DTCC 2010-2022

第十三届中国数据库技术大会

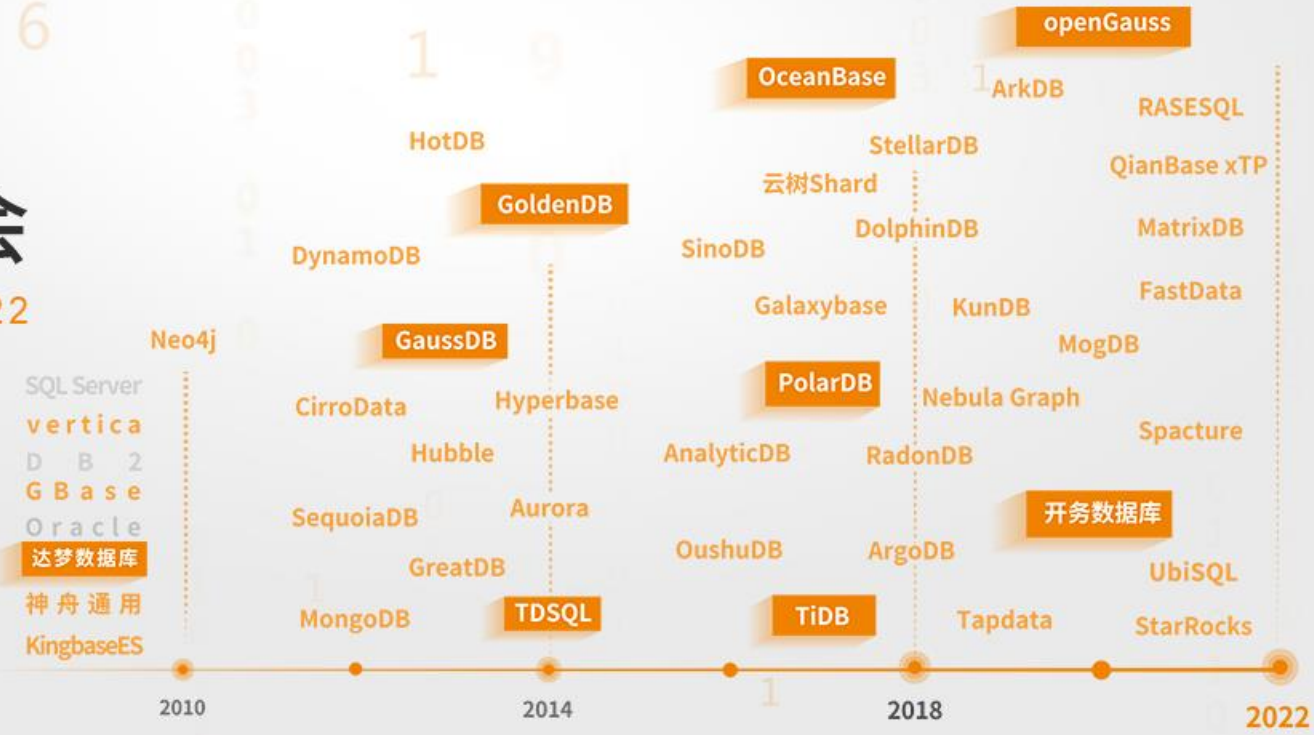
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



图数据库构建长城汽车数据一元化OneID数据底座

长城汽车-产业数智化中心（IDC）

陈晓 数据中台大数据工程师

目录

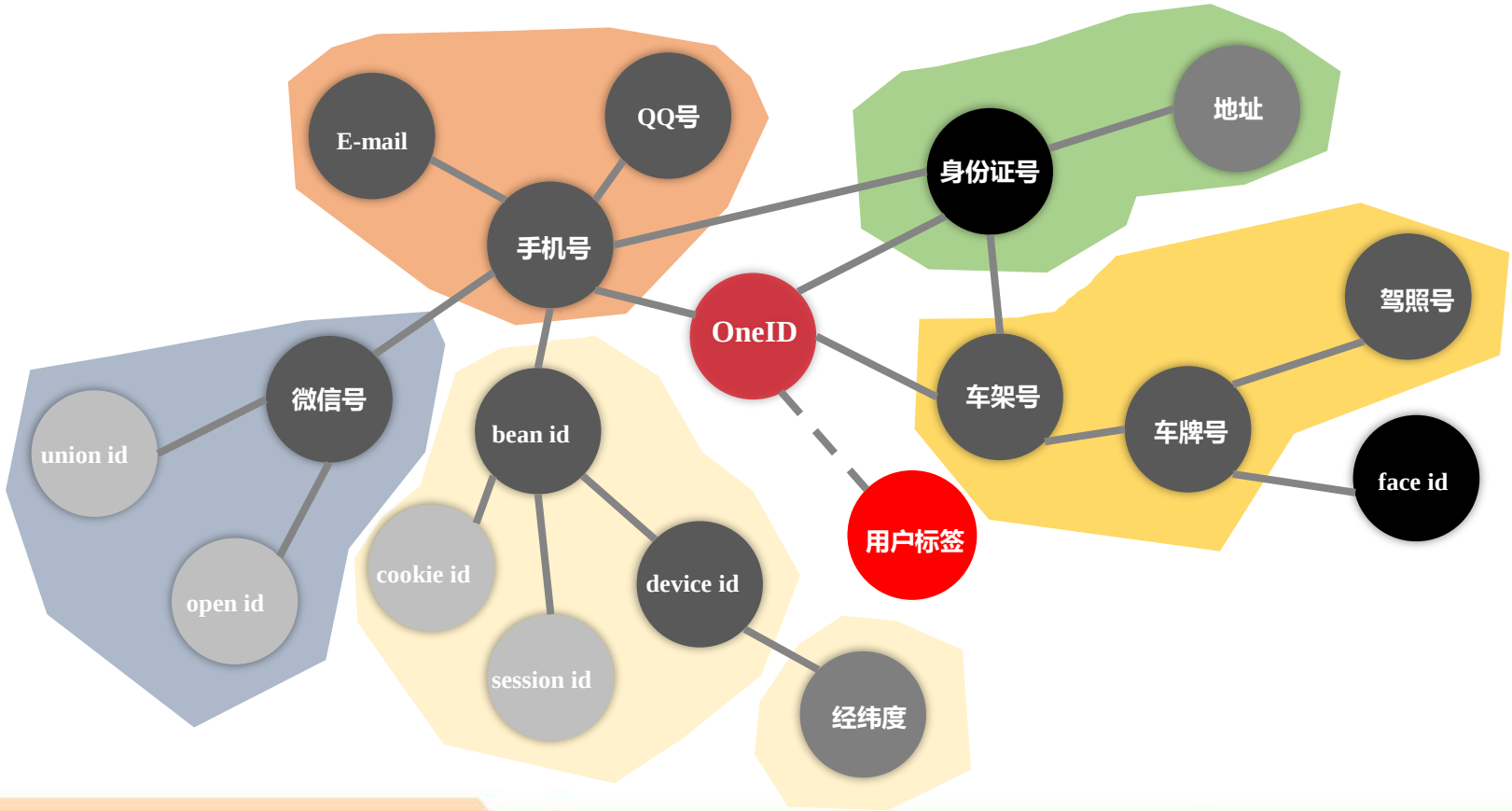
CONTENTS

Part 1

图数据库在OneID项目中的应用

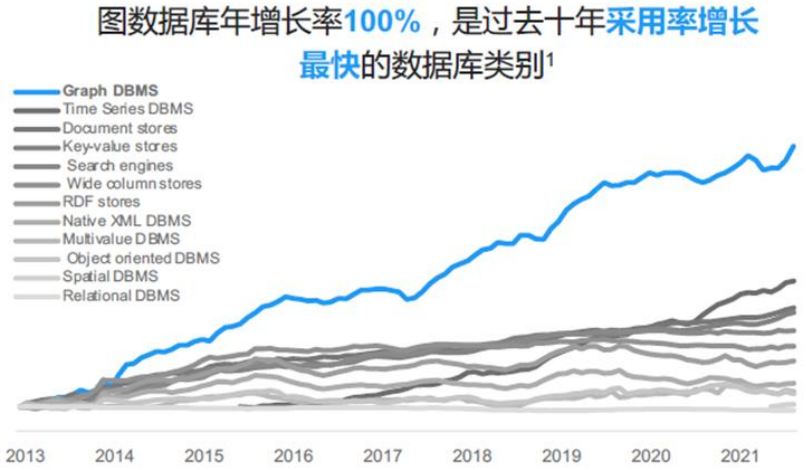
OneID实现数据一元化

- 数据孤岛问题；例如：PC端、APP端、微信小程序、相关系统单元等.
- 统一用户识别；用ID-mapping技术将多ID归一管理，解决作为用户的唯一身份识别.



