To Revisit Benchmarking Graph Analytics

A work collaborated by Shanghai Jiao Tong University and Alibaba Damo Academy

Presented by Longbin Lai, on behalf of Prof. Xuemin Lin

LDBC Graphalytics Benchmