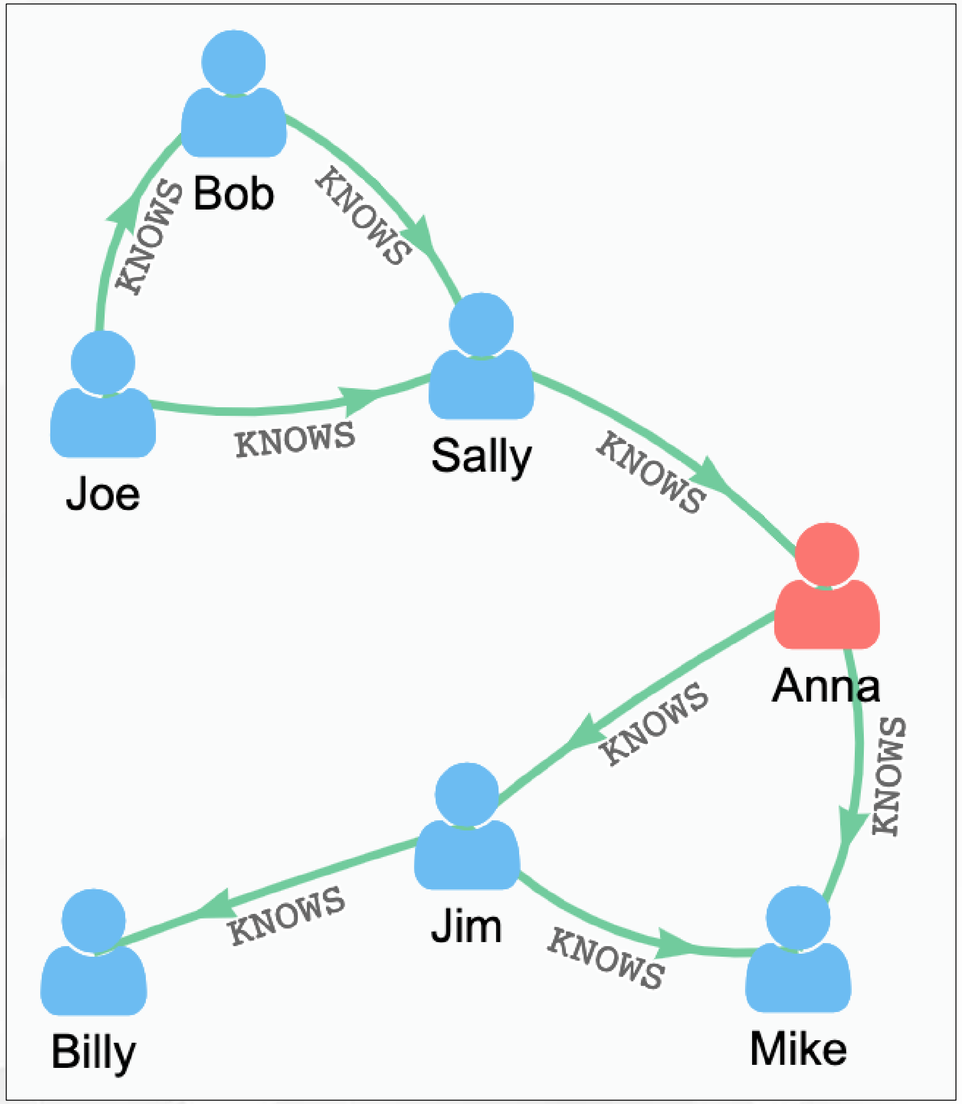
缺少数学工具来评测网络的表达能力

因果关系推理

美国增加关税会影响中国GDP吗

博士毕业的薪水会比硕士要高吗？

图



$$G = (V, E)$$

V = Vertex E = Edge

同构图	异构图
有向图	无向图

常见的图计算

图数据库

Neo4J，Titan

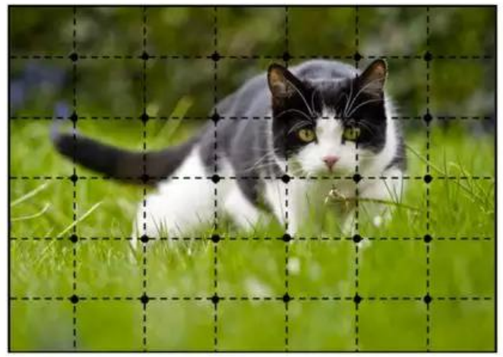
实时图分析查询

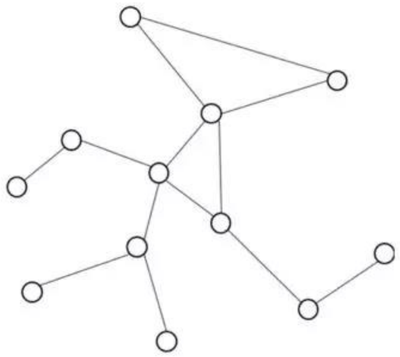
Aliyun GraphCompute， Amazon Neptune

离线图处理

Pregel， PowerGraph，Spark GraphX

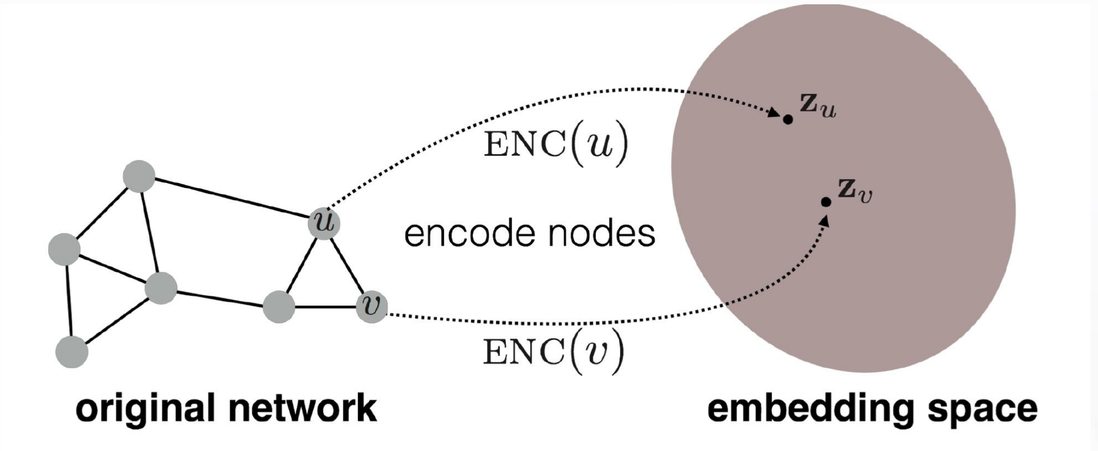
图+深度学习 = 图神经网络