图数据库

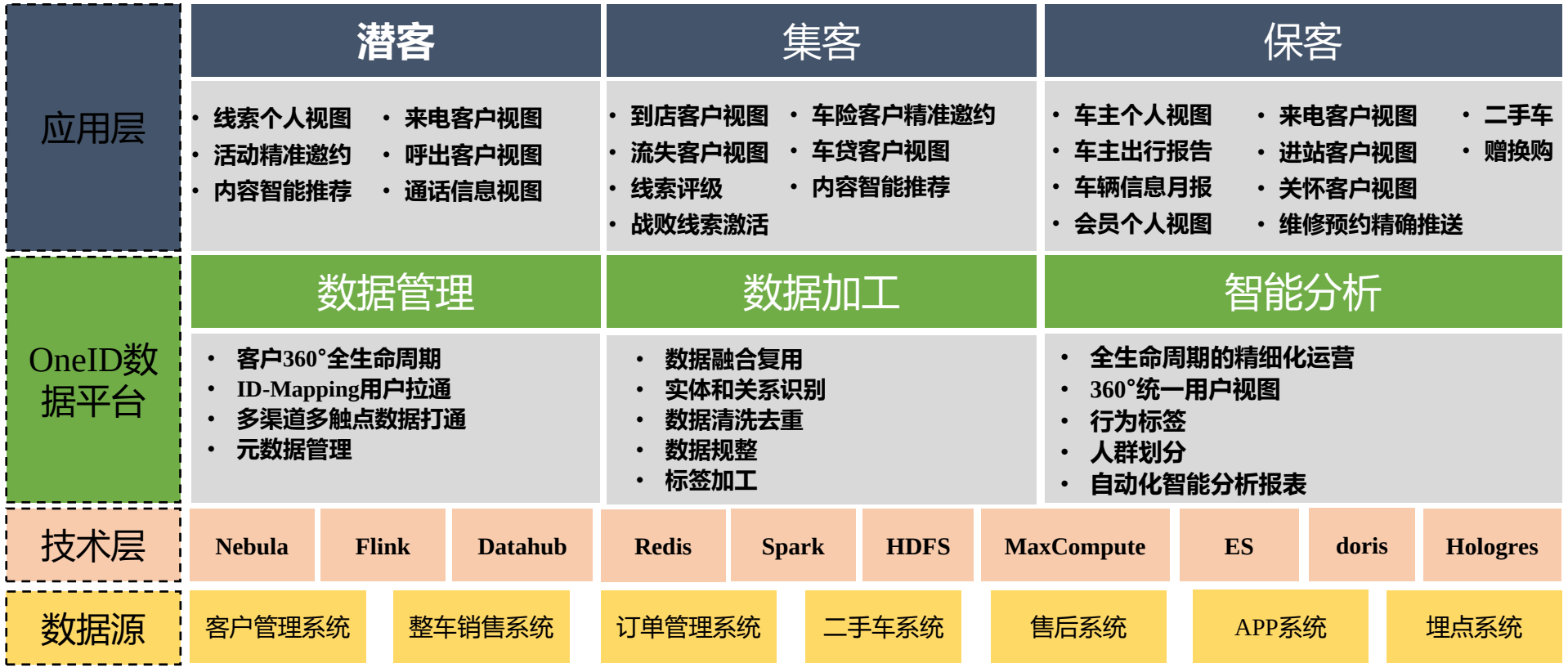
- 图数据库本身就是基于事物关联关系的模型表达工具，与OneID的拉通数据的理念相契合；
- 图数据库在解决大规模数据实体间复杂关系的查询问题上，具备天然优势，查询效率提升显著；
- 图数据库过去10年增长趋势最快，验证了图数据库成为一种趋势；
- NebulaGraph 在处理超大数据集（千亿节点万亿条边）的查询上保持毫秒级查询延时.



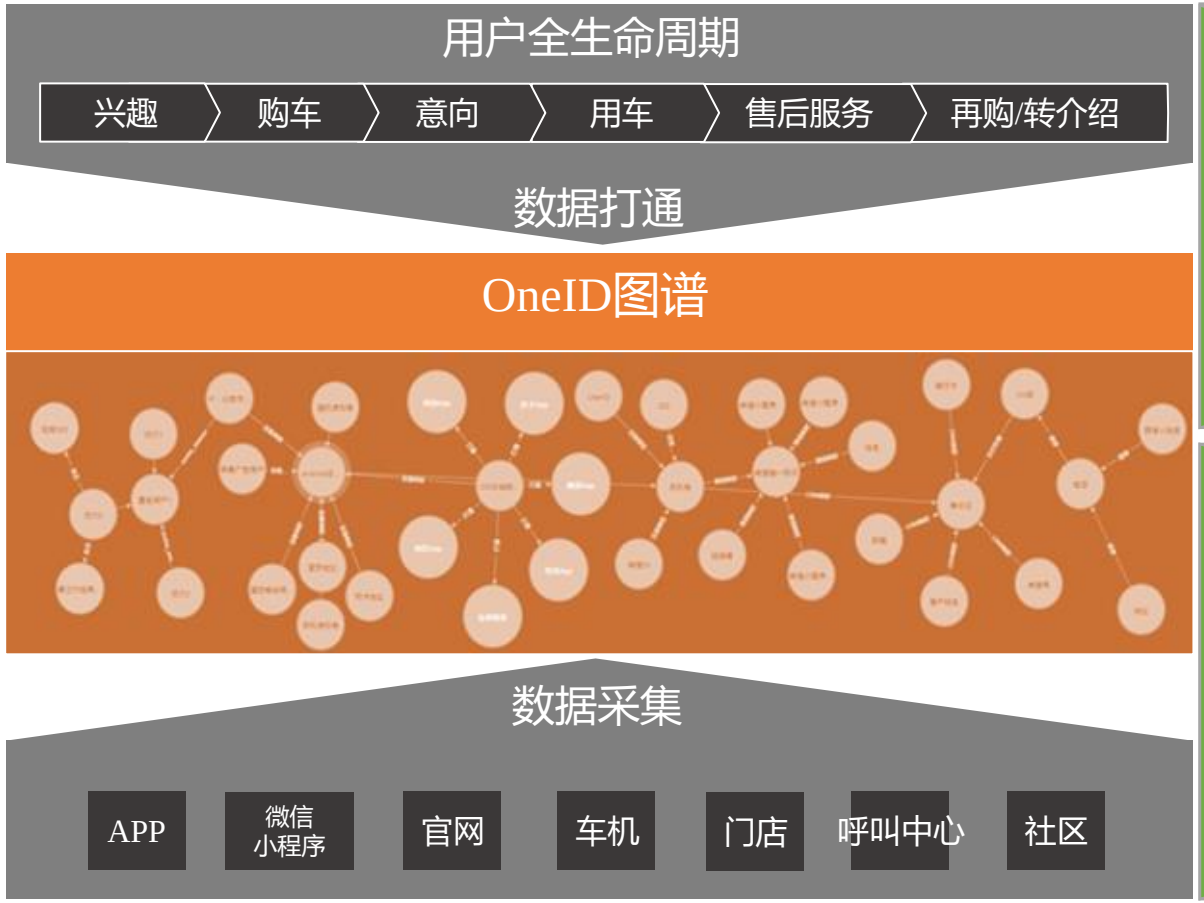
75%世界百强企业采用图数据库，起到标杆作用，未来快速辐射腰部企业[2]

