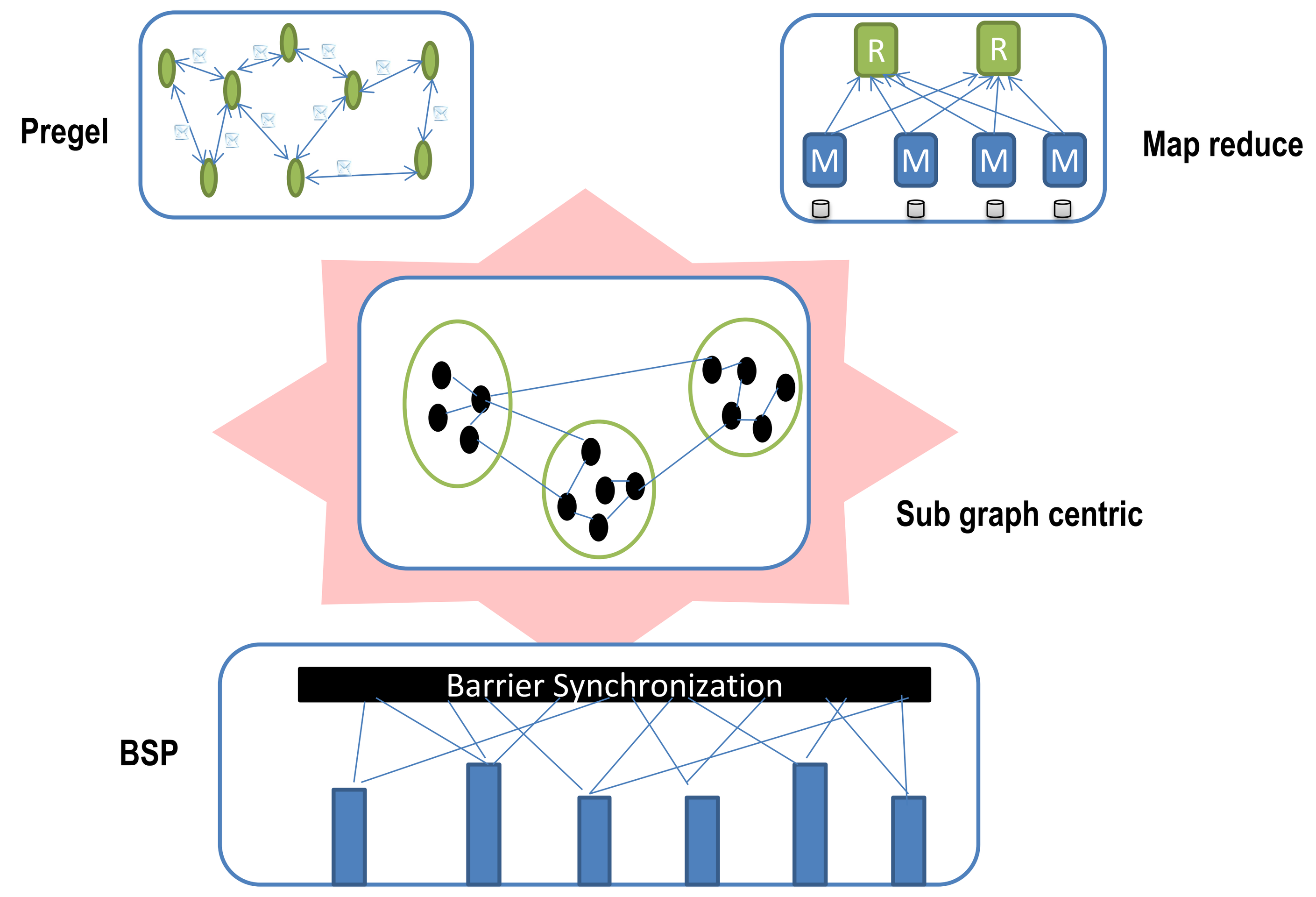
Sub-graph centric programming framework for large scale time series graphs

Charith Wickramaarachchi, Alok Kumbhare

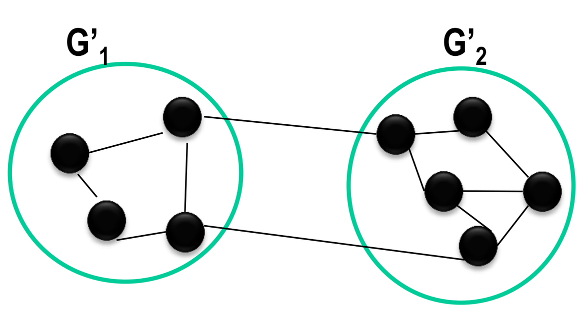
Advisors: Yogesh Simmhan & Viktor Prasanna

