DataFun.

图算法在风控以及平台能力建设

汪浩然 互联网行业资深风控和图计算专家

目录 CONTENT

DataFun.

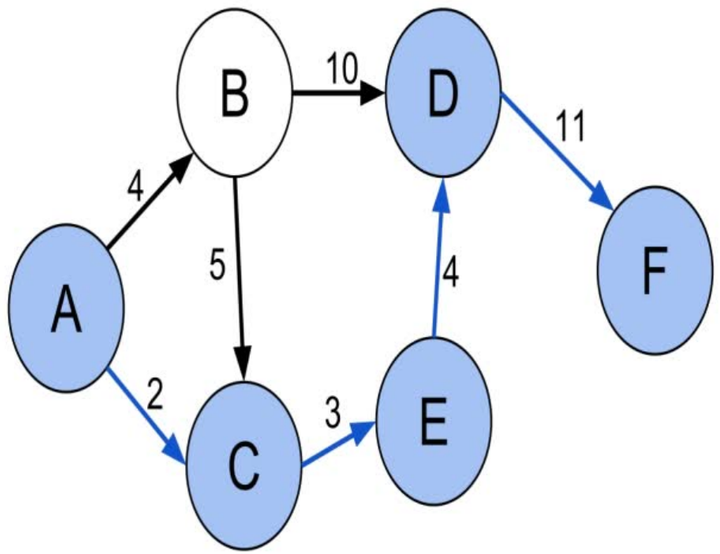
DataFun.

01

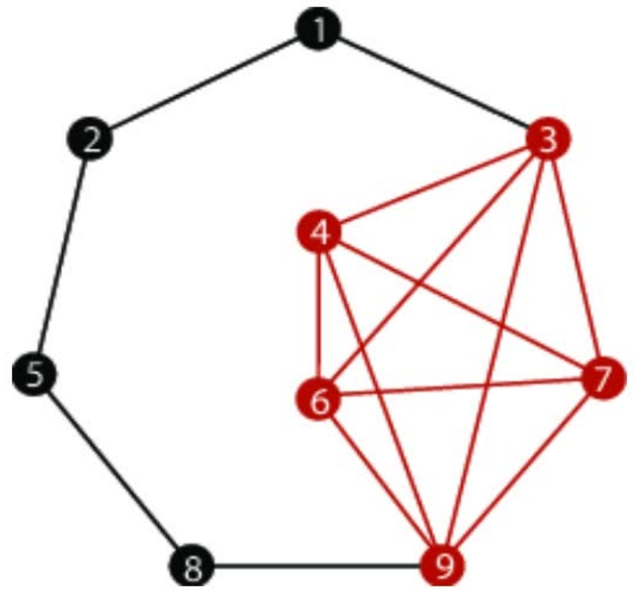
图算法和风控简介

什么是图算法-图论算法

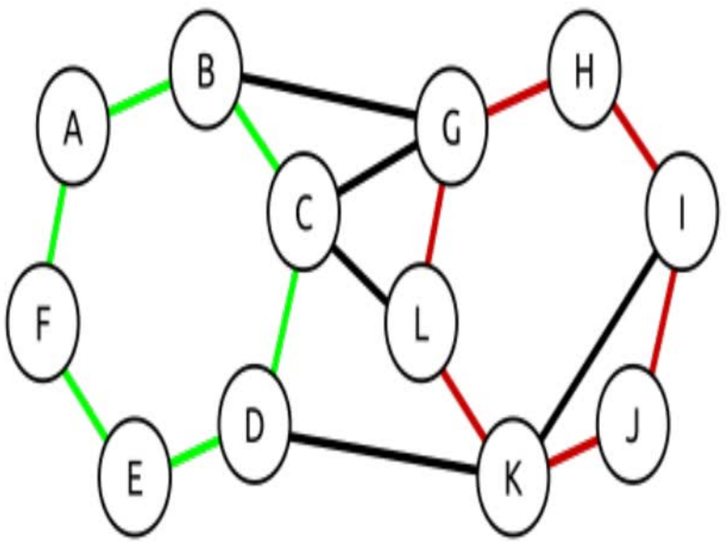
最短路径发现



团的识别发现



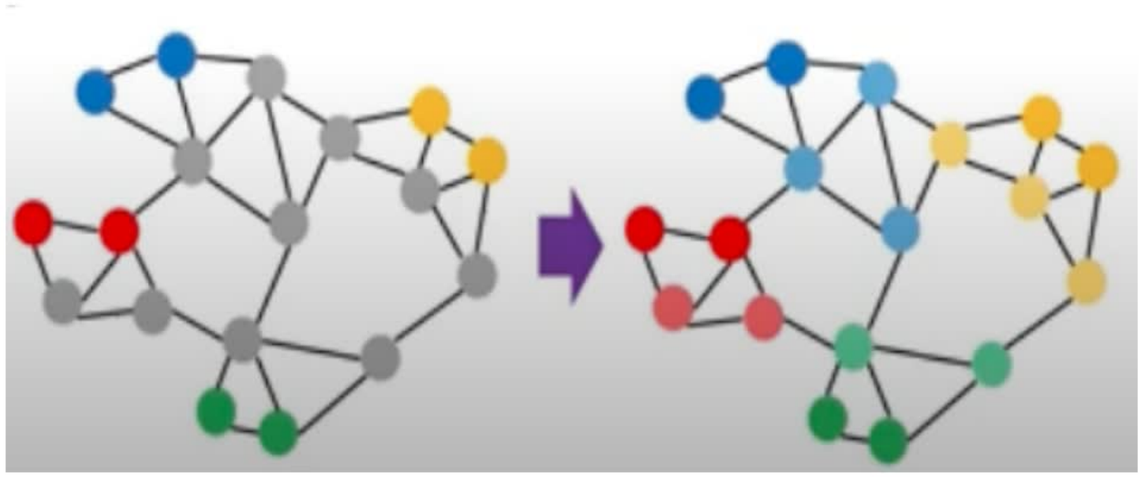
圈的识别发现



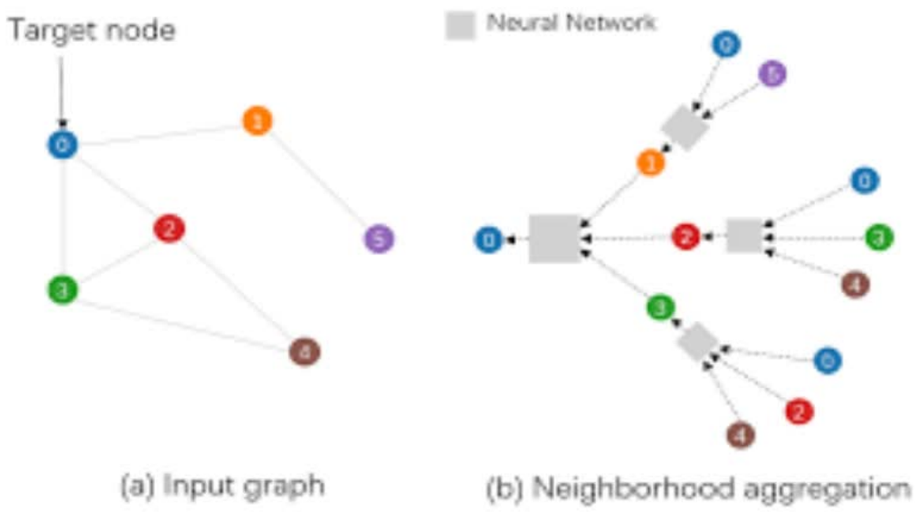
DataFun.