20/25	7/10	7/10	4/5
金融企业	零售企业	互联网企业	电信企业

OneID平台架构


应用层
潜客
线索个人视图
来电客户视图
活动精准邀约
呼出客户视图
内容智能推荐
通话信息视图
集客
到店客户视图
车险客户精准邀约
流失客户视图
车贷客户视图
线索评级
内容智能推荐
战败线索激活
保客
车主个人视图
来电客户视图
二手车
车主出行报告
进站客户视图
赠换购
车辆信息月报
关怀客户视图
会员个人视图
维修预约精确推送
OneID数据平台
数据管理
客户360°全生命周期
ID-Mapping用户拉通
多渠道多触点数据打通
元数据管理
数据加工
数据融合复用
实体和关系识别
数据清洗去重
数据规整
标签加工
智能分析
全生命周期的精细化运营
360°统一用户视图
行为标签
人群划分
自动化智能分析报表
技术层
Nebula
Flink
Datahub
Redis
Spark
HDFS
MaxCompute
ES
doris
Hologres
数据源
客户管理系统
整车销售系统
订单管理系统
二手车系统
售后系统
APP系统
埋点系统

OneID标签图谱赋能用户增长和个性化运营



推荐引擎
- 有用
- 有利
- 有趣
- 有朋友

→ 用户运营（个性化运营）

自动化运营
- 拉新
- 激活
- 转化
- 裂变

→ 增长运营（精细化运营）

各大推荐场景：

① 首页个性化推荐
② 频道首页信息流推荐
③ 相关推荐
④ 搜索推荐
⑤ 热门推荐
⑥ 猜你喜欢
⑦ 社交推荐
⑧ 商品图谱推荐

OneID赋能业务场景

OneID为客服人员、销售顾问、运营人员，在流量、线索、场景等价值流转化上提供落地服务.

价值流 → **潜客**(流量转化) → **集客**(线索转化) → **保客**(场景转化)

客服人员

呼叫中心 – 来电弹屏页
车辆查询页

销售顾问

登录明细页

运营人员

用户增长模型页

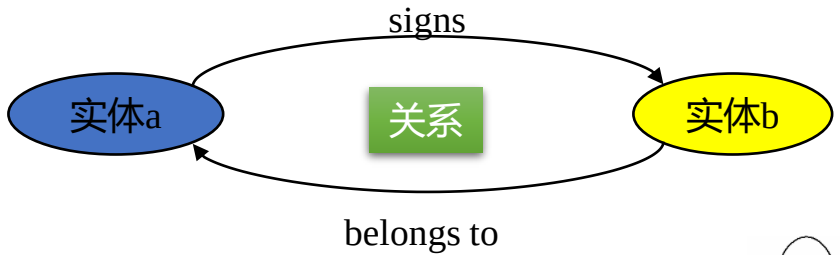
OneID实现了用户360°视图，应该将**OneID**的建设经验和能力复用到车辆360°、零部件360°、供应商360°上，为更多领域业务提供数据服务.

Part 2

OneID项目中的图挖掘技术

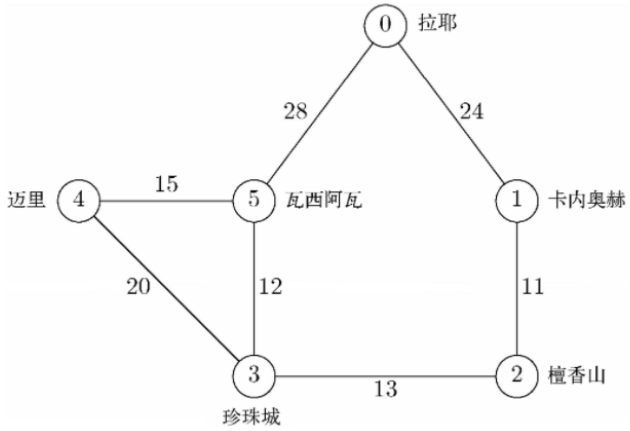
什么是图

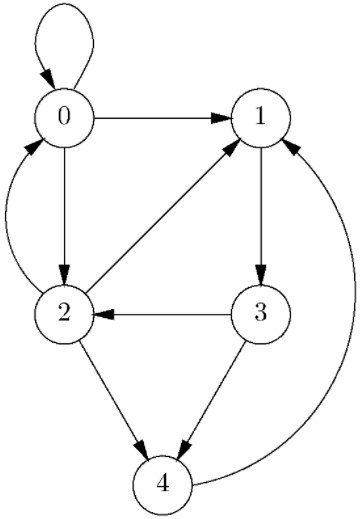
图就是二元关系，它利用一系列由线（称为边）或箭头（称为弧）连接的点（称为节点）提供了强大的视觉效果.
图的本质是由二元关系组构成，实体-关系-实体模型：



图有多种形式：

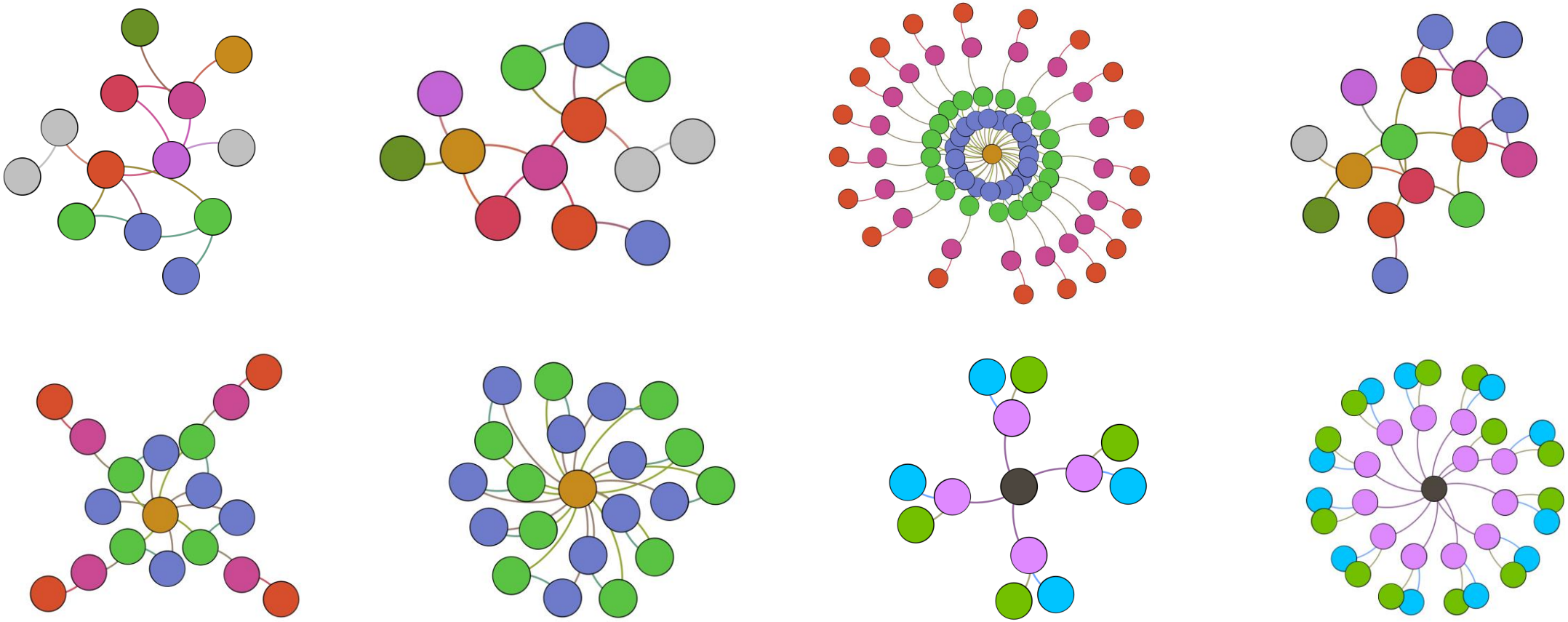
- 无向图、有向图
- 加权、未加权
- 同构 、异构
- 单边、多重边
- 静态图、动态图





