Graph Analytics for Community Detection with GraphLab

Petko Georgiev

Motivation

- Community detection algorithms
  - tools for the analysis and understanding of network data
  - applications in social, technological and biological networks
- High-quality algorithms are slow!
- Some algorithms can be run only on graphs with hundreds of vertices

GraphLab's execution model comes to the rescue

- Data graph (data/computation dependencies)
- Update functions (local computation)
- Sync mechanism
- Consistency model (full, edge, vertex)
- Scheduling primitives

Think-like-a-vertex as in Pregel

- Each vertex has user defined functions:
  - Gather
  - Apply
  - Scatter

- GraphLab also supports asynchronous convergence testing

GraphLab Toolkits

Toolkit	Algorithms
Topic Modeling	LDA
Graph Analytics	PageRank, K-cores Decomposition, Triangle Counting, Connected Components, Graph Colouring
Clustering	K-means++, Spectral Clustering
Collaborative Filtering	ALS, SGD, SVD++ and variants
Graphical Models	Structured Prediction
Computer Vision	Image-Stitching

GraphLab Toolkits++

Toolkit	Algorithms
Topic Modeling	LDA
Graph Analytics	PageRank, K-cores Decomposition, Triangle Counting, Connected Components, Graph Colouring
Clustering	K-means++, Spectral Clustering
Collaborative Filtering	ALS, SGD, SVD++ and variants
Graphical Models	Structured Prediction
Computer Vision	Image-Stitching
Community Detection	**TBA**

Aim of study

- Build a community detection toolkit
- Evaluate the flexibility of GraphLab's API
- Extract commonalities in the parallel/distributed algorithm design
- Measure speed-up on multicore and distributed environments
- Evaluate performance benefits for large graphs

Community detection algorithms

Algorithm	Type	Status
Kernighan-Lin Modularity Maximisation	Divisive	Implemented
Spectral Modularity Maximisation	Divisive	In Progress
Louvain Fast Modularity	Agglomerative	Tentative
Betweenness-based	Divisive	Tentative
Radicchi et al.	Divisive	Tentative
Simulated Annealing	Optimisation	Tentative
Genetic Algorithms	Optimisation	Tentative
Hierarchical Clustering	Agglomerative	Tentative

Challenges

- Not all algorithms fit into the “think-like-a-vertex” model
- Algorithms have several phases
- Overhead of parallel implementations for small graphs
- One algorithm is already quite fast (Louvain fast modularity is $O(n \log^2 n)$ for sparse graphs)

Further work

- More algorithms…
- Distributed deployment (EC2)
- Performance analysis
  - Multicore environment
  - Distributed environment

References

- Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein (2010). "**GraphLab: A New Parallel Framework for Machine Learning**." *Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin and Joseph M. Hellerstein (2012). "**Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud**."*PVLDB*.
- M. E. J. Newman (2010). ***Networks: An Introduction***. Oxford: Oxford University Press. ISBN 0-19-920665-1