USC Viterbi School of Engineering

1. Introduction



- Limitations of Map Reduce[1] and Pregel[2] models for graph algorithms.
  - Map reduce – Do not consider the topology and locality of graphs.
  - Pregel - Costs large number of coordination steps for some class of graph algorithms due to vertex centric nature
- Bulk Synchronous parallel - Abstract computer model for designing data parallel algorithms.
  - Computation proceeds in super steps which consists of
    - Computation
    - Communication
    - Barrier Synchronization



Sub-graph centric programming model

- Sub-graph G' =(V',E')
  - connected subset of vertices of a Graph G = (V,E)

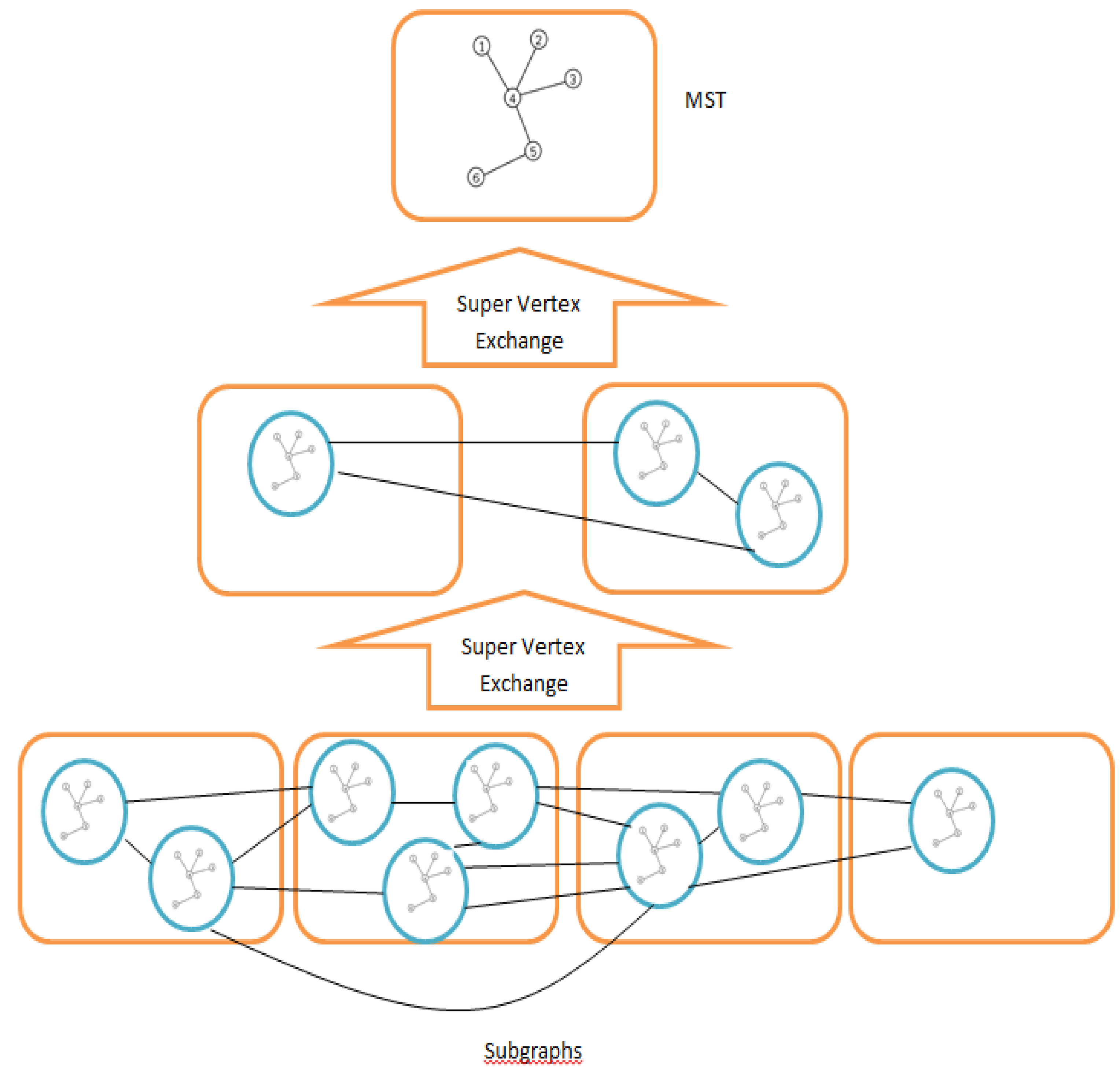
$$V' \subset V, E' \subset E$$

- Sub-Graph centric programming model
  - Operates on a sub-graphs concurrently
  - Communicate between sub-graphs

