

# Centrality Measures with Apache Flink

Nguyen Duc Hieu

December 07, 2018

Flink

# Flink

- ▶ framework and distributed processing engine
- ▶ stateful computations over unbounded and bounded data streams

# DataSet

- ▶ Immutable finite collection of data
- ▶ Elements can not be added, removed or inspected
- ▶ Elements can be either Simple Data(Integer, Double) or Tuples

# Transformation

Datasets can be transformed into new Dataset to get desired results

## Some Transformations

- ▶ Map
- ▶ FlatMap
- ▶ MapPartition
- ▶ Filter
- ▶ Projection of Tuple DataSet
- ▶ Reduce
- ▶ Combine
- ▶ Aggregate on full Tuple DataSet
- ▶ MinBy / MaxBy on full Tuple DataSet
- ▶ Distinct
- ▶ Join
- ▶ OuterJoin
- ▶ Cross
- ▶ CoGroup
- ▶ Union

## Centrality

# Centrality

- ▶ Identifying which vertices in a graph are most important:
  - ▶ most influential person(s) in social network
  - ▶ key infrastructure nodes in a computer network
  - ▶ point of spreading diseases

# Various Measures

- ▶ Degree Centrality
- ▶ Closeness Centrality
- ▶ Betweenness Centrality
- ▶ Eigenvector Centrality



## Degree Centrality

# Degree Centrality

- ▶ Number of ties a node has
- ▶ Degree can be interpreted as
  - ▶ Immediate risk to a node in a network catching a virus
  - ▶ Number of friendships or collaborations

Let  $G(V, E)$  and  $v^* \in G$  be the node with the highest degree.

The Degree Centrality of a Graph  $G$  can be defined as:

$$C_D(G) = \frac{\sum_{i=1}^{|V|} [deg(v^*) - deg(v_i)]}{(|V| - 2) * (|V| - 1)}$$

# Used Transformation from Flink API

- ▶ Map
- ▶ Cross

(data flow image here)

## K-Betweenes Centrality

## Betweenes Centrality

- ▶ Number of shortest Paths from all vertices to all other vertices going through a vertex

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

# K-Betweenes Centrality

- ▶ For Distributed Betweenes
- ▶ Borgatti and Everett (2006)
- ▶ K-Betweenes of a vertex  $v$  as the sum of pairs at most  $k$  apart, which are passing  $v$

$$C_{B,k}(v) = \sum_{s \neq v \neq t \in V: \text{dist}(s,t) \leq k} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- ▶ Graph API for Flink
  - ▶ containing set of utility functions for graph analysis
  - ▶ supports iterative graph processing
  - ▶ introduces a library of graph algorithms

Pregel, BSP in Gelly



## Implementation with Gelly