High Quality, Scalable and Parallel Community Detection for Large Real Graphs

Arnau Prat-Pérez, David Dominguez-Sal, Josep-Lluis Larriba-Pey

To be published in WWW 2014

GraphTA 2014
A brief introduction

• **Real networks** are structured into modules called **communities**.
• Informally, communities are groups of nodes more **highly connected** between them than with the rest of the graph.
• Several applications:
  • **Recommend** users, products, sites, etc...
  • Find **similar proteins**.
  • **Visualization** of large data.
• However:
  • **Non consensuated definition**.
  • **Computationally expensive**.
Goal

• Find algorithms with a strong focus on:
  - Quality
  - Scalability
  - Parallelism

• **Scalable Comunity Detection (SCD)** is our first proposal for scalable **disjoint community** detection algorithm for **SMP machines**.
Weighted Community Clustering (CIKM12)

- The **Weighted Community Clustering** (WCC) ranks the quality of a graph partition into **disjoint communities**
- Strongly based on **triangles**.
- Good structural indicators:
  - large internal edge density
  - small diameter
  - small conductance
Scalable Community Detection (SCD)

- **WCC** as heuristic.
- Transfers can be computed in parallel.
- We approximate the WCC using a proposed estimator.
- Algorithm's complexity: $O(m \cdot \log n)$
Some results

- 4-Core Xeon with 32GB of RAM.
- **Quality**: better than state of the art, using **real graphs** with **ground truth communities**.
- **Scalability**: faster than the fastest. Graphs with **1.8 Billions edges** processed in **just 4.3 hours**.
- **Parallelism**: almost **linear speed-up** for large graphs.
- Verified **quasi-linear complexity**.
Summary

• Triangles allow us to find communities fast and reliably.
• We exploit the characteristics of current multicore microprocessors.
• Paper and code available at www.dama.upc.edu.