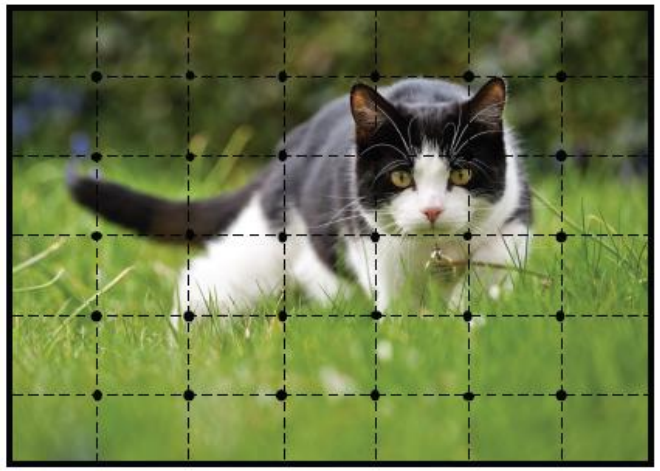
OneID图库中不同形态的图

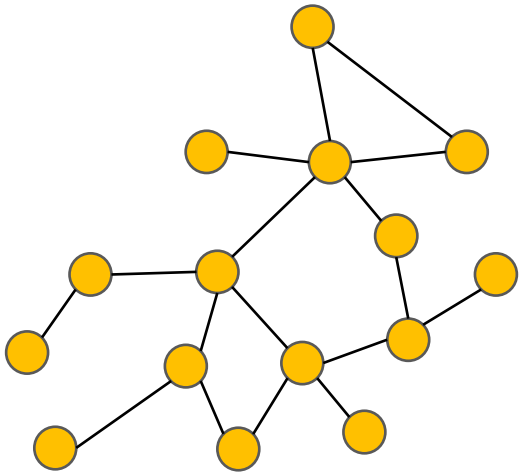
形态各异的图网络：



如何进行图建模

图像和拓扑图对比：





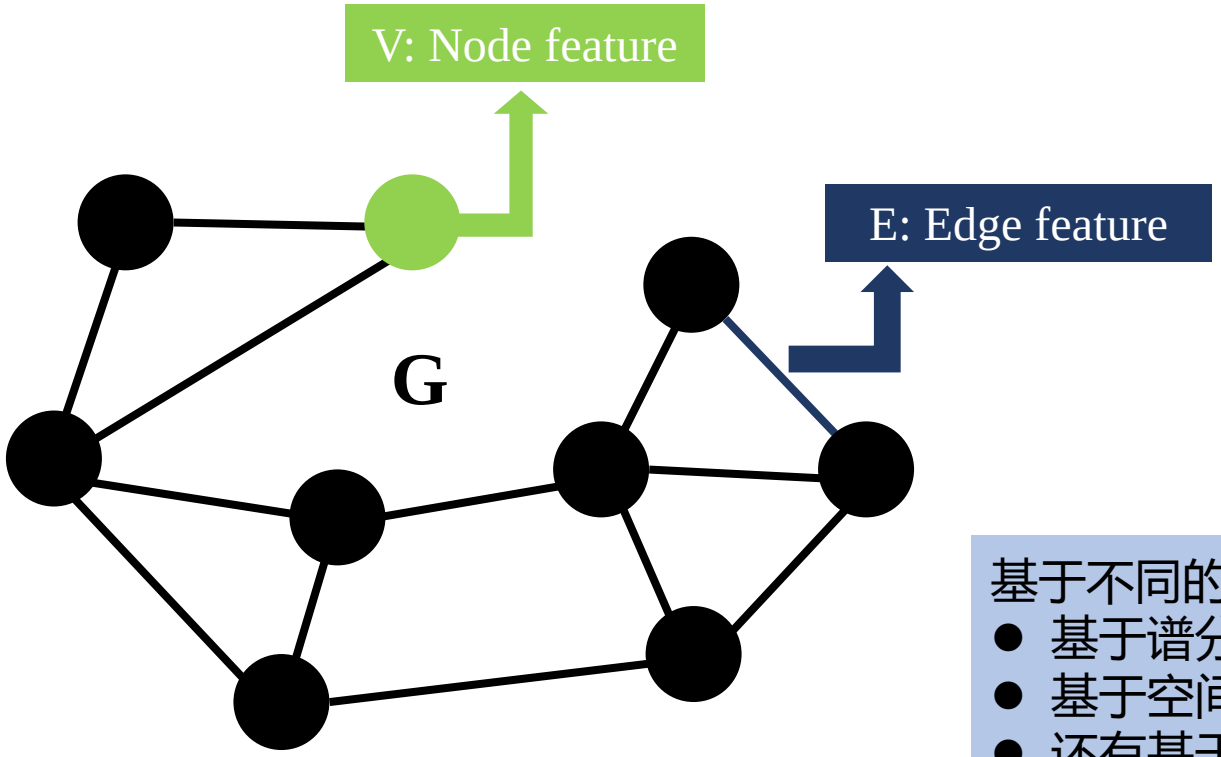
CNN在非欧氏空间这里不适用.

因此，不是所有的事情可以表示为序列或网格。我们将如何应用神经网络?

左:欧氏空间中的图像　　右:非欧氏空间中的图

如何进行图建模

图卷积的应用：



图表示：G = （V，E）

邻域矩阵Aij：

$$A_{ij} = \begin{cases} 1 & If\ \{v_i, v_j\} \in E \quad and \quad i \neq j \\ 0 & otherwise \end{cases}$$

度矩阵D: $D_{ii} = d(v_i)$

图卷积神经网络卷积层的基本传播规则：

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

基于不同的传播规则和采样技术：

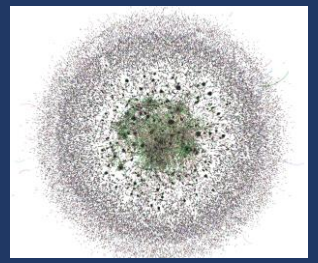
- 基于谱分解的方法：GCN、ChebNet、Spectral Network
- 基于空间结构的方法：GraphSAGE、GAT
- 还有基于递归运算和跳链接的其他传播模块

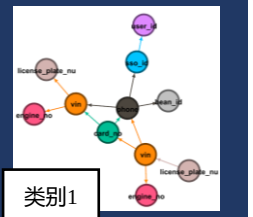
图算法在OneID项目中的应用

应用层

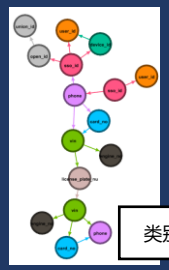
OneID项目应用案例

图融合
图切割
图分类

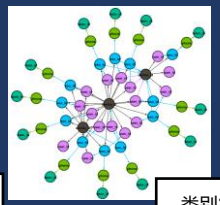




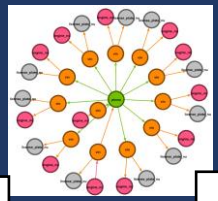
类别1



类别2



类别3



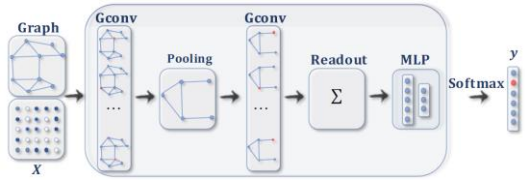
类别4

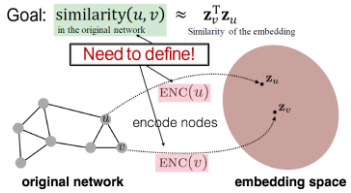
自定义类别 …

计算层

图算法

GNN	Graph Embedding	传统图挖掘算法
GCN SAGPool DiffPool …	DeepWalk Node2Vec Line …	Centrality Algorithms Betweenness Centrality Pathfinding algorithms …





