

# ビッグデータを瞬時に解析する

Data Skewnessを捉えた超高速・省メモリな大規模データ処理

塩川浩昭（筑波大学 計算科学研究センター・助教）

✉ shiokawa@cs.tsukuba.ac.jp



筑波大学  
University of Tsukuba

## 研究概要

### ● 解決したい課題

限られた計算資源の中でいかにして  
大規模データを高速に計算するか？

#### 【現在の大規模データ処理の課題】

- 大規模データ処理に必要な計算コストと普及した計算機の性能にはギャップがある

#### 【本研究の目的】

- 高速・省メモリなデータ処理アルゴリズムの開発により、ギャップの解消を目指す

### ● 研究成果の概要

**30億件**のデータを**6.7秒**で計算可能に

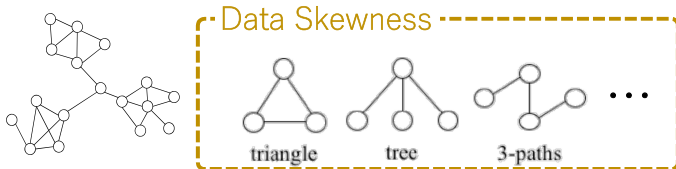
- グラフデータを対象とした、超高速・省メモリな大規模データ処理アルゴリズムの設計法を提案
- Desktop PCやLaptopなどの計算環境で、従来技術の精度を維持しつつ、**約1,000倍の高速化**を達成
- 理論的な解析により、スケールフリーなグラフにおいて、計算回数を**従来比で数千分の1程度**に抑えられることを証明

## 研究成果の詳細

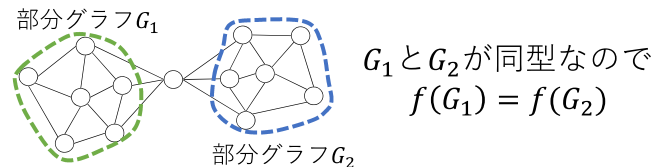
### ● 基本的なアプローチ：Data Skewnessを捉えた計算コスト削減

- Data Skewness** = 「実データのみが持つ頻出する（特徴的な）部分構造」に着眼
- Data Skewnessを捉えることで、冗長な計算を大幅に削減する

#### 【着眼点①】 実データ偏った局所構造を持つ

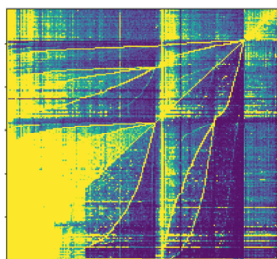
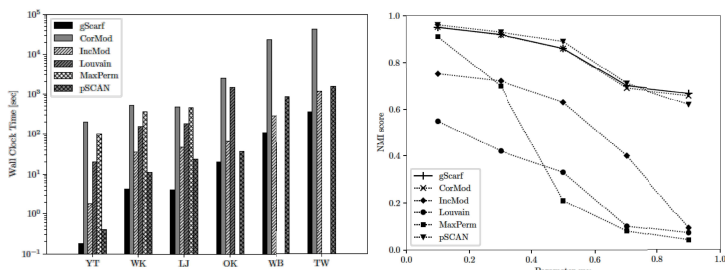


#### 【着眼点②】 局所構造に対して決定性がある

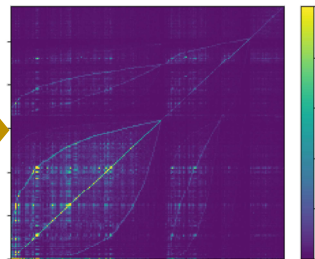


### ● 研究成果①：gSscarf法

- Modularityクラスタリングに対する高速化手法
- 正確な処理結果を**1,000倍以上高速**に計算可能



1%未満  
に抑制



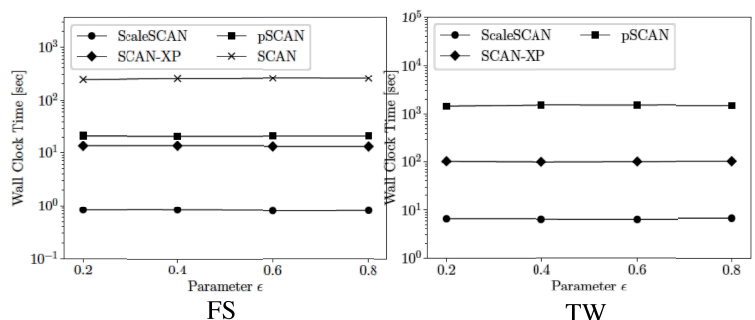
[Duan et al., KDD 2015]

提案手法 gSscarf

計算回数ヒストグラムの比較

### ● 研究成果②：ScaleSCAN法

- 密度ベースクラスタリングに対する高速化手法
- Twitterデータ(**約30億エッジ**)に対して正確なクラスタを**6.7秒**で計算可能 (500~1,000倍の高速化)



### ● その他の成果：

- ラベル付きグラフでのキーワード検索の高速化
- 分散グラフ処理の効率化
- 多次元データクラスタリング、機械学習の高速化