什么是图算法-图机器学习

标签传播算法

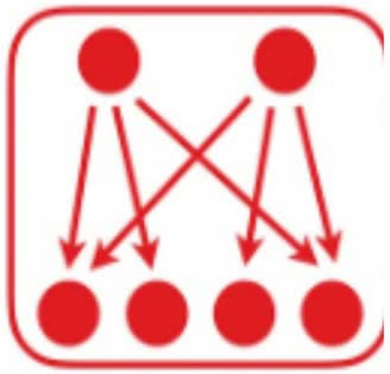


图神经网络算法

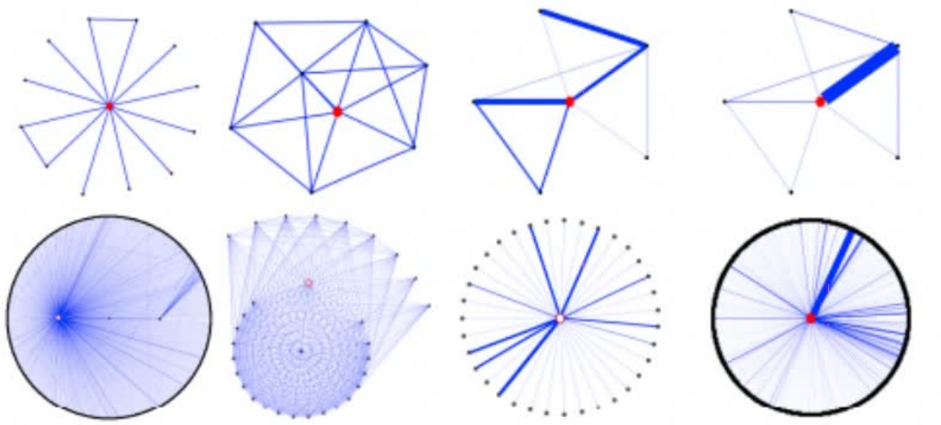


什么是图算法-图挖掘算法

高密度子图发现

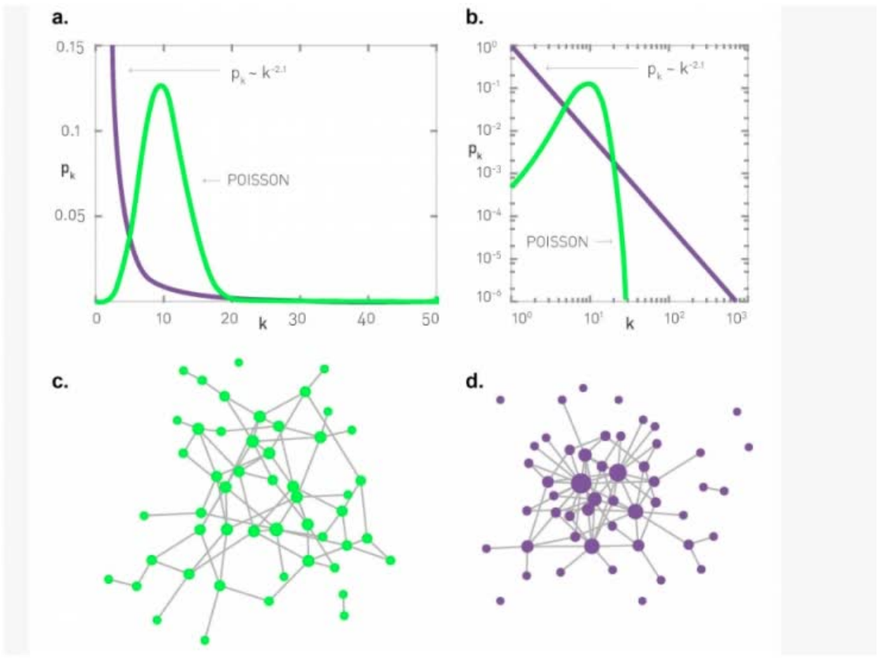


邻居域异常



(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

度分布模式



什么是风控

互联网风控干什么


羊毛党
薅羊毛


密码
你的微信帐号于18:27在HUAWEI-HUAWEI C8813Q设备上通过微信密码登录。如果这不是你的操作，你的微信帐号已经被盗号者登录。请尽快登录微信修改微信密码。或访问weixin110.qq.com冻结微信。
确定


现金贷
利滚利
以贷还贷
知乎 @米袋金融


异地上线：买家查询物流信息，会发现所购商品始发地为海外或者上海、深圳等地，但实际不是。
什么值得买


贷款诈骗
冒充客服诈骗
婚恋交友诈骗
刷单扫码诈骗
虚假购物诈骗

图算法和风控的相遇

“作案”有团伙特性

“作案”有相似性

“作案”需要大量账号和设备资源配合

“作案”具有成本因素

物以类聚，人以群分

DataFun.

02

图算法在风控的演化

几个核心趋势

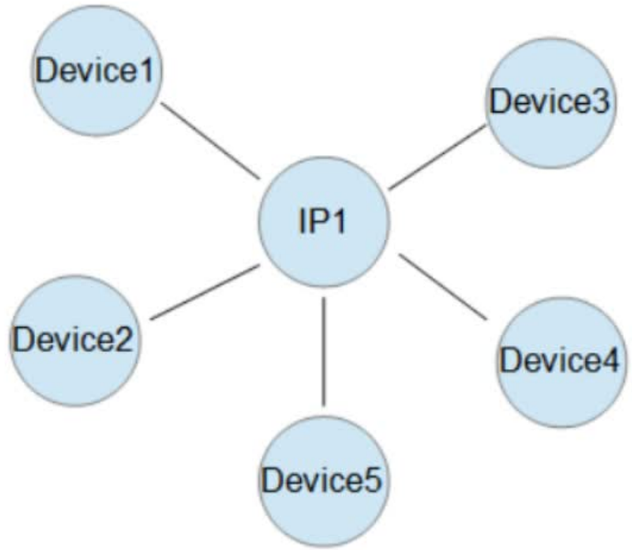
从规则到算法， **rule writer → algorithmic model**

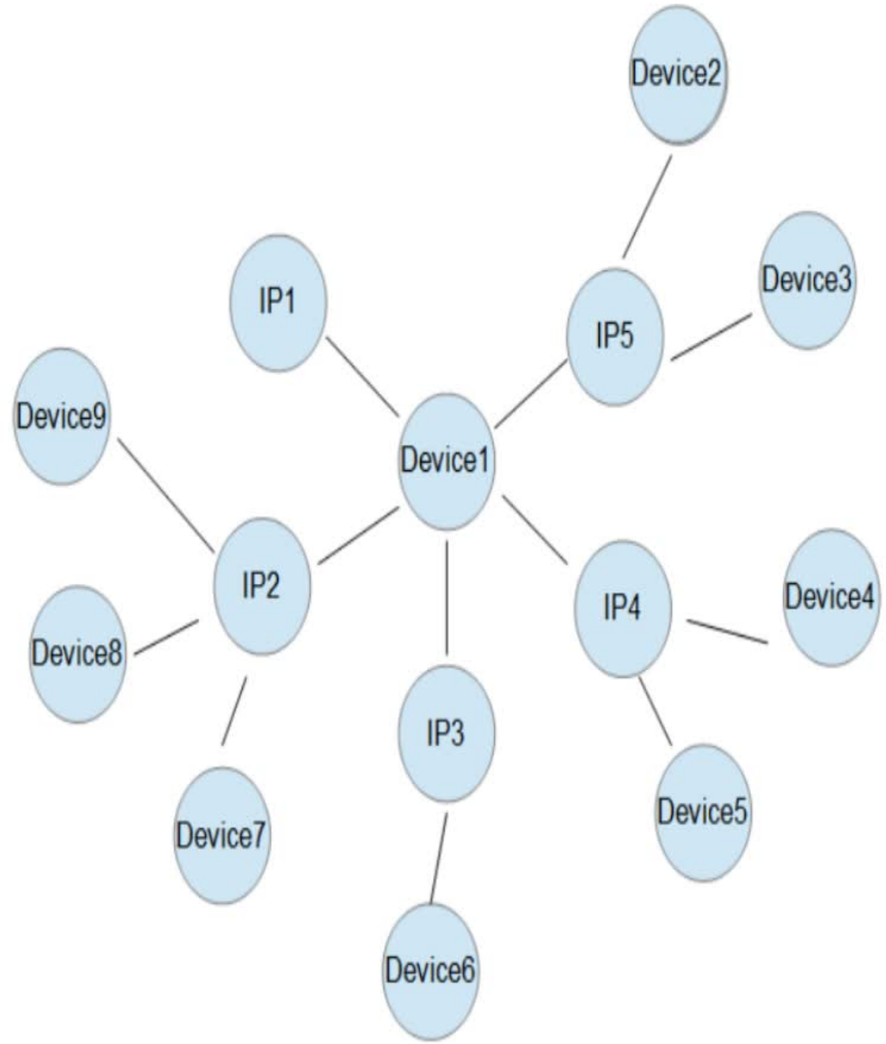
从经典一阶的**Velocity**变量到图神经网络 **one order velocity → neutral net aggregator**