计算框架

Pytorch | Tensorflow | PYG | DGL | Networkx | Sklearn

基础层

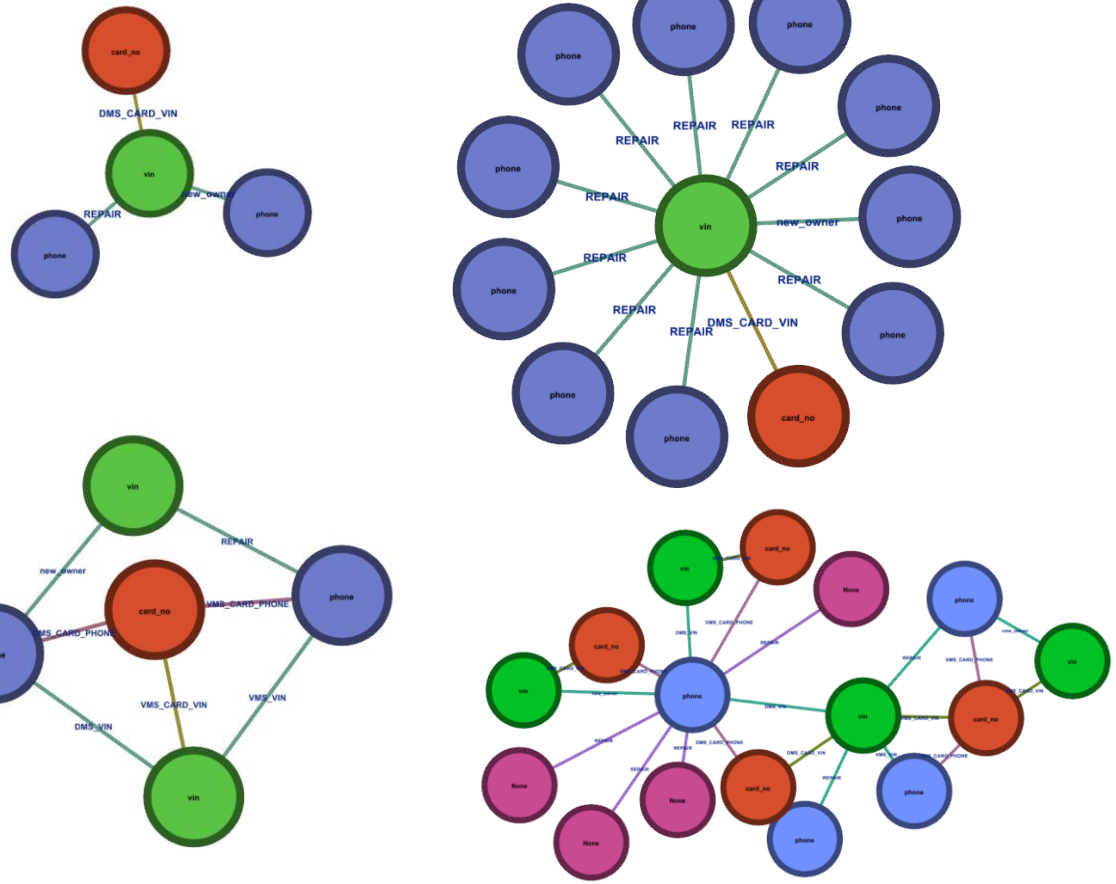
基础硬件

CPU | GPU（cuda）

图算法应用1—车主信息融合

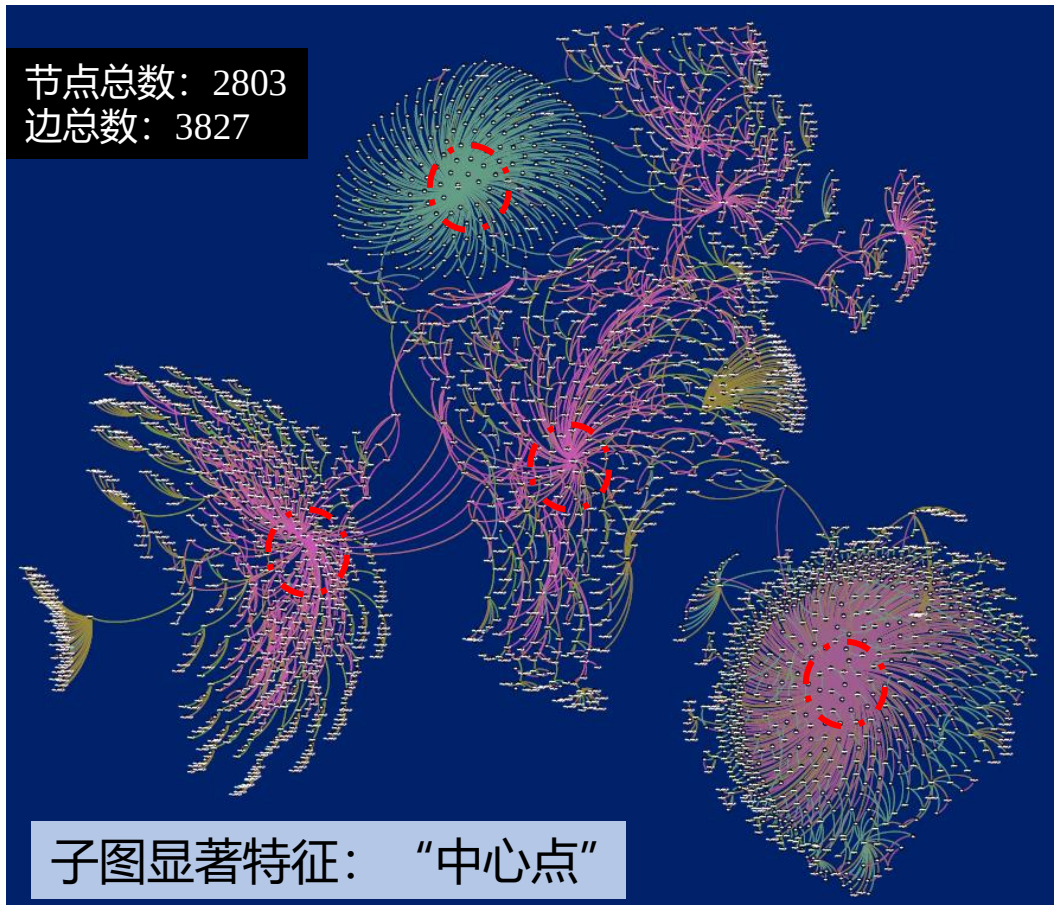
例如：利用图技术，对车和人是一对多的关系进行融合、优化.

补全：利用融合信息，对缺失信息推断;
去重：重复、真伪信息过滤;
推断：先验知识辅助推断.



图算法应用2—图切割提纯

信息融合过程中存在的“信息冗余”问题：



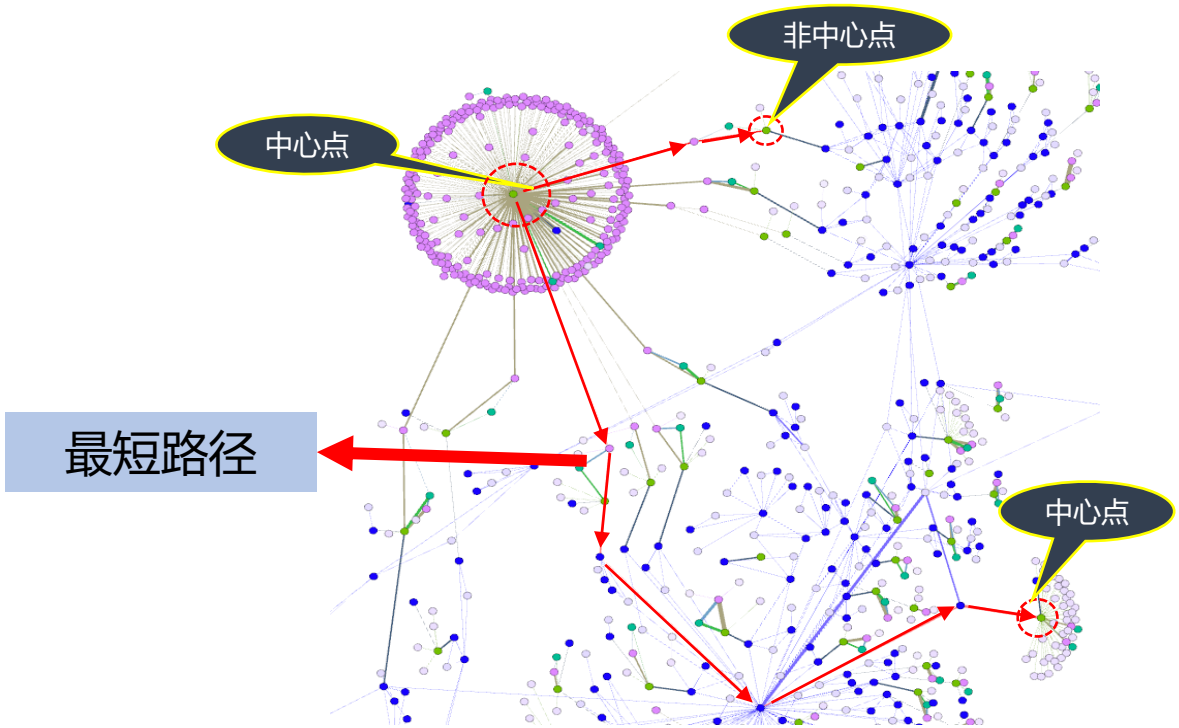
“信息冗余”问题：

- 销售、企业 通过手机、身份证将多人信息糅在同一张子图中；
- “羊毛党”刷单，将多人信息绑定到同一张子图中；
- “脏数据”流入绑定多人信息.

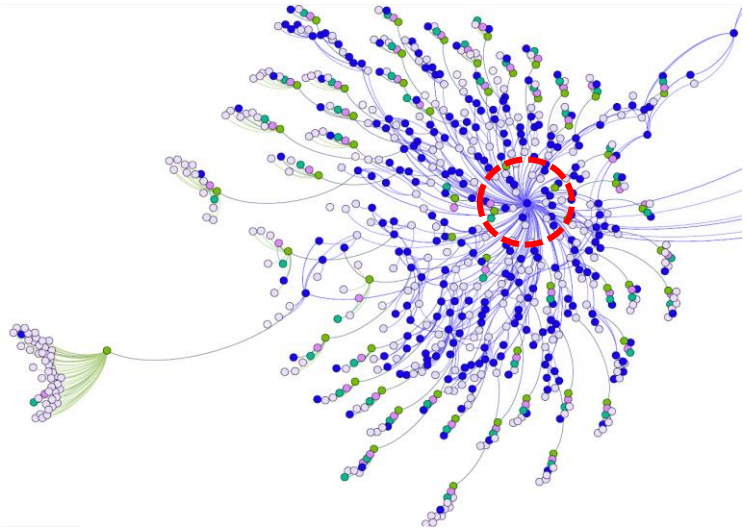
思路：在同一子图中,存在出入度比较高的一类节点，结合业务将此类节点作为“中心节点”，探索它与临近节点的关系、与他人信息之间的最短路径，进而将子图切割，实现无关信息的隔离.

图算法应用2—图切割提纯

利用最短路径算法、中心点算法等实现图切割：



- 识别出中心点节点和非中心点节点
- 使用最短路径算法完成切割
- 识别出问题设备“羊毛党”

