

# ScaleGraph/Python Overview

このドキュメントは、大規模グラフデータに対するグラフ計算ライブラリである ScaleGraph の Python 拡張である ScaleGraph/Python 拡張について、概要を説明するものです。

## Programmer Interface

ScaleGraph/Python は、ScaleGraph ライブラリがもつ、下記のグラフアルゴリズムおよび Pregel 計算機能を、Python から利用できるようにしたものです。

## Graph Algorithm Computation

以下の8つのアルゴリズムを Python function として利用できる Python module を提供しています。

- PageRank
- Degree Distribution
- Betweenness Centrality
- HyperANF
- Strongly Connected Component
- Minimum Spanning Tree
- Max Flow
- Spectral Clustering

## Pregel Model Computation

グラフ計算で利用するアルゴリズムを、ユーザが定義することができます。ユーザは以下の計算を Python コードとして記述します。この Pregel Model Computation では、すべての頂点に対して、1回ずつ Compute を実行し、Compute の結果を Aggregator で集約することを Super Step と呼びます。Super Step を複数回繰り返すことでグラフ計算をすすめます。

- Compute

- 各頂点において実行する計算の記述です。
- Aggregator
  - Computeの結果を集約する計算の記述です。
- Terminator
  - 次のSuper Stepを実行するか判定する式の記述です。

## Computation Environment

### Task

ScaleGraph/Pythonは計算のバックエンドとしてX10で記述されているScaleGraphライブラリを利用します。ScaleGraphが動くクラスタ環境とネットワーク接続されているデスクトップで、ユーザがScaleGraph/Pythonを利用することを想定しています。計算タスクはクラスタ環境で実行されます。ユーザがApache PySpark等他のグラフ計算ライブラリと一緒にScaleGraph/Pythonを利用することを想定しています。

### Storage

ScaleGraphは大規模なグラフデータに対する高速なグラフ計算を提供するライブラリですが、グラフデータの入出力は標準的なUnix File Systemのみを想定したものでした。すなわち、計算を実行するクラスタ環境が高速な共有ファイルシステムをもつことを前提としていました。

ScaleGraph/Pythonにおいては、標準のUnix File Systemだけではなく、Hadoop File System (HDFS)をサポートしています。前述のとおりユーザがApache PySparkを同時に利用することを想定していますが、Apache PySparkにおけるRDD計算の結果はHDFS経由でScaleGraph/Pythonと受け渡しすることができます。