从数学严格定义的网络结构到图神经网络 **strict definition → probabilistic inference**

经典一阶的Velocity变量

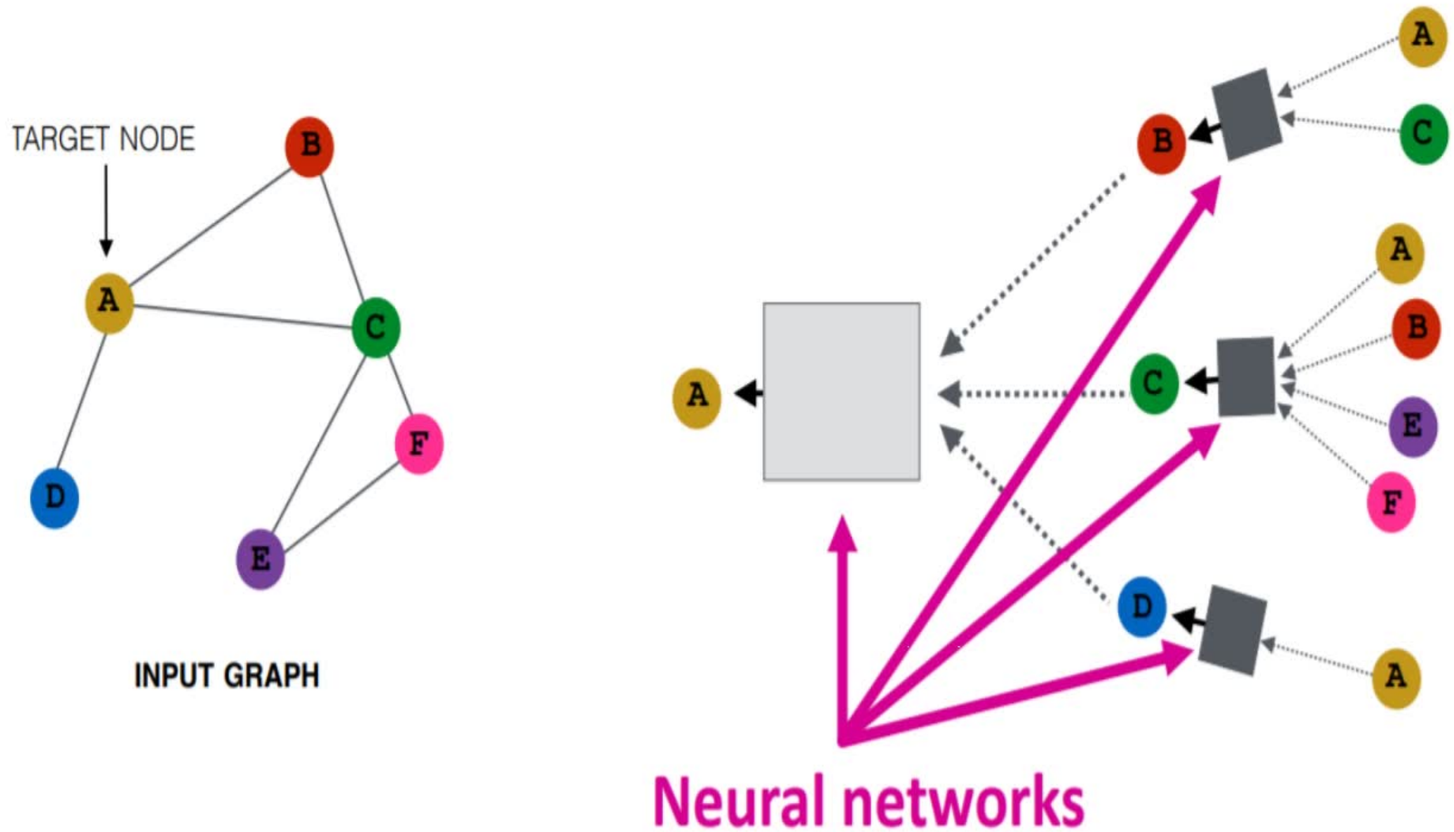
常见相关变量：
最近7天交易笔数
最近7天交易金额
最近30天交易笔数最近
30天交易金额最近
7天高危时段交易笔数
最近7天高危商品交易笔数
1小时内同ip交易用户数
最近7天同设备登录用户数





设计聚合函数。例如Min，Max, Mean 来聚合相关设备或IP的各种特征，成为风控因子。

神经网络的聚合



Aggregator算子的突破

代表性突破，Deepmind关于聚合函数的研究

Theorem 1 (Number of aggregators needed). *In order to discriminate between multisets of size n whose underlying set is $\mathbb{R}$, at least n aggregators are needed.*

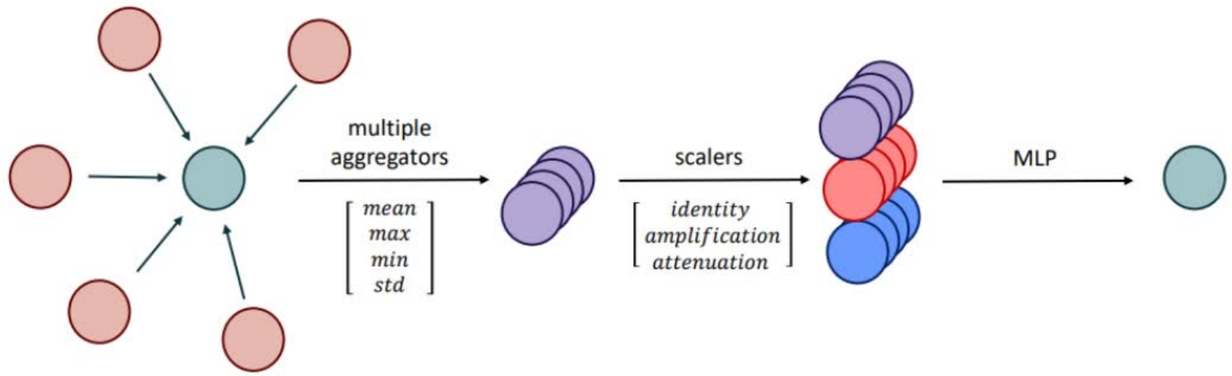
$$M_n(X) = \sqrt[n]{\mathbb{E}\left[(X-\mu)^n\right]}\ ,\quad n>1$$