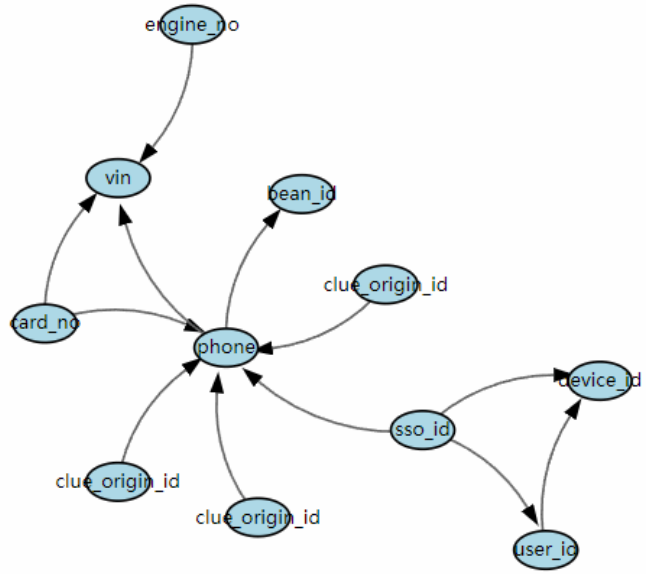


图算法应用2—图切割提纯

切割后效果展示：

案例1：在保证图信息的完整、独立的情况下，提取出某公司的相关信息

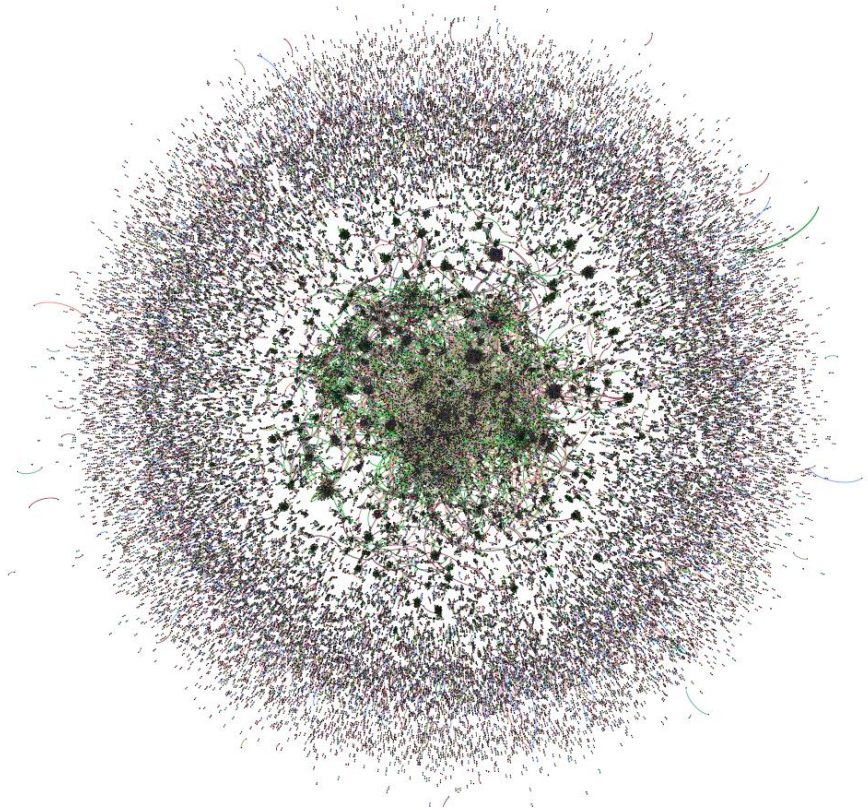
案例2：某个人的信息提纯

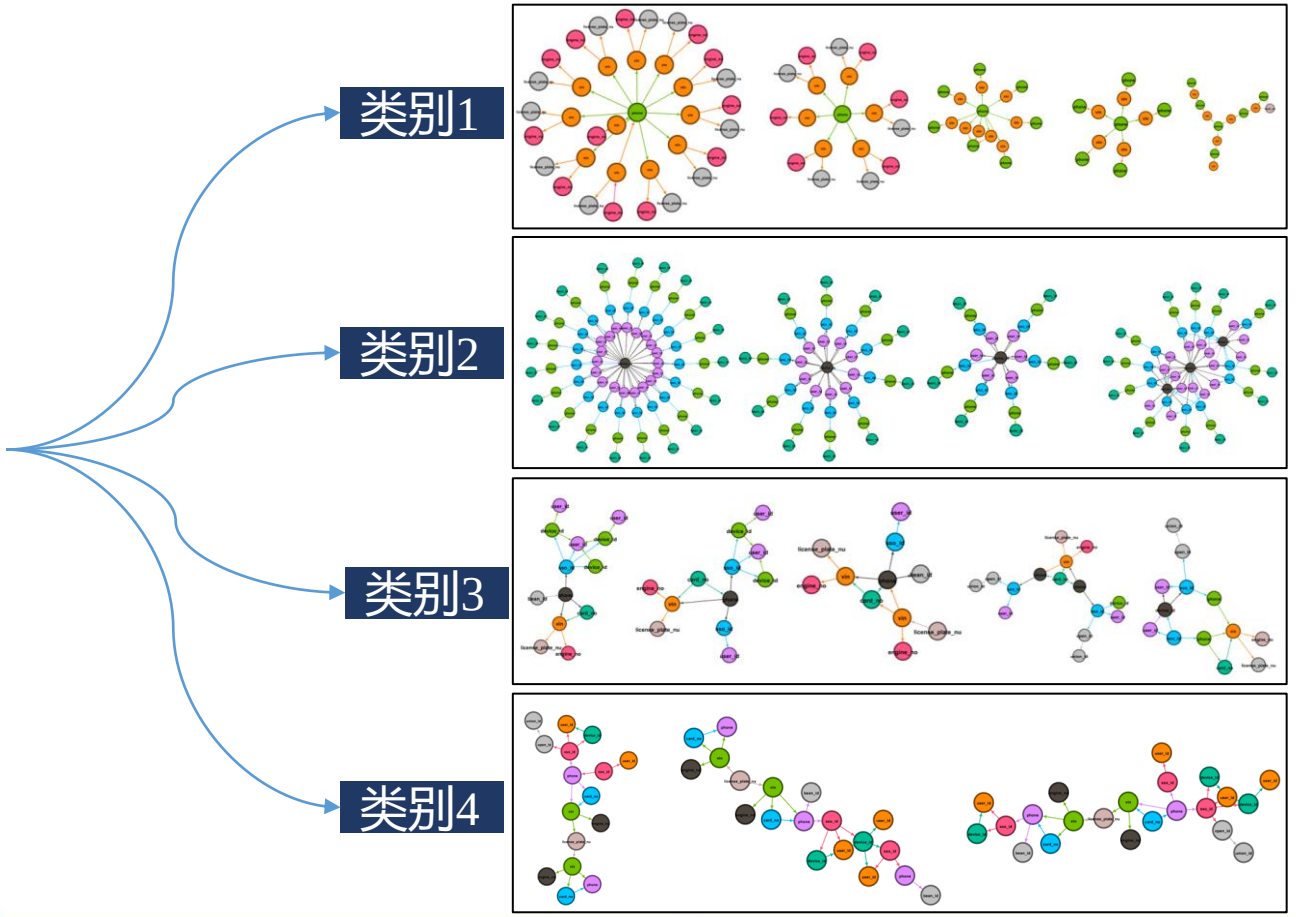


- 提升图的数据质量
- 支撑推荐系统、用户画像、图打分等应用服务体系建设

图算法应用3—图分类

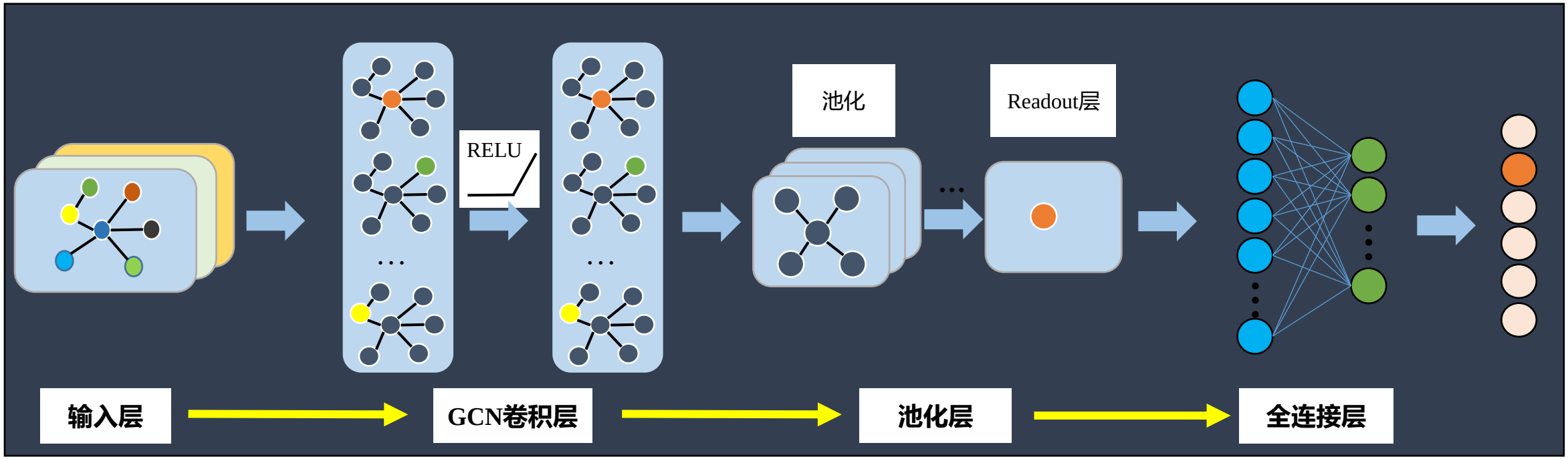
子图分类：物以类聚人以群分
例如：家庭用户、销售人员、公司大客户等.





图算法应用3—图分类

图神经网络子图分类模型架构：



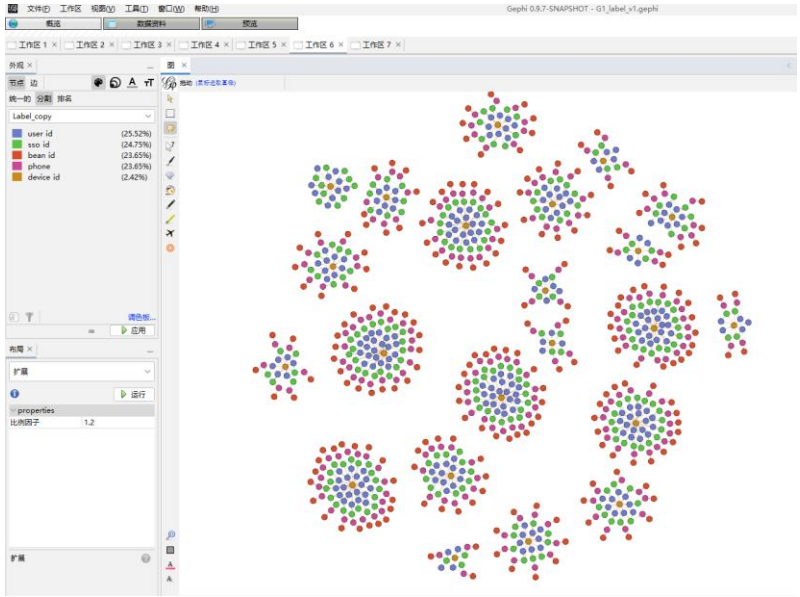
- 输入层：度矩阵D、邻域矩阵A、节点特征$H_{(0)}$
- 卷积层：$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$
- 池化层：$\text{idx} = \text{top-rank}(Z, \lceil kN \rceil), \quad Z_{mask} = Z_{\text{idx}}$ $s = \frac{1}{N}\sum_{i=1}^{N} x_i \,||\, \max_{i=1}^{N} x_i$
- 全连接层

图算法应用3—图分类

数据准备工作：

图采样模块设计

- 动态图：对采样数据要做离线存储，对子图进行随机采样；
- 信息冗余：对子图进行预剪枝，过滤掉无关节点；
- 信息缺失：保留了缺失属性的节点.