2. Motivation

Sub graph centric graph algorithms

- Operates on sub-graphs which will run in parallel.
- Sub graphs can communicate with each other in each super step.

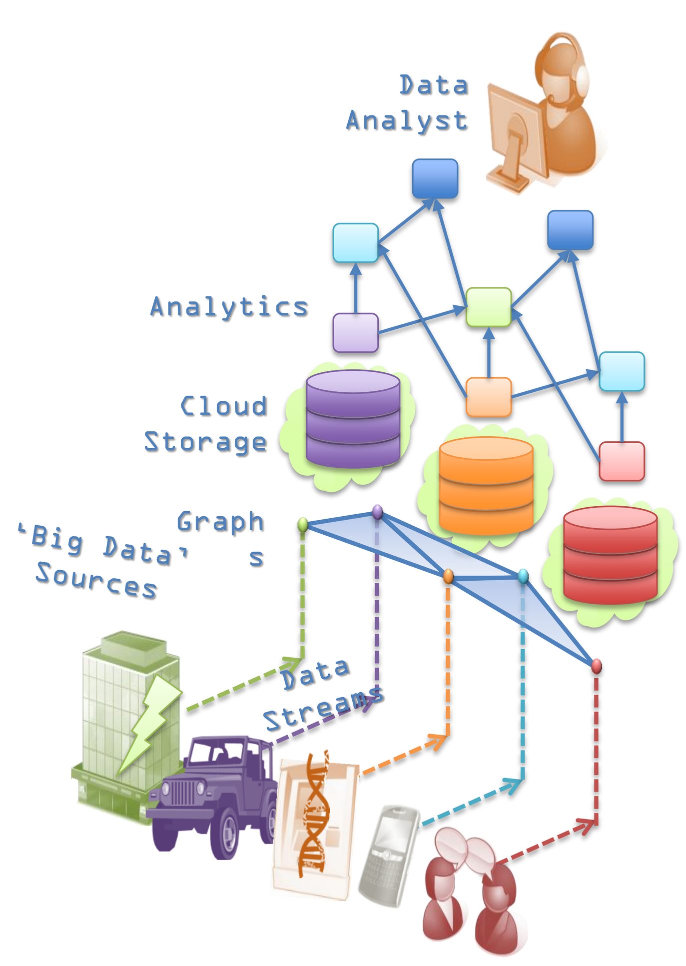


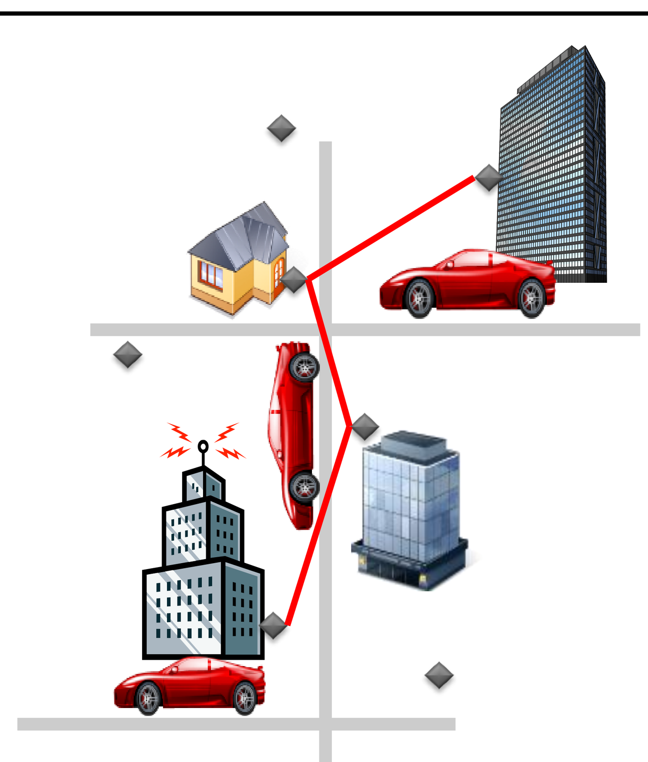
Sub graph centric Minimum Spanning Tree algorithm

- Build Min spanning forest in each sub -graph locally with Borůvka's algorithm
- Extends Karloff et al. [3] s MST algorithm using a sub-graph centric approach and reduce the graph size.