- The proof is (in my opinion) **really cool**! (relies on **Borsuk-Ulam** theorem)

- PNA proposes empirically powerful **combination** of aggregators for general-purpose GNNs:

$$\bigoplus = \underbrace{\begin{bmatrix} I \\ S(D,\alpha=1) \\ S(D,\alpha=-1) \end{bmatrix}}_{\text{scalers}} \otimes \underbrace{\begin{bmatrix} \mu \\ \sigma \\ \max \\ \min \end{bmatrix}}_{\text{aggregators}}$$

$$S(d,\alpha) = \left(\frac{\log(d+1)}{\delta}\right)^{\alpha},\quad d>0,\quad -1\le\alpha\le 1$$

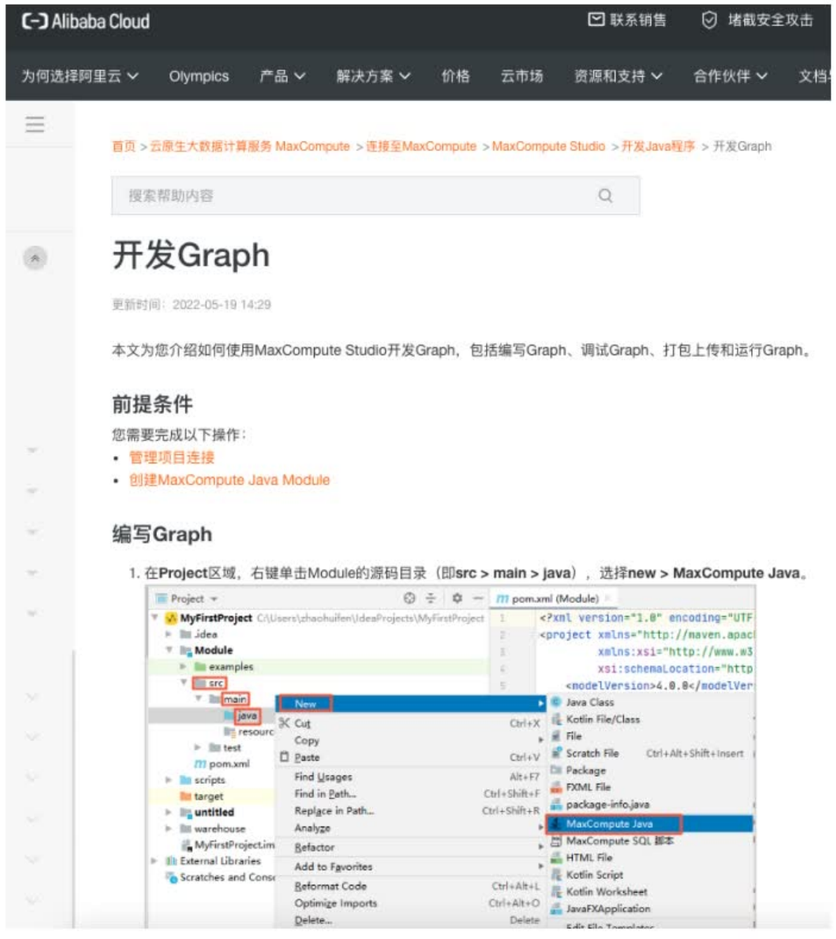


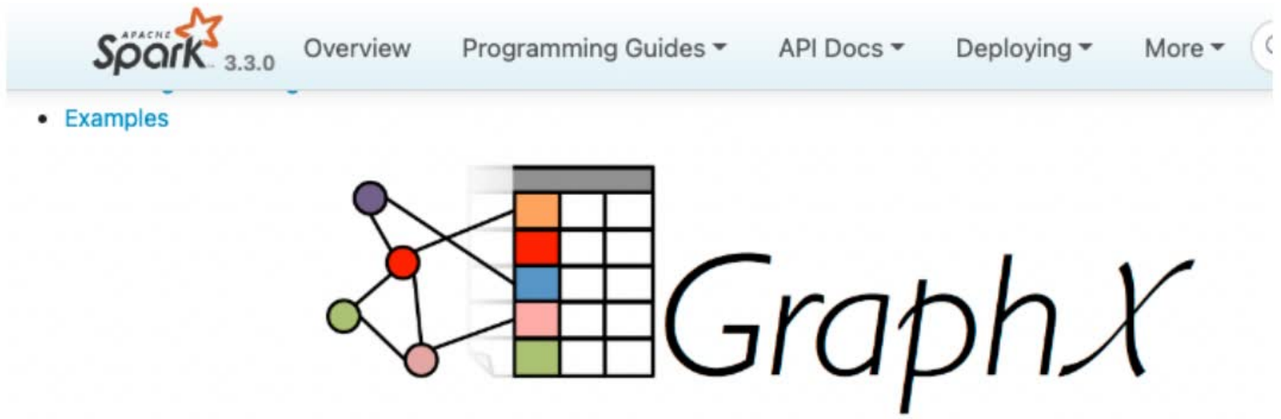
DataFun.

03

相应平台的演化

业界常用的完成离线图算法和图机器学习算法框架





About

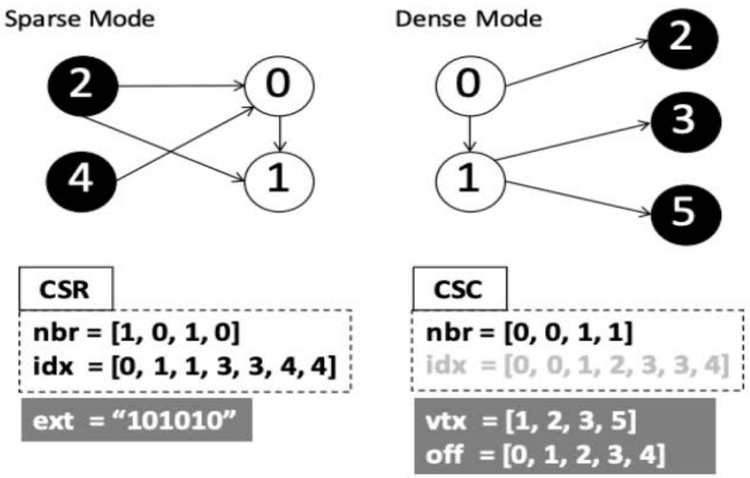
腾讯高性能分布式图计算框架Plato

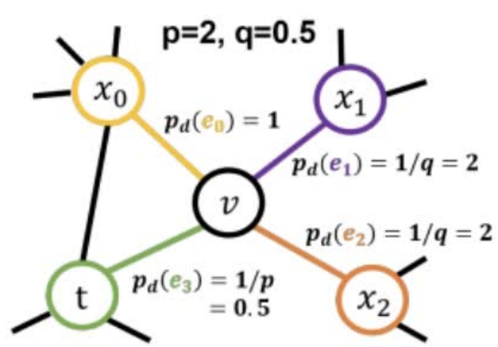
plato graph-computing

Push & Pull Sparse & Dense

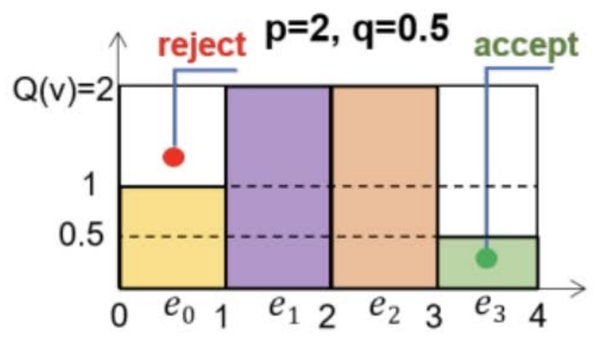
特殊边概率权重下的拒绝采样

Dual-Mode Edge Representation



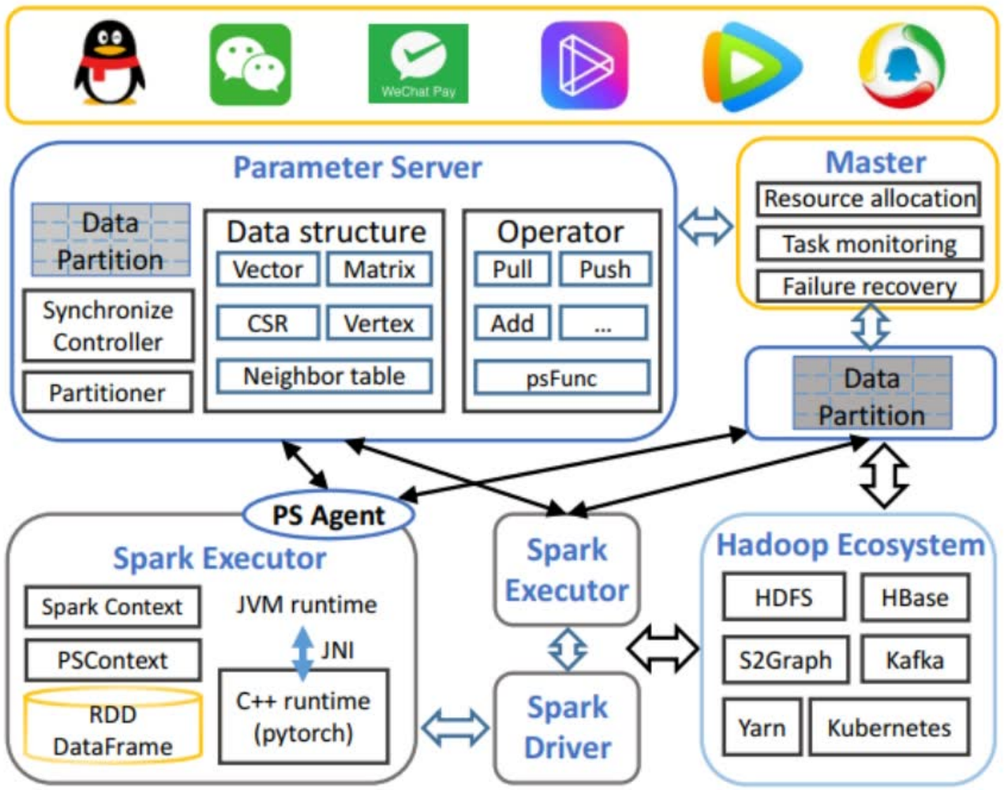


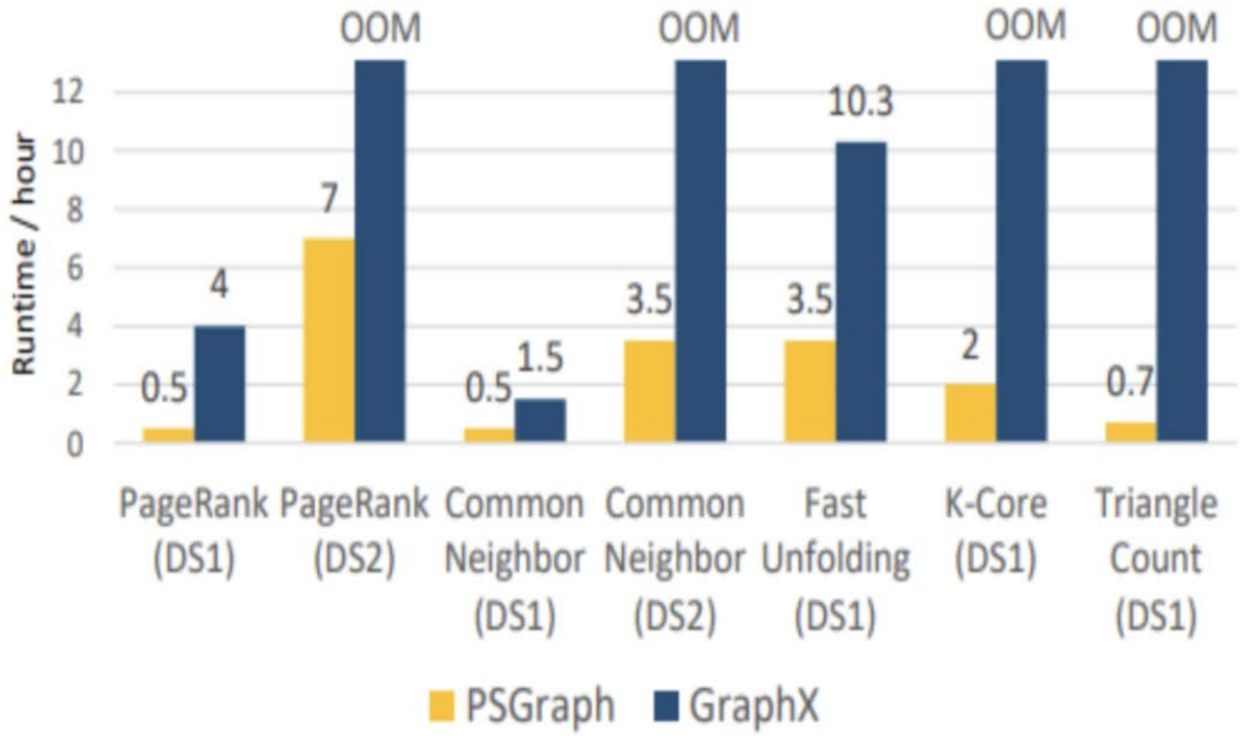
(a) Sample graph



(b) Rejection sampling at v

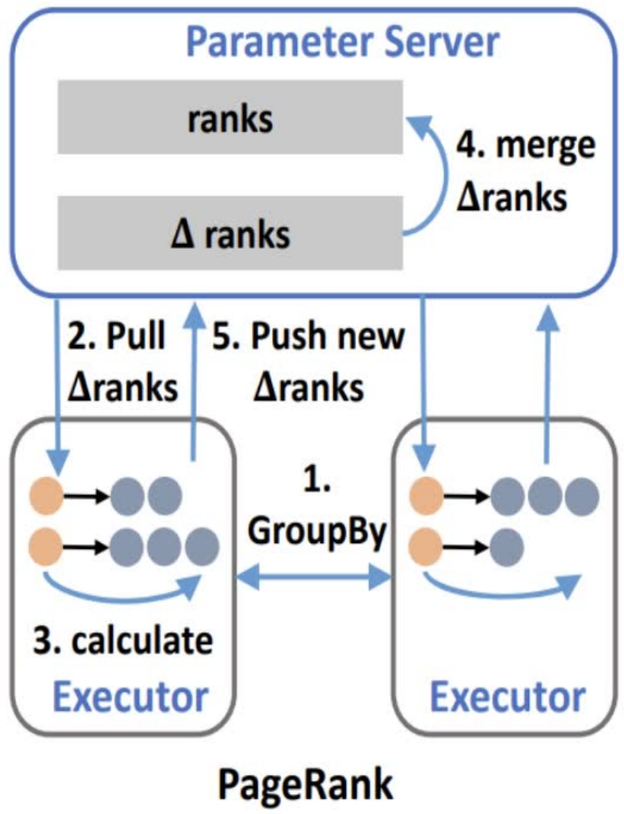
腾讯 Angel图计算-简介

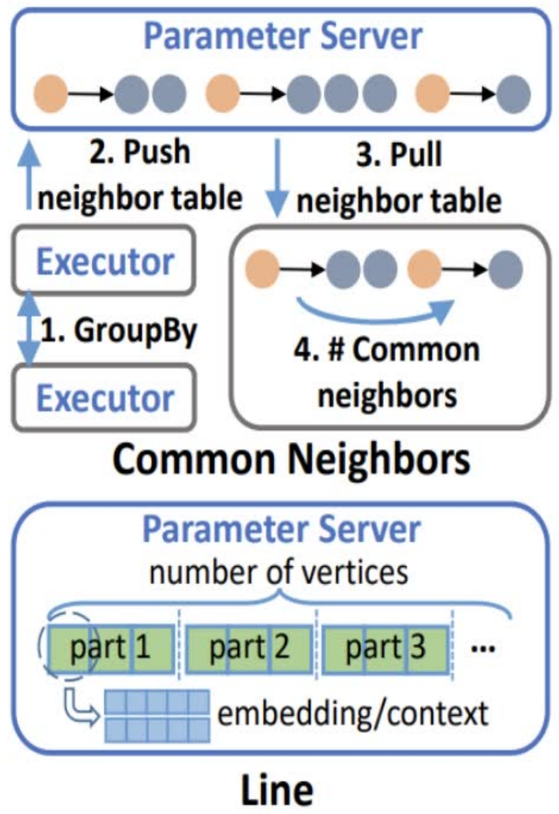




腾讯 Angel图计算-代码样例

```
class GraphRunner {

    def main(args: Array[String]): Unit = {
        val params = ArgsUtil.parse(args)
        SparkContext.getOrCreate()