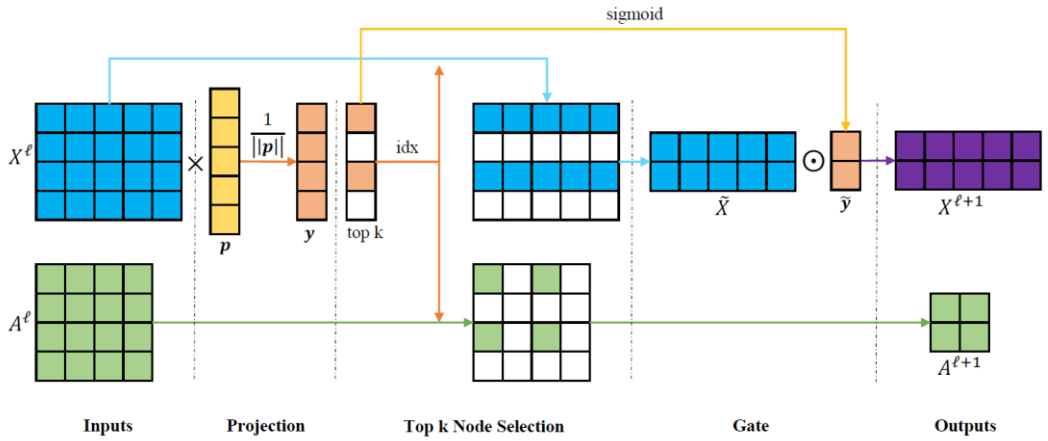
图数据加工模块：Dataset

利用nebula3、networkx 、Gephi完成图内容的解析、图重构、图可视化展示等：

- 邻接矩阵
- 节点的图标识
- 图类型
- 节点特征

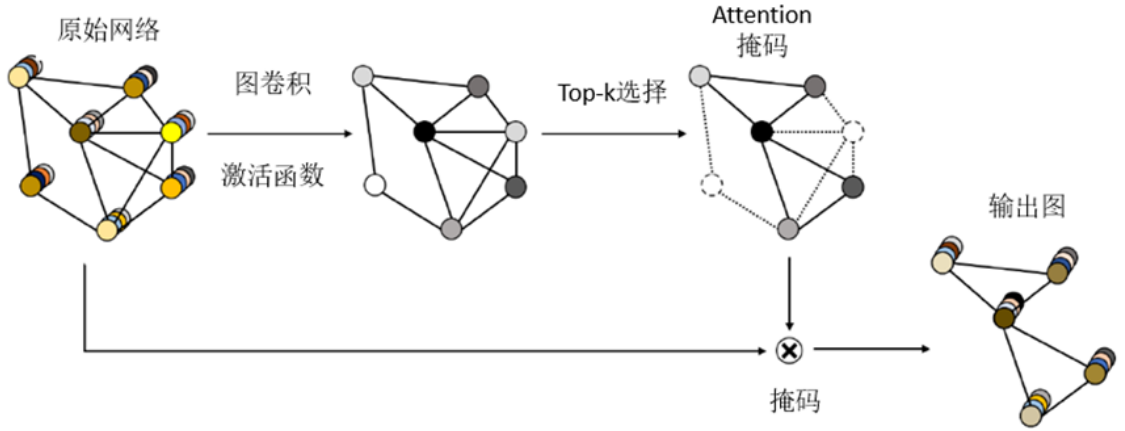
图算法应用3—图分类

Pooling常用于图分类任务中，常用的池化方式：



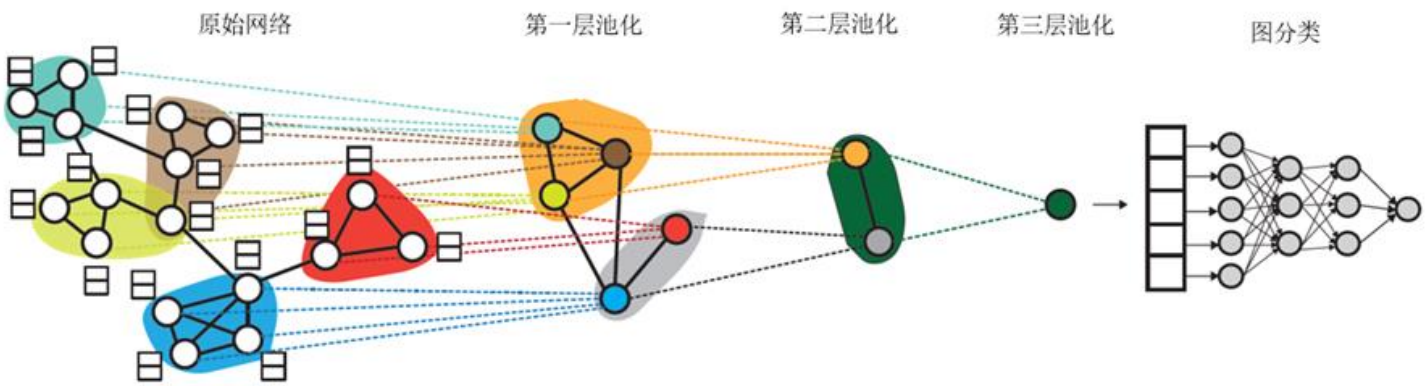
TopkPool：全局筛选，难以捕捉到局部信息.

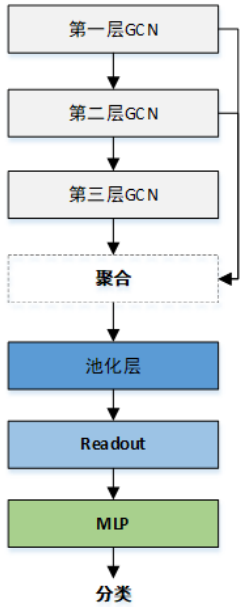
SAGPool：采用GNN获取节点的self-attention得分，融合图的特征和拓扑结构.

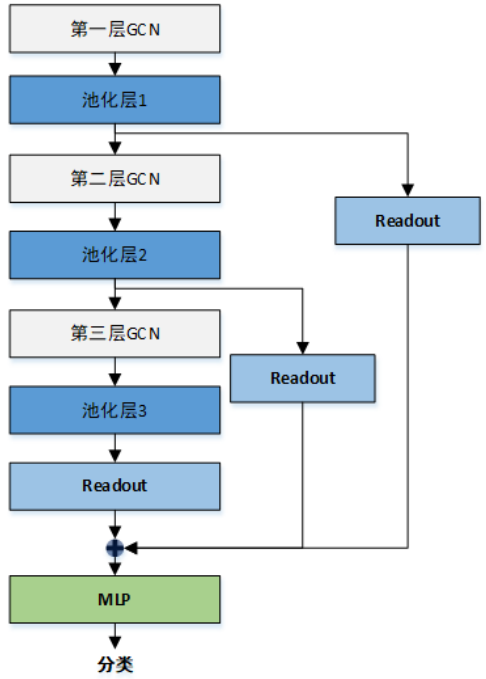


图算法应用3—图分类

DiffPool：采用GNN获取节点embedding表达，然后对节点进行聚类，生成一个粗粒度的图，重复以上过程.







Part 3

未来技术方向思考

未来技术方向思考

图技术能力建设的四个方向拓展：

- OneID技术在生产制造、零部件供应，等领域复用
- 面对制造业的图计算平台的开发
- 图神经网络在节点属性预测，链路预测应用
- 多GPU分布式计算的应用

文献综述

图神经网络综述类：

【1】Jie Zhou et al. Graph neural networks: A review of methods and applications.

【2】Ziwei Zhang et al. Deep Learning on Graphs: A Survey.

【3】Jie Zhou et al. Graph Neural Networks: A Review of Methods and Applications.

【4】Zonghan Wu et al. A Comprehensive Survey on Graph Neural Networks.

图表示学习Embeddings：

【5】**DeepWalk:** Perozzi et al. Online Learning of Social Representations.

【6】**LINE:** Jian Tang et al. Large-scale Information Network Embedding.

【7】**node2vec:** Aditya Grover et al. Scalable Feature Learning for Networks.

不同卷积核：

【8】**GCNConv:** Kipf and Welling et al. Semi-Supervised Classification with Graph Convolutional Networks.

【9】**ChebConv:** Defferrard et al. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering.

【10】**GATConv:** Veličković et al. Graph Attention Networks.

【11】**SAGEConv:** Hamilton et al. Inductive Representation Learning on Large Graphs.

文献综述

不同池化方式：

【12】**Top-K Pooling:** Gao and Ji et al. Graph U-Nets, Cangea et al. Towards Sparse Hierarchical Graph Classifiers and Knyazev et al. Understanding Attention and Generalization in Graph Neural Networks.

【13】**DiffPool**: Ying et al. Hierarchical Graph Representation Learning with Differentiable Pooling.

【14】**Set2Set**: Vinyals et al. Order Matters: Sequence to Sequence for Sets.

【15】**SAG Pooling:** Lee et al. Self-Attention Graph Pooling and Knyazev et al. Understanding Attention and Generalization in Graph Neural Networks.

【16】**ASAPooling:** Ranjan et al. ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations.


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS