TP Spark - GraphX

Data management for Big Data

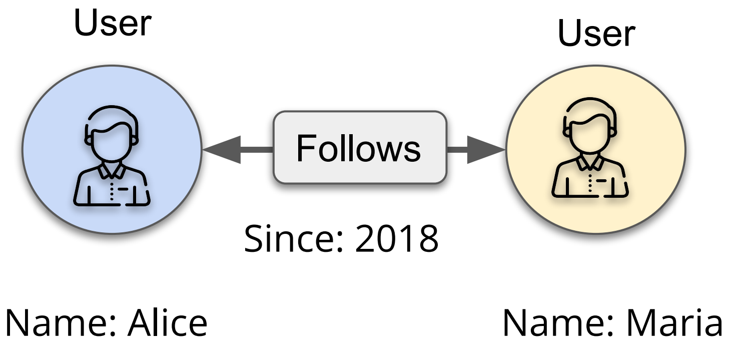
ECOLE SUPERIEURE
D'INGENIEURS DE RENNES

UNIVERSITÉ DE RENNES 1

Maria Massri
maria.massri@irisa.fr

Property graph model

A property graph is a collection of nodes and edges having each a set of properties.



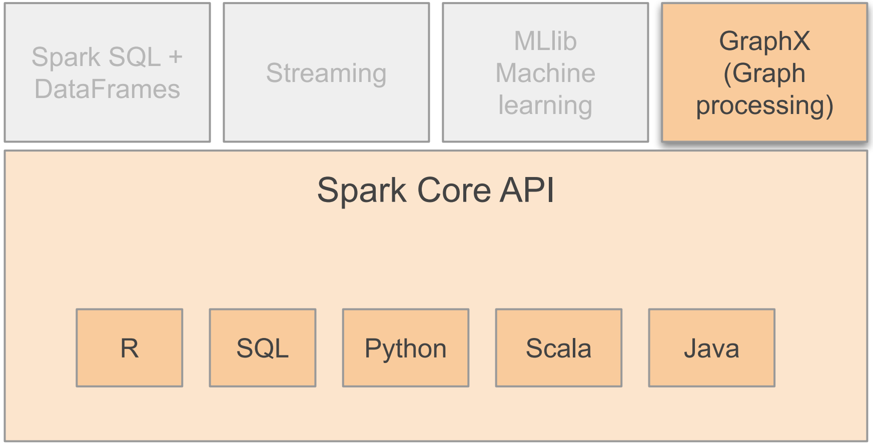
Graph processing frameworks

Graph processing framework is the set of tools that allows practitioners analyze graphs.

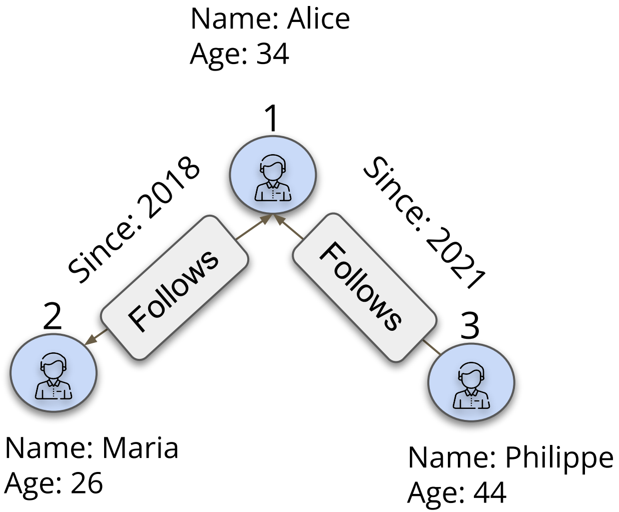
- **Social networks:** Rumor propagation, Community detection
- **Transaction networks:** Fraud detection (Anomaly detection)
- **Transportation networks:** Shortest paths, Finding trips, Congestion analysis

GraphX: Introduction

- GraphX is Spark library allowing distributed graph processing.
- It offers a graph abstraction and special graph operations to develop graph processing algorithms.



Graph creation



Graph[VD, ED]

Graph(UserRDD, followersRDD)

Graph constructor

RDD[(VertexId, VD)]

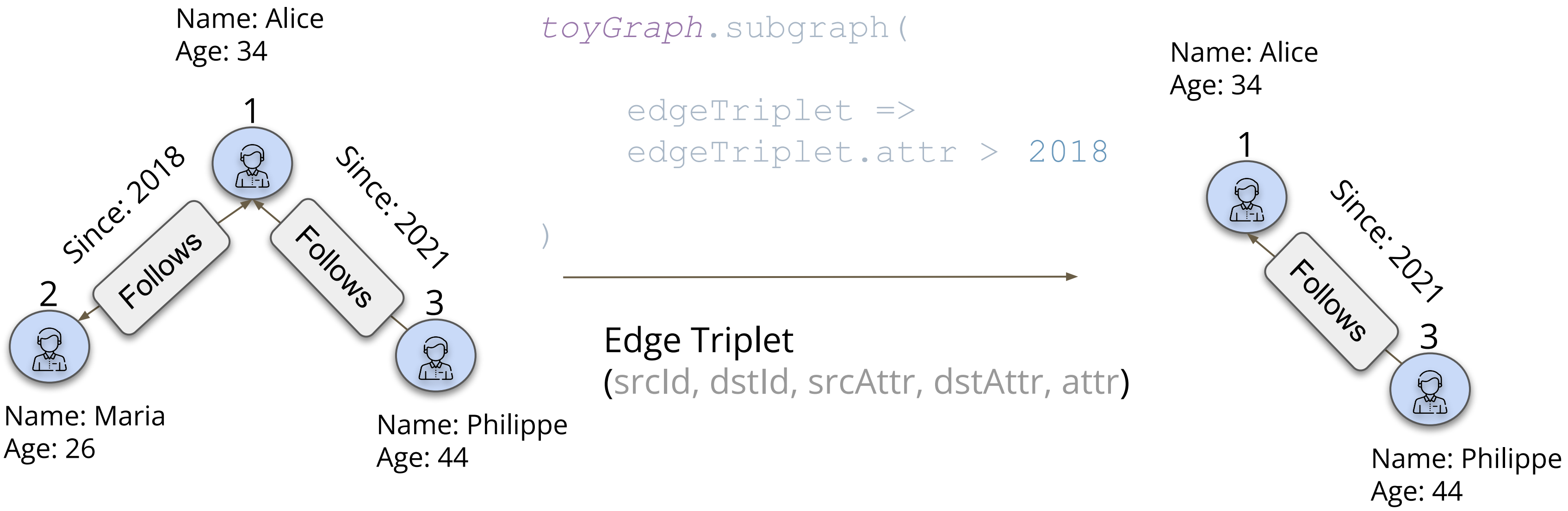
VertexId	User
1	User(“Alice”, 34)
2	User(“Maria”, 26)
3	User(“Philippe”, 44)

RDD[Edge[ED]]

Edge[int]
Edge(2, 1, 2018)
Edge(1, 2, 2018)
Edge(1, 3, 2021)

Subgraph extraction

The subgraph operator extracts a subgraph based on a condition of the triplet.

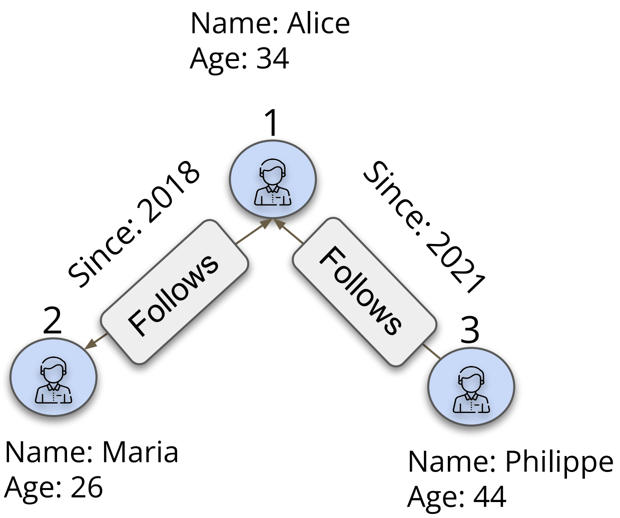


```
toyGraph.subgraph(

    edgeTriplet =>
    edgeTriplet.attr > 2018

)
```

Edge Triplet
(srcId, dstId, srcAttr, dstAttr, attr)

Aggregate message

Aggregate message operator allows to compute a local aggregated value (e.g. degree) for each node based on the information on the node itself, his edges, or neighbors.



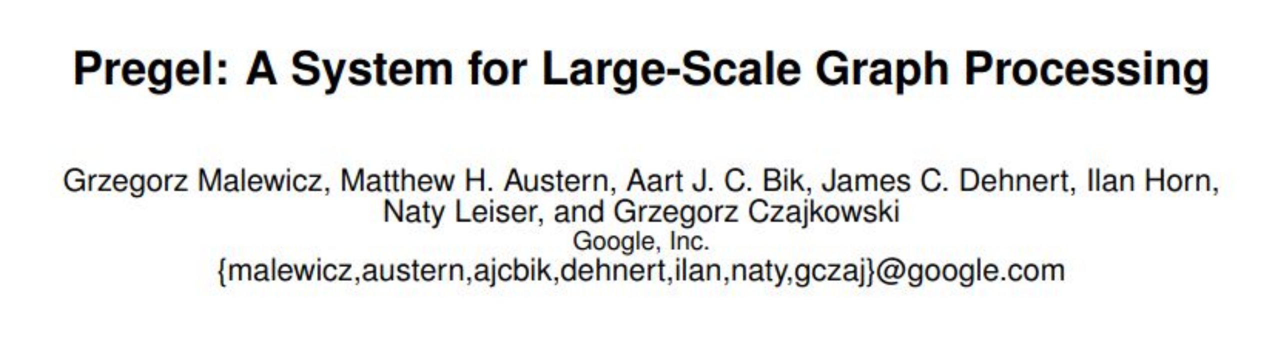
Extract incoming node degrees

```
toyGraph.aggregateMessages[Int](

    x => x.sendToDst(1),  (sendMsg)
    (x, y) => x+y        (mergeMsg)

)
```

VertexId	Int
1	2
2	1
3	0

Pregel

- Pregel was first outlined in a paper published by **Google** in 2010.
- It is a system for **Iterative large scale graph processing**.
- It inspired the development of Giraph for **Facebook** and **GraphX** as a library in **Spark**.

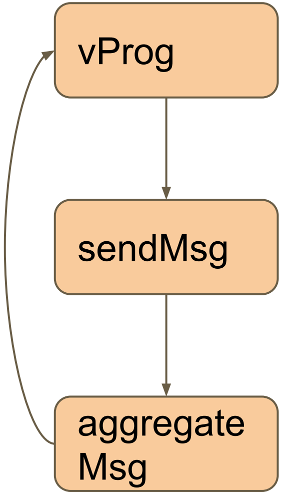


Pregel: A System for Large-Scale Graph Processing

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn,
Naty Leiser, and Grzegorz Czajkowski
Google, Inc.
{malewicz,austern,ajcbik,dehnert,ilan,naty,gczaj}@google.com

Cited by ~ 5000 references

Pregel: Iterative algorithm



vProg

sendMsg

aggregate Msg

Repeat until no message are received

Computes a new value for the node

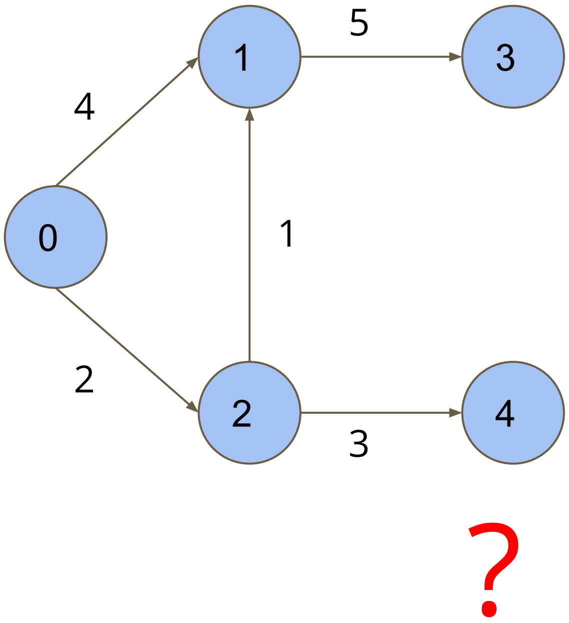
Invoked on all the out-edges and computes an optional message to destination vertex

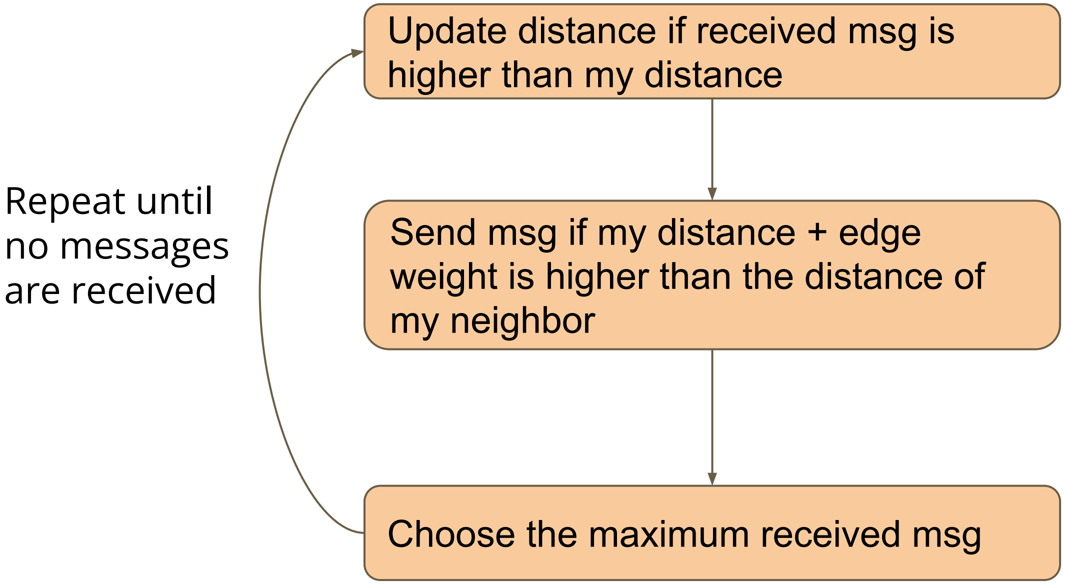
Used to combine message received by the vertex

- The Pregel algorithm will terminate when all the nodes stop receiving messages!

Pregel: Example of iterative algorithm

Compute, for each vertex, the maximum distance it can be reached with.





Repeat until no messages are received

Update distance if received msg is higher than my distance

Send msg if my distance + edge weight is higher than the distance of my neighbor

Choose the maximum received msg

TP SPARK

Data management for Big Data

Maria Massri
maria.massri@irisa.fr