        PSContext.getOrCreate()
        val algo = new GraphAlgo(params)
        val graph = GraphIO.load(params)
        val output = algo.transform(graph)
        GraphIO.save(output)
    }

}

class GraphAlgo(args: Array[String]) {

    def transform(dataset: Dataset[_]): DataFrame = {
        val edges = GraphOps.loadEdges(dataset)
        val neighborTable = GraphOps.toNeighborTable(edges)
        val model = PSContext.matrix(row, col, DataType)
        val delta = ...     // do calculation
        model.update(delta)
        SparkContext.createDataFrame(model)
    }

}
```





腾讯 Angel图计算-GraphSage实现

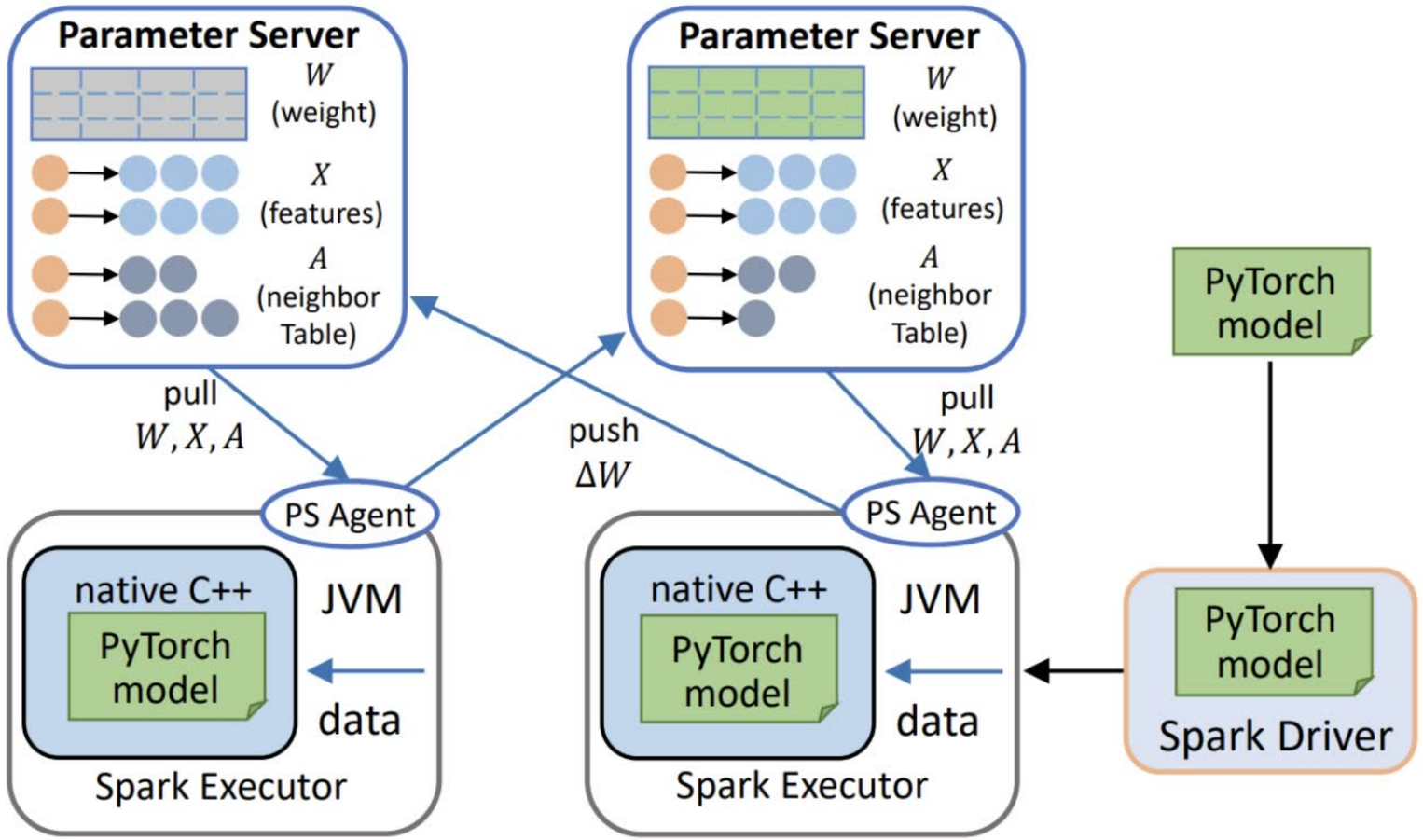
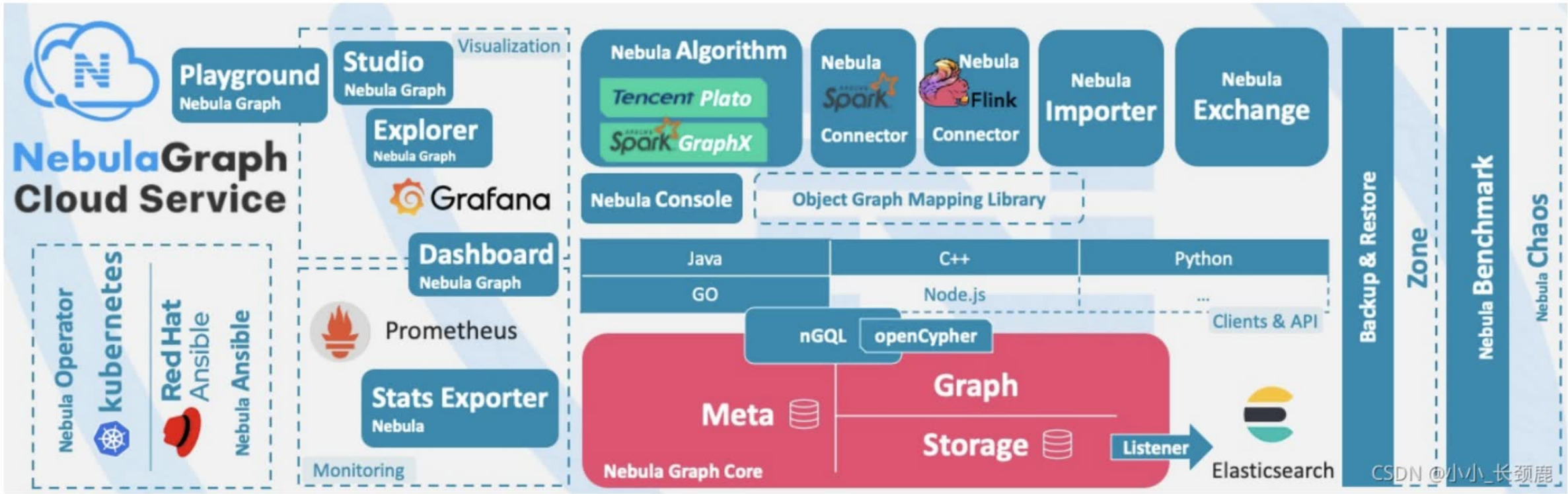