非规则化的数据

Graph Embedding



图神经网络的应用领域

社交领域

推荐领域

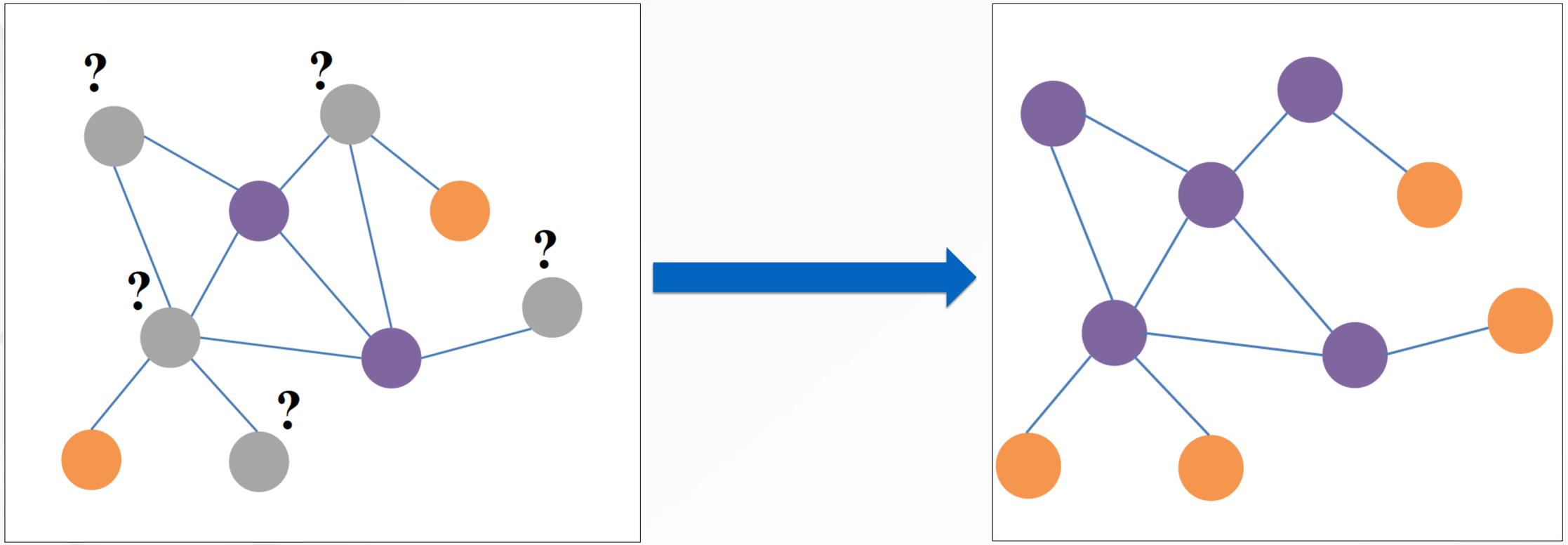
知识图谱

生命科学

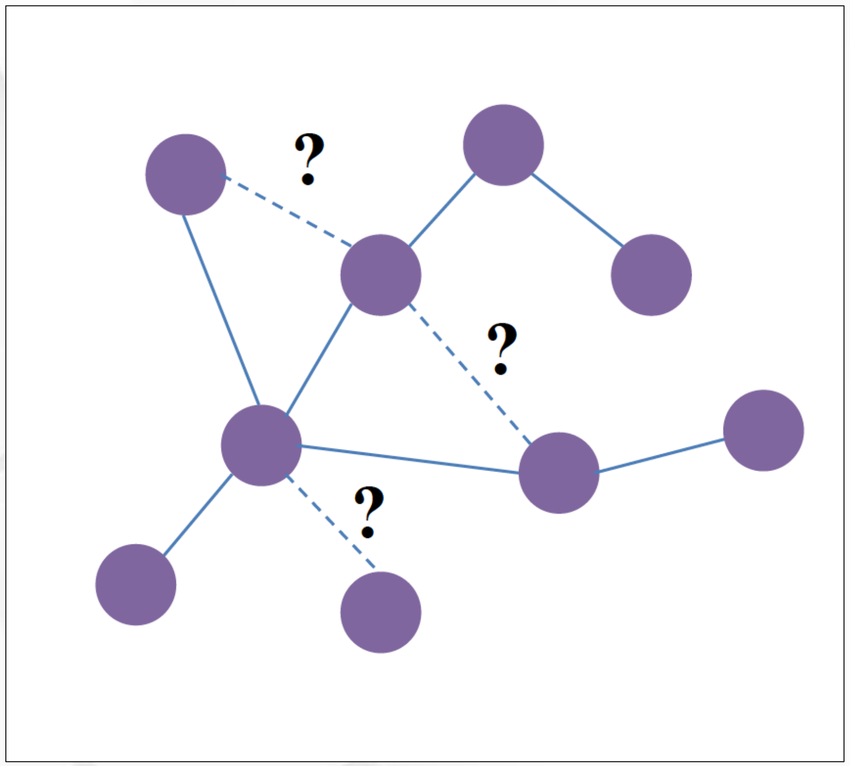
反作弊

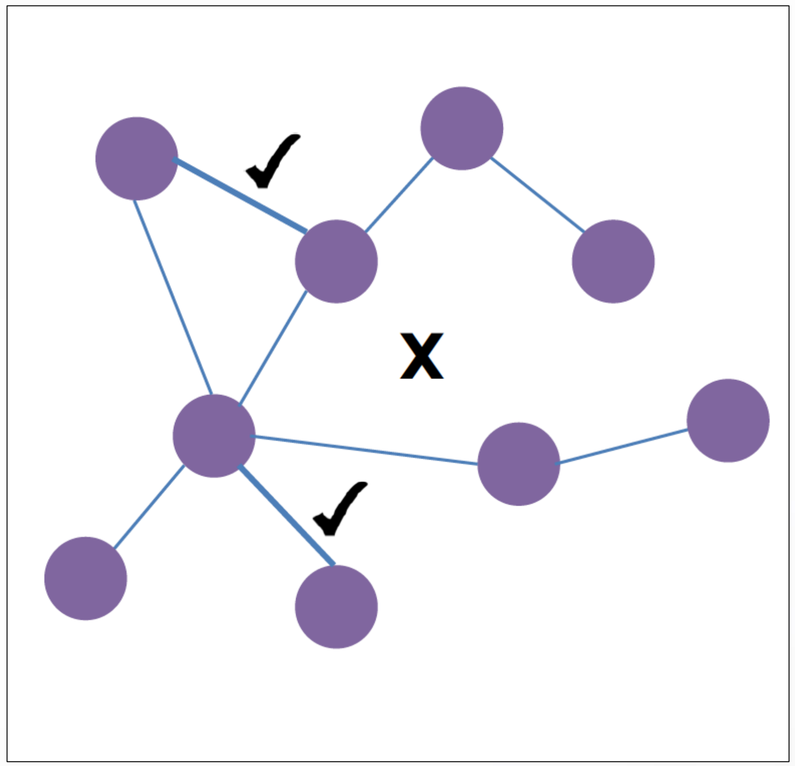
线上支付

GNN的应用 – 分类



GNN的应用 – 关系预测





GNN大规模应用的四大挑战

规模庞大

数百亿甚至数千亿点，数千亿甚至数万亿的边

点边异构

同构、异构、多边、多属性

属性丰富

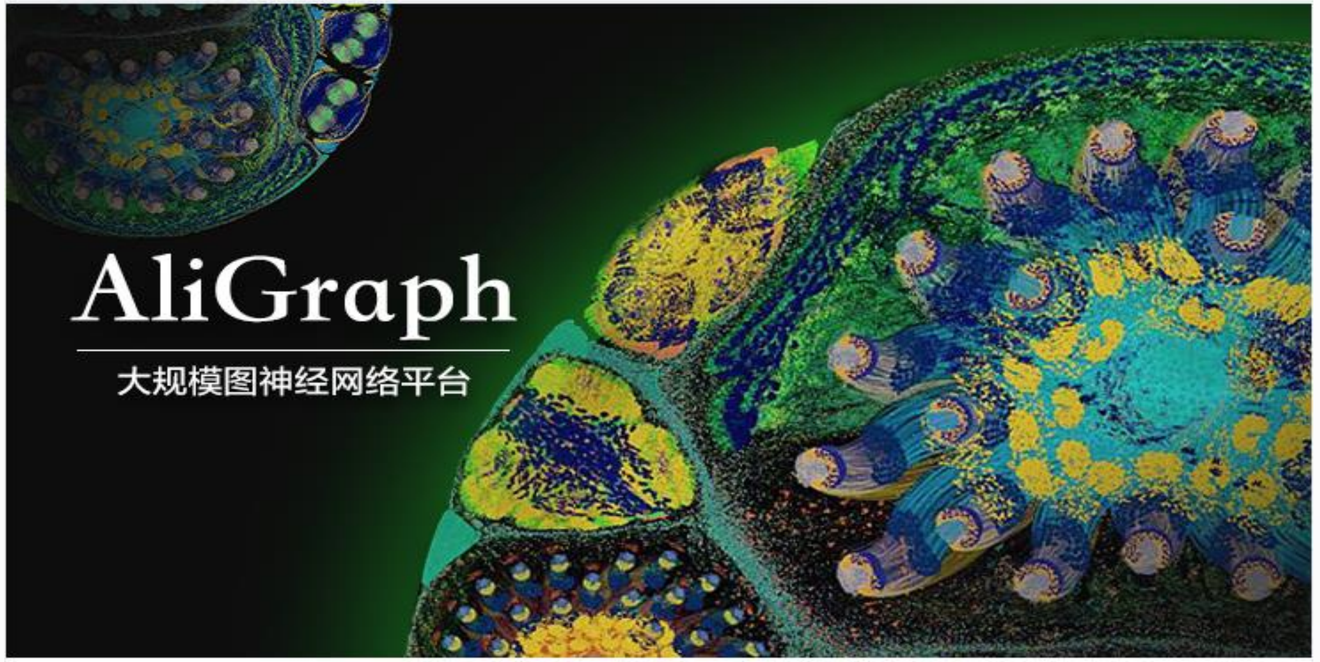
点属性，边属性

动态变化

节点、边的增删更新

AliGraph：大规模GNN平台

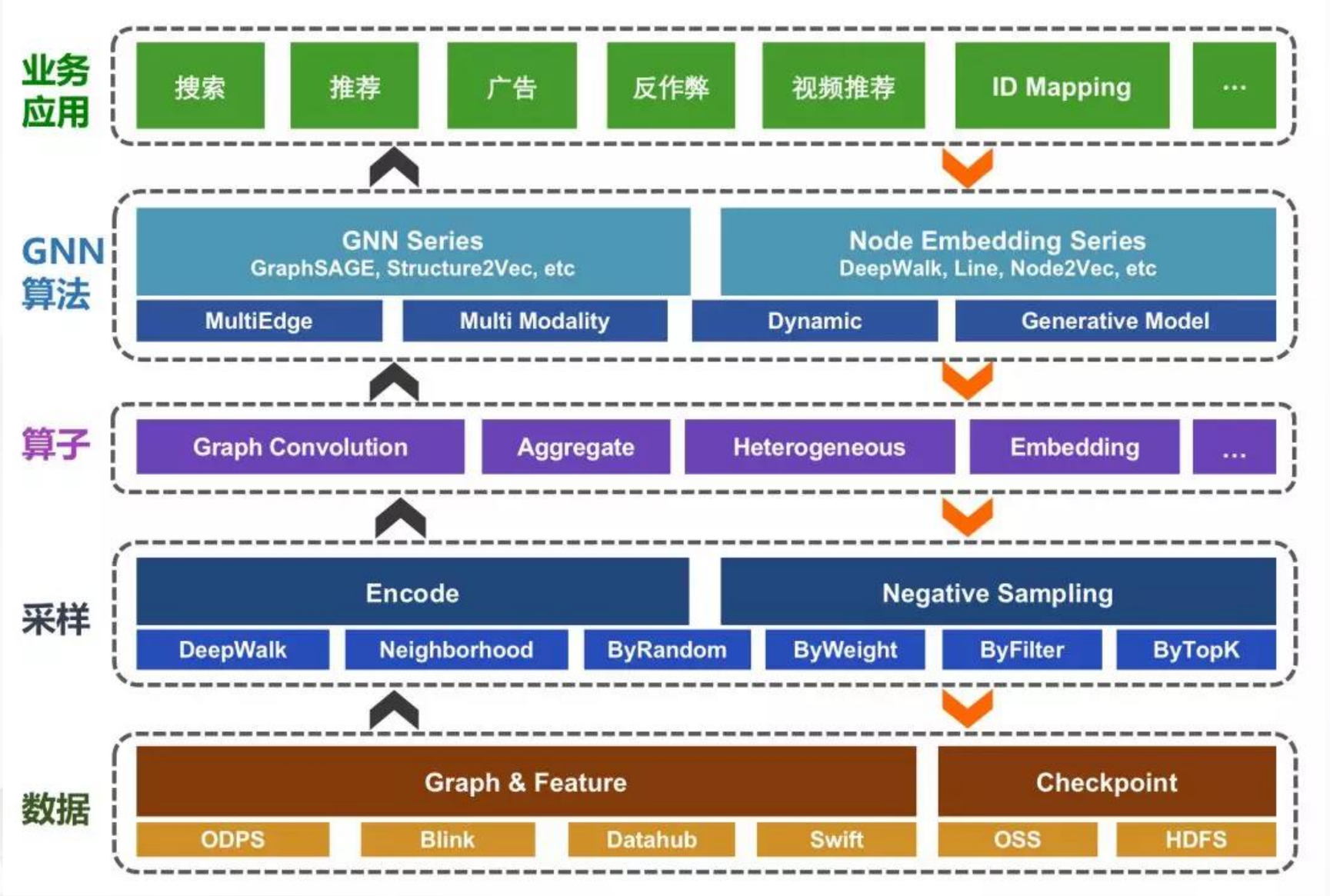
采样&建模&训练一体化的GNN平台



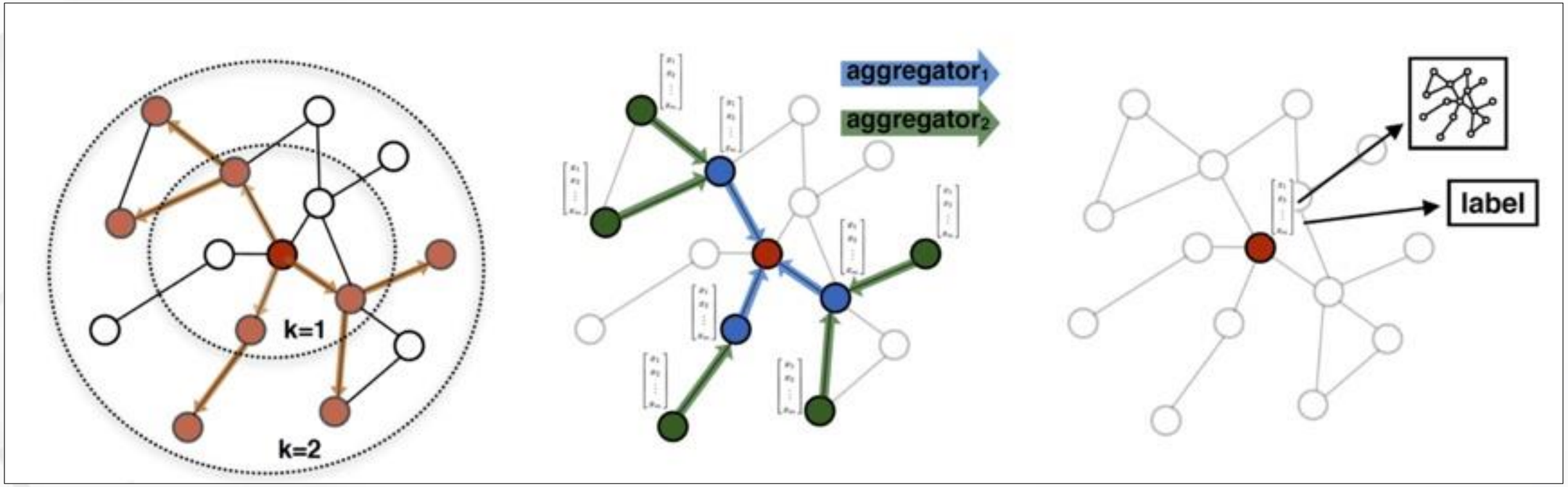
AliGraph：大规模图神经网络平台

AliGraph是新一代大规模图神经网络平台，其认知计算模型比起现有的深度学习技术有了突破性进展，被誉为人工智能2.0模型。

AliGraph系统架构



AliGraph编程范式



AliGraph编程范式

- sample
- aggregate
- combine

Algorithm 1: GNN Framework

Input: network $\mathcal{G}$, embedding dimension $d \in \mathbb{N}$, a vertex feature $\mathbf{x}_v$ for each vertex $v \in \mathcal{V}$ and the maximum hops of neighbors $k_{max} \in \mathbb{N}$.

Output: embedding result $\mathbf{h}_v$ of each vertex $v \in \mathcal{V}$

$\mathbf{h}_v^{(0)} \leftarrow \mathbf{x}_v$

for $k \leftarrow 1$ *to* k_{max} **do**

  for *each vertex* $v \in \mathcal{V}$ **do**

    $S_v \leftarrow \text{SAMPLE}(Nb(v))$

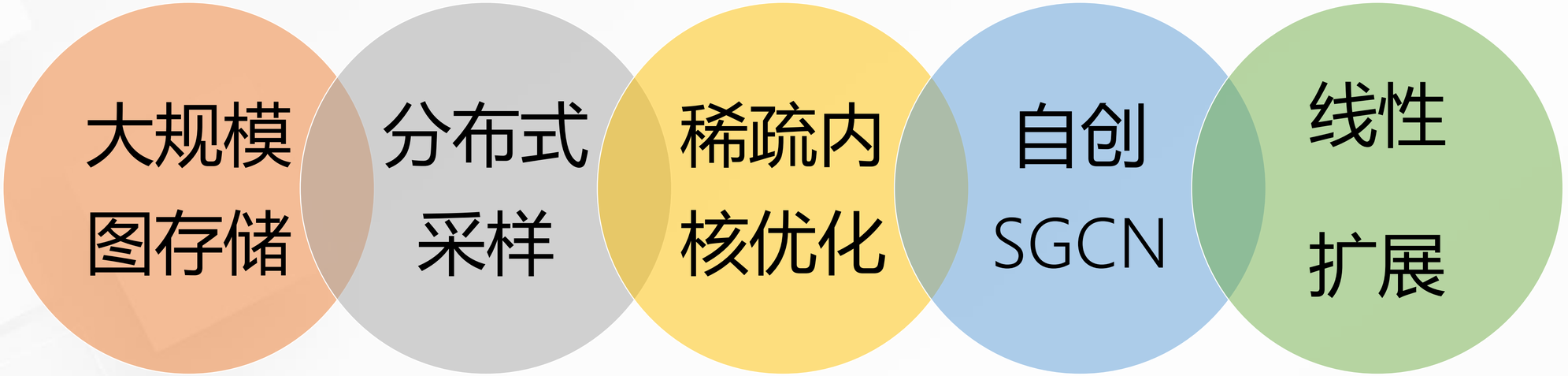
    $\mathbf{h}'_v \leftarrow \text{AGGREGATE}(\mathbf{h}_u^{(k-1)}, \forall u \in S)$

    $\mathbf{h}_v^{(k)} \leftarrow \text{COMBINE}(\mathbf{h}_v^{(k-1)}, \mathbf{h}'_v)$

  normalize all embedding vectors $\mathbf{h}_v^{(k)}$ for all $v \in \mathcal{V}$

$\mathbf{h}_v \leftarrow \mathbf{h}_v^{(k_{max})}$ for all $v \in \mathcal{V}$ return $\mathbf{h}_v$ as the embedding result for all $v \in \mathcal{V}$

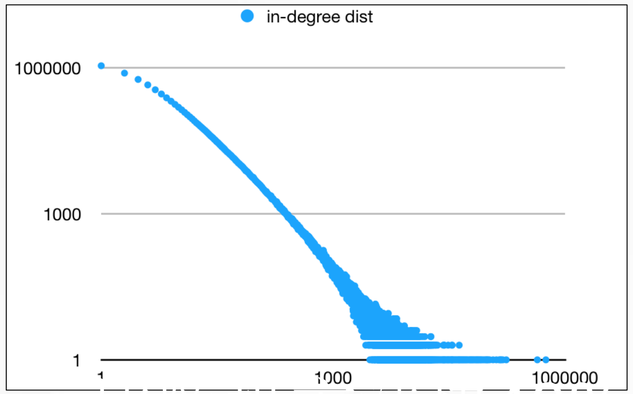
AliGraph五大特点



基于内存的图存储

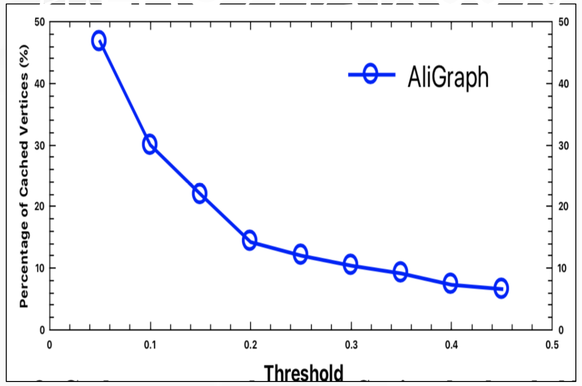
- 分布式的图存储
- 支持百亿点的规模，可伸缩
- 基于Vertex Cut的分片
- Worker基于出入度的缓存优化

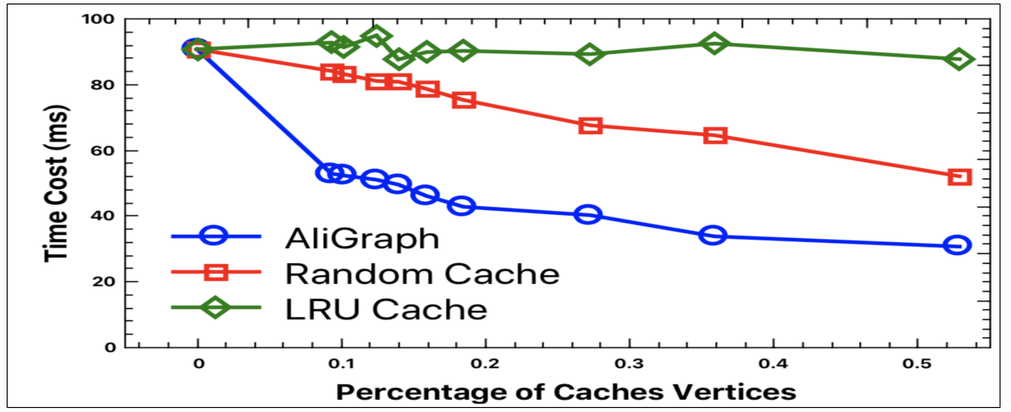
缓存效果



Input: graph $\mathcal{G}$, partition number p, cache depth h, threshold $\tau_1, \tau_2, \ldots, \tau_h$
Output: p subgraphs
Initialize p graph servers
for *each edge* $e = (u, v) \in \mathcal{E}$ **do**
  $j = \text{ASSIGN}(u)$
  Send edge e to the j-th partition
for *each vertex* $v \in V$ **do**
  for $k \leftarrow 1$ *to* h **do**
    Compute $D_i^{(k)}(v)$ and $D_o^{(k)}(v)$
    if $\frac{D_i^{(k)}(v)}{D_o^{(k)}(v)} \geq \tau_k$ **then**
      Cache the 1 to k-hop out-neighbors of v on each partition where v exists

缓存加速：比随机方法快40%-50%，比LRU方法快50%-60%





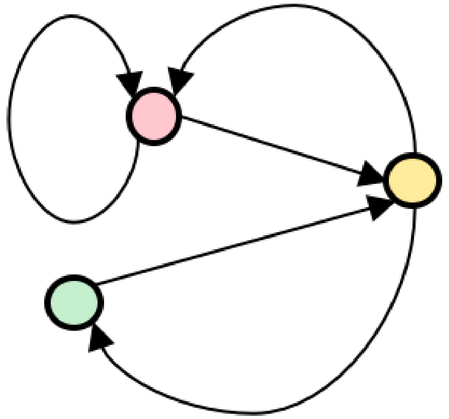
缓存加速：比随机方法快40%-50%，比LRU方法快50%-60%

分布式采样

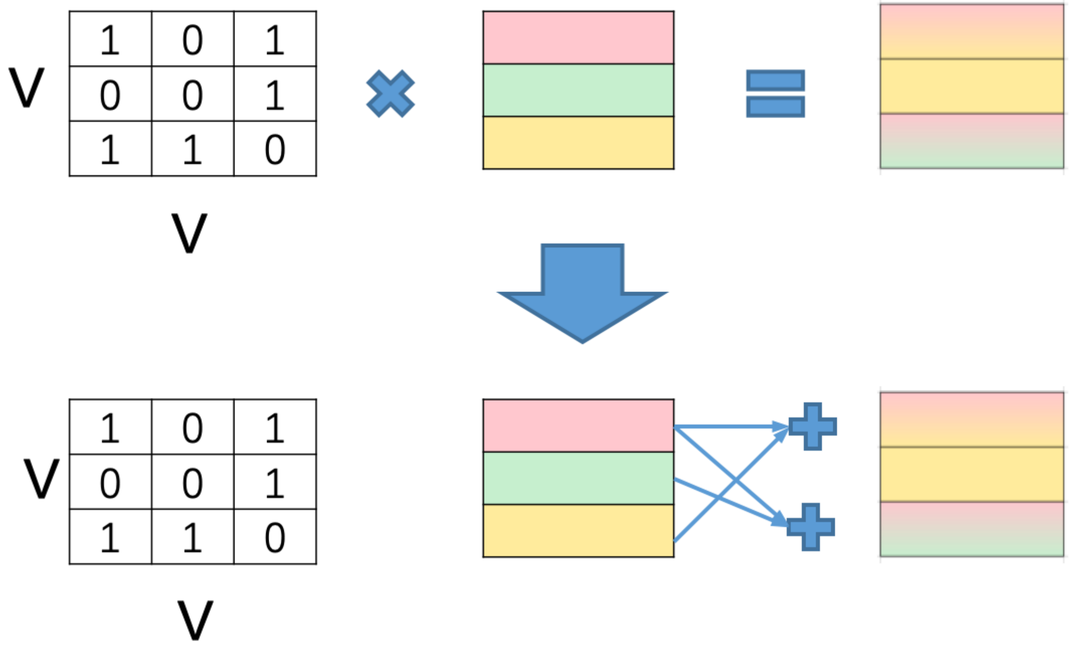
- 支持丰富的采样策略
- 支持多跳的采样功能
- 毫秒级的采样性能
- 模块化的采样设计
- 本地/全局的负采样