3. Time Series Graphs

- Fixed graph of event sources
  - Known relationships between sources E.g. pathway
  - E.g. "license plate detected" event generated by a camera
- Streams of events form graph time-series
  - Snapshots of graphs over time
- Graph Template G_T=(V,E)
- Graph instance G_i = (V_t,E_t,t_i)
- Time Series Graph T_G= (G_T, G_1,G_2,..G_k)





Time series graph applications

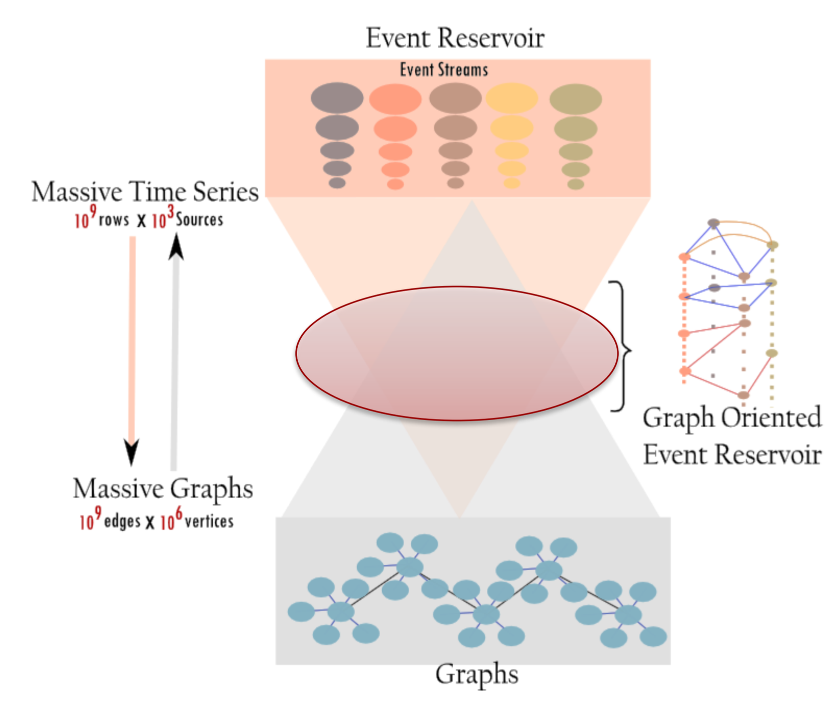
- License plate recognition systems scan the license plates of moving or parked vehicles.
  - Large amount of time series data
  - Applications :
    - Finding hot routes / spots for Auto theft.

4. GoFFish

Analytics & Time-series Graph Applications

Gopher
Graph-oriented Programming Framework

GoFS
Graph-oriented Distributed File System



- Graph-oriented File System
  - Relationships *between* event sources and *across* time are captured in the data model
  - Layout is key: How can we maximize parallelism & minimize disk access for analytics
- Graph Programming Framework
  - Abstractions sensitive to event relationships & time-series
  - Sub-graph centric operations that span graph instances
  - Analytics composed as a dataflow
  - Framework aware of distributed layout
  - Limit coordination overhead
- Works on Intersection of time-series event data & graph data structures

References

[1] MapReduce: Simplified Data Processing on Large Clusters , Jeffrey Dean and Sanjay Ghemawat , http://dl.acm.org/citation.cfm?id=1327492
[2] Pregel: a system for large-scale graph processing , Grzegorz Malewicz et al. http://dl.acm.org/citation.cfm?id=1807184
[3] A Model of Computation for MapReduce,Karloff et al. ,http://dl.acm.org/citation.cfm?id=1873677

This work is supported by Defense Advanced Research Projects Agency - XDATA program and the National Science foundation Co – PI s : Viktor Prasanna , Yogesh Simmhan , Raghu Raghavendra
The views of authors expressed herein do not necessarily reflect those of the sponsors.

http://ganges.usc.edu

University of Southern California