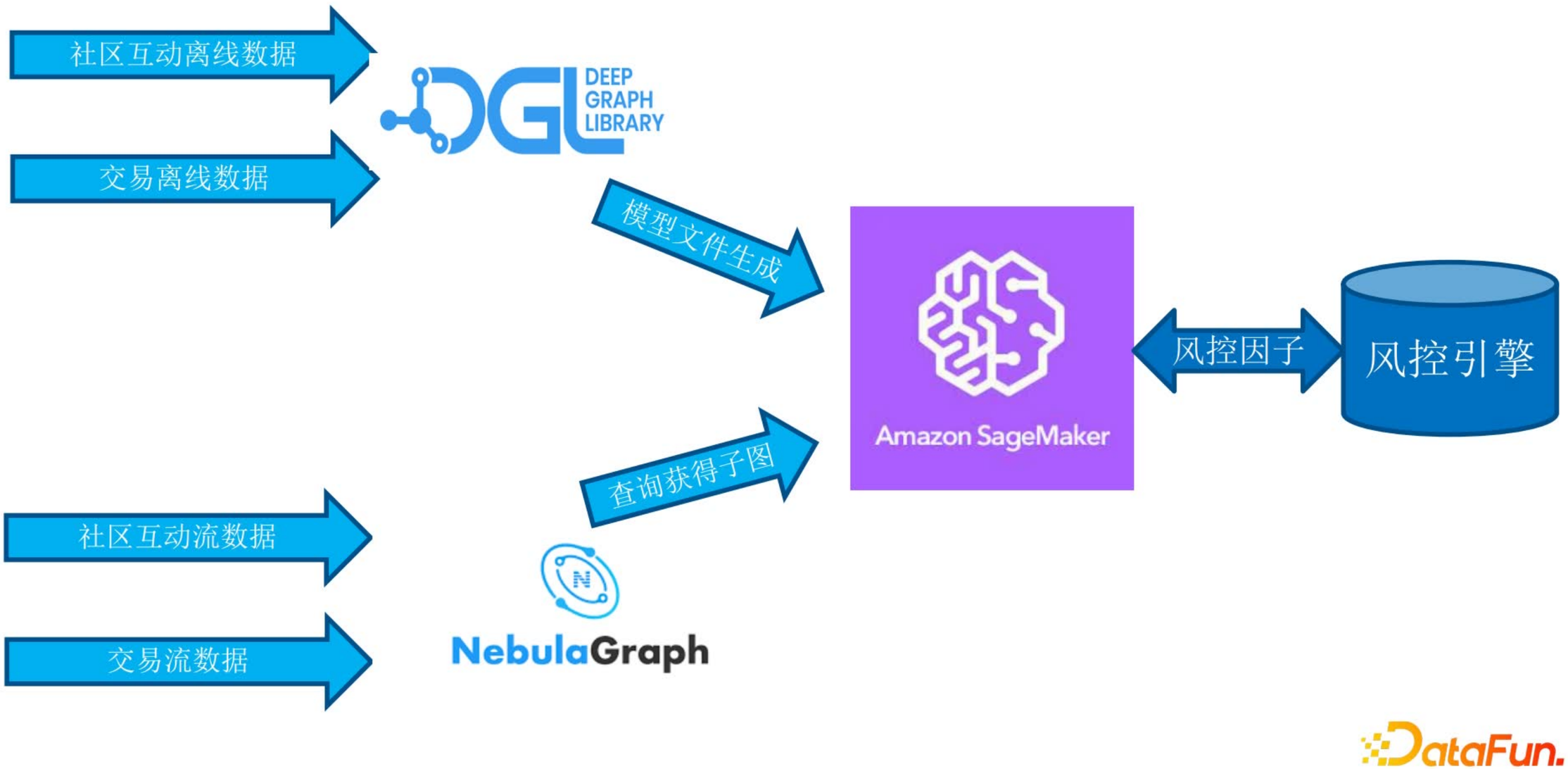
Fig. 5. Implementation of GraphSage.

开源图数据库生态完善

NebulaGraph



DGL+SageMaker+Nebula, 快速搭建一个可以线上实时模型查询推理系统

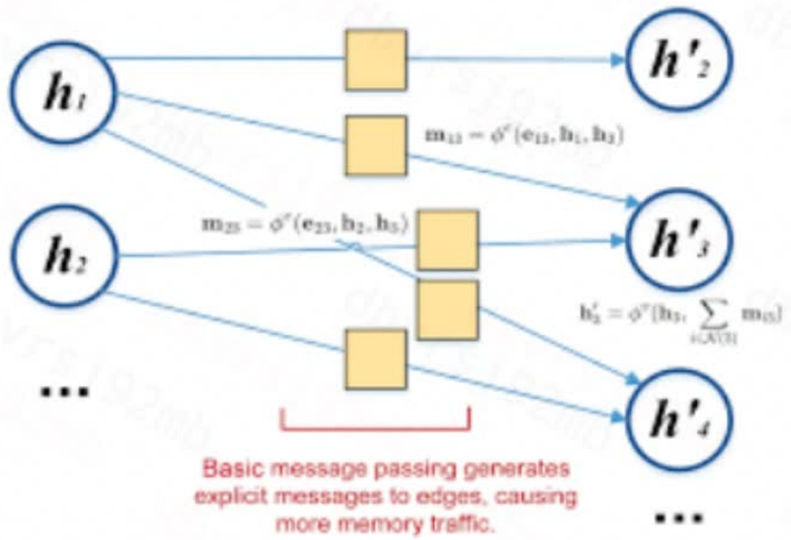


DGL – message passing 角度的编程和实现优化，减少通讯和存储开销，利用GPU特性

内部支持的Message Passing

```
def gcn_msg(edge):
    msg = edge.src['h'] * edge.src['norm']
    return {'m': msg}
```

```
def gcn_reduce(node):
    accum = torch.sum(node.mailbox['m'], 1) *
node.data['norm']
    return {'h': accum}
```



$$\mathbf{M} = \text{send}(E, \phi^e, \mathbf{H}_v) \qquad \mathbf{H}'_v = \text{recv}(V, \Sigma, \mathbf{M})$$

Basic message passing

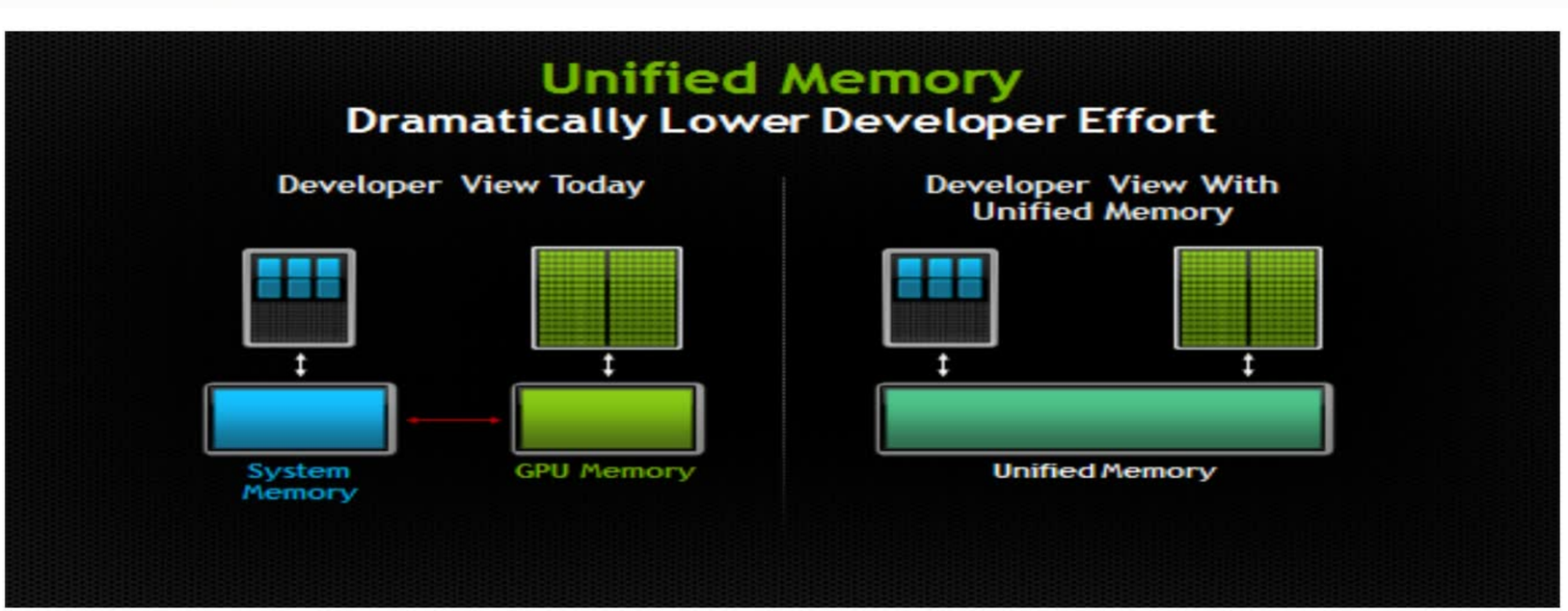
$$\mathbf{H}'_v = \text{send\_and\_recv}(E, \phi^e, \Sigma, \mathbf{H}_v)$$

DGL **fuses** the computation as `send_and_recv` to avoid explicit message storage, thus is **faster** and more **scalable** for large graphs.

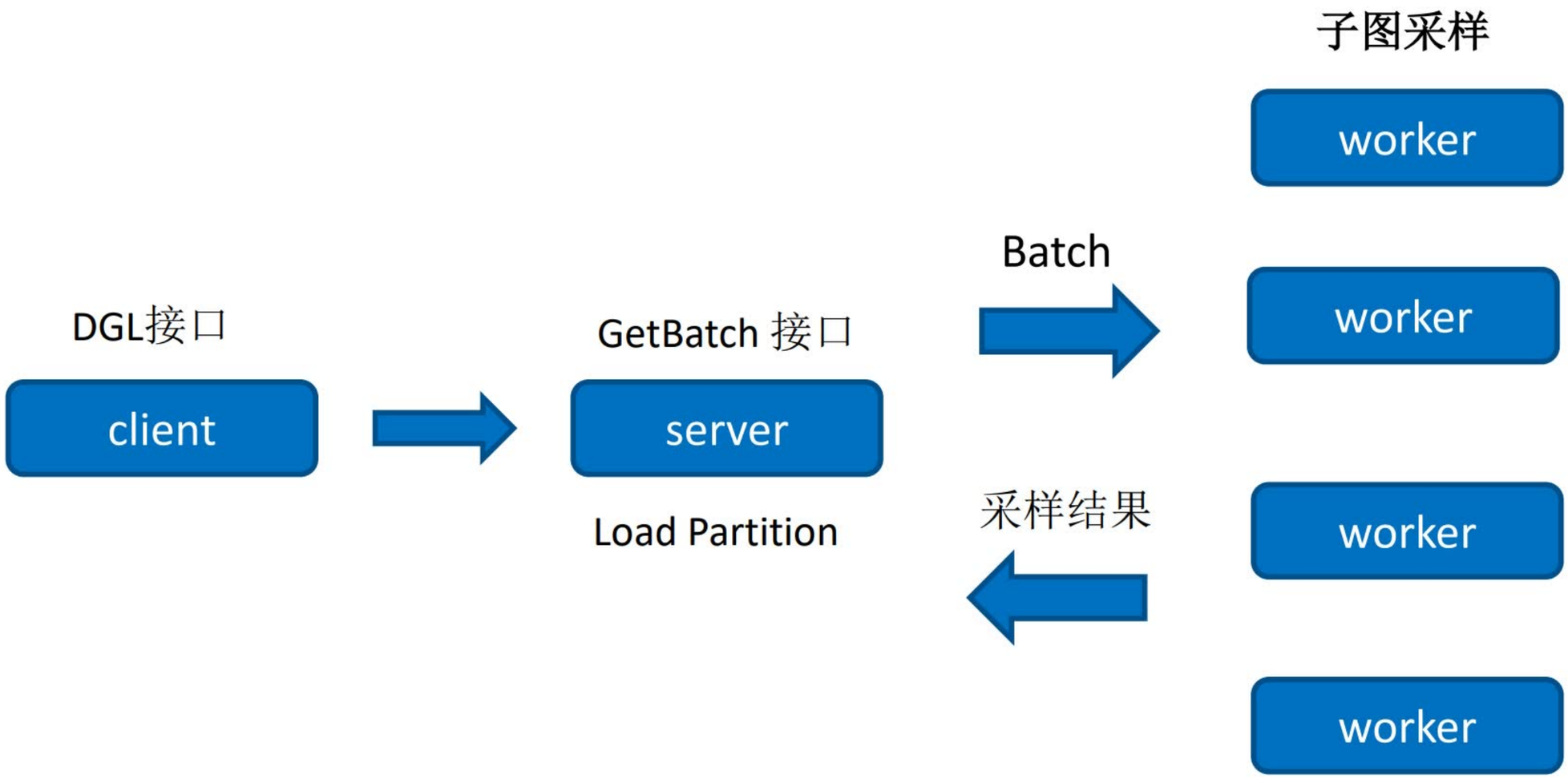
DGL-GPU优化，更多的操作交给GPU来完成

One can use GPU-based neighborhood sampling with DGL data loaders via:

- Put the graph onto GPU.
- Put the `train_nid` onto GPU.
- Set `device` argument to a GPU device.
- Set `num_workers` argument to 0, because CUDA does not allow multiple processes accessing the same context.



DGL-Adapter 突破内存限制，少量的通信成本牺牲换取更大的图数据规模训练能力。



DataFun.

04

展望未来

图算法和图神经网络算法的融合

图神经网络算法学习能力的攻克

图神经网络算法鲁棒性

图神经网络算法可解释性

平台易用性和整合性

应用算法和系统算法上下融会贯通和统筹

非常感谢您的观看

DataFun.