稀疏内核优化

图连接关系：

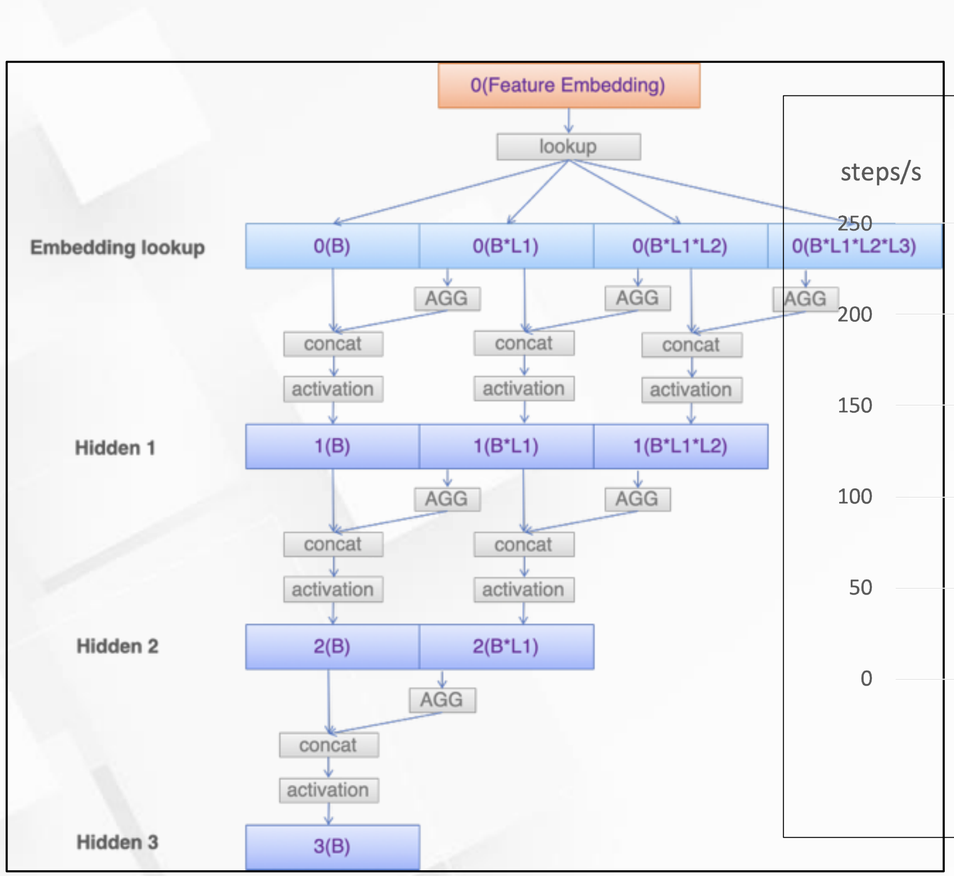


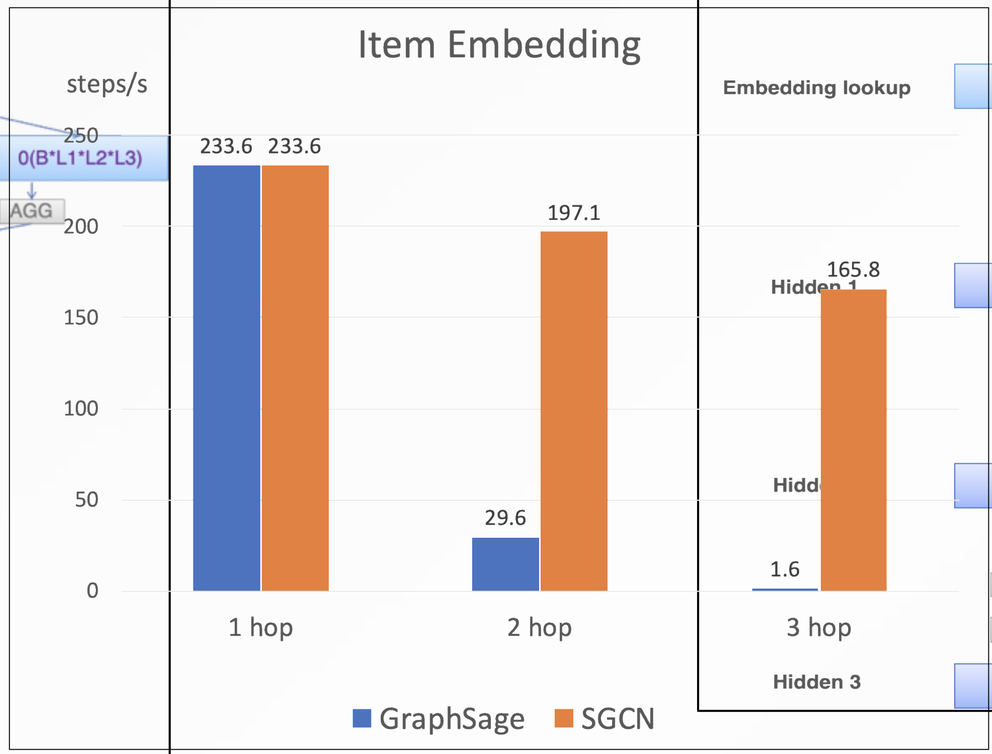
图矩阵表示：

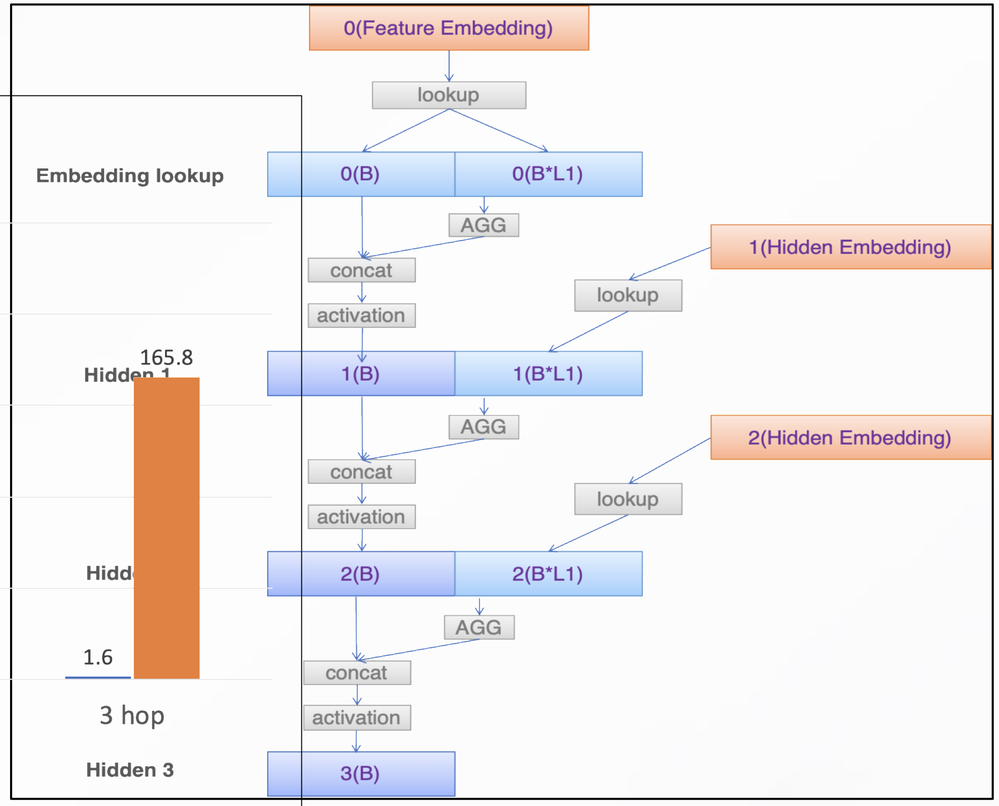


加速原始版本GCN达3X!

自创SGCN

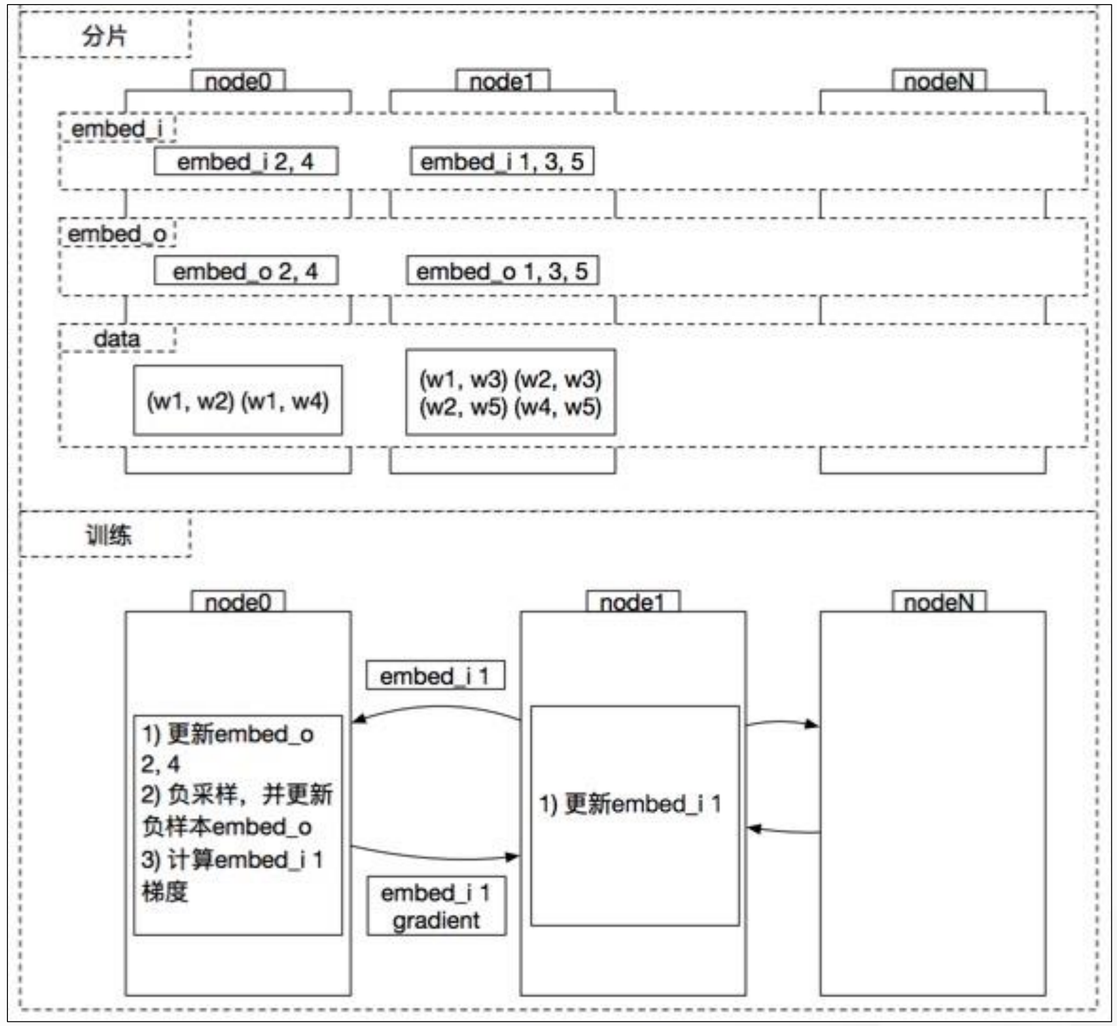

0(Feature Embedding)
lookup
Embedding lookup
0(B)
0(B*L1)
0(B*L1*L2)
0(B*L1*L2*L3)
AGG
concat
activation
Hidden 1
1(B)
1(B*L1)
1(B*L1*L2)
Hidden 2
2(B)
2(B*L1)
Hidden 3
3(B)


Item Embedding
steps/s
250
200
150
100
50
0
233.6
233.6
29.6
197.1
1 hop
2 hop
GraphSage
SGCN


Embedding lookup
Hidden 1
Hidd
Hidden 3
165.8
1.6
3 hop
0(Feature Embedding)
lookup
0(B)
0(B*L1)
AGG
concat
activation
1(Hidden Embedding)
1(B)
1(B*L1)
2(Hidden Embedding)
2(B)
2(B*L1)
3(B)

分布式扩展

- 去中心化
- Distribute Negative Sampling
- 多机异步训练



AliGraph API

- 以点为中心
- 边表混合
- 采样&建模&训练一体

Field	Type	Label	Comment
node_id	bigint		
feature	string		

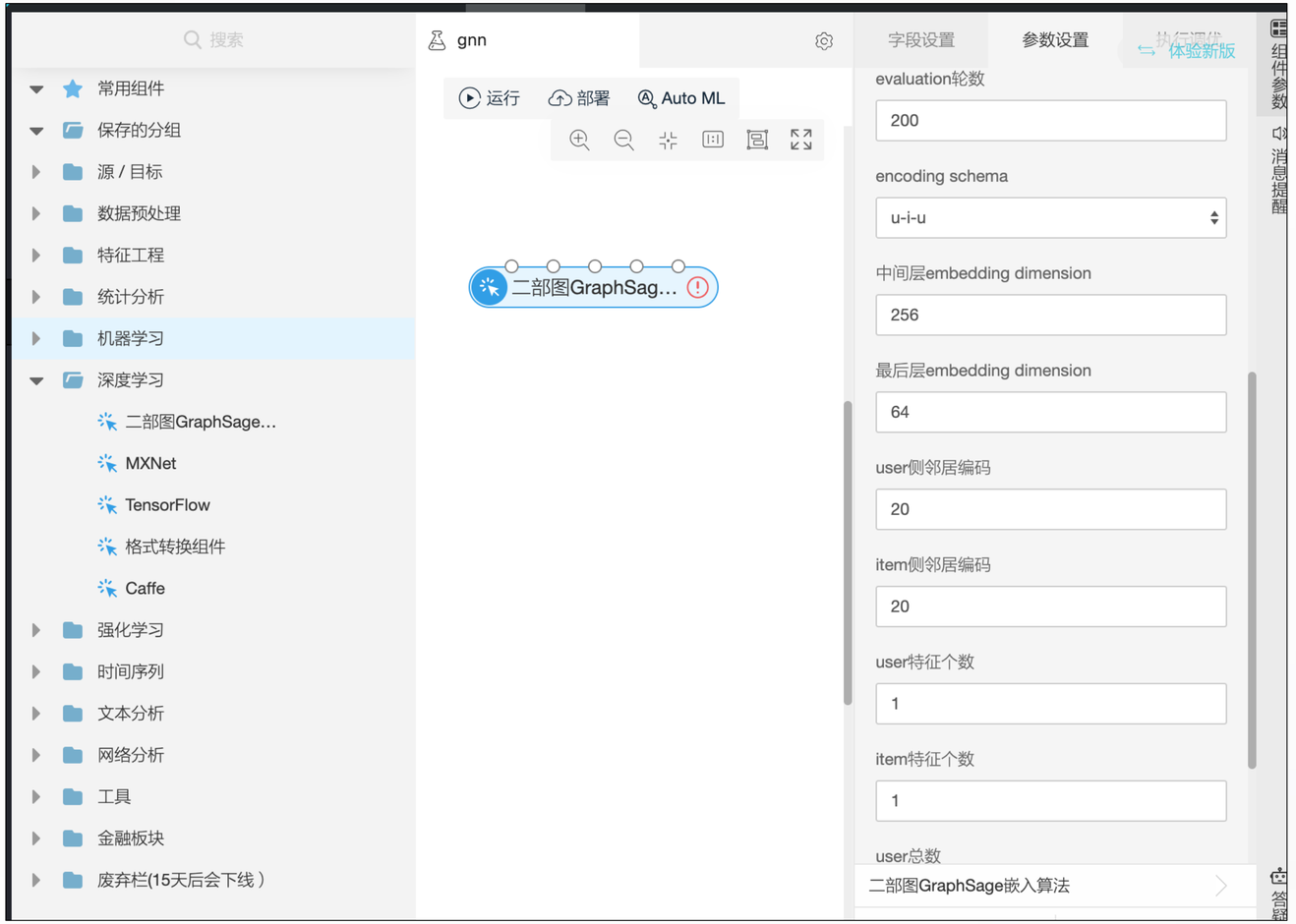
Field	Type	Label	Comment
src_id	bigint		
src_type	string		
dst_id	bigint		
dst_type	string		
edge_type	string		
val	double		

```
g = ss.Graph()
        .add_nodes("item")
        .sample("topk")
        .batch(128)
        .model(GraphSage())

g.train()
```

如何使用


机器学习PAI交流5群
71人
扫一扫群二维码，立刻加入该群。


搜索
gnn
常用组件
保存的分组
源 / 目标
数据预处理
特征工程
统计分析
机器学习
深度学习
二部图GraphSage...
MXNet
TensorFlow
格式转换组件
Caffe
强化学习
时间序列
文本分析
网络分析
工具
金融板块
废弃栏(15天后会下线)
运行
部署
Auto ML
二部图GraphSag...
字段设置
参数设置
体验新版
evaluation轮数
200
encoding schema
u-i-u
中间层embedding dimension
256
最后层embedding dimension
64
user侧邻居编码
20
item侧邻居编码
20
user特征个数
1
item特征个数
1
user总数
二部图GraphSage嵌入算法
组件参数
消息提醒

阿里云 2019阿里云峰会·上海 开发者大会

阿里云开发者社区

扫码加入社群
与志同道合的码友一起
Code Up

大数据计算开发者...

该群属于"阿里云ACE"部门群，仅组织内部成员可以加入，如果组织外部人员收到此分享，需要先申请加入该组织。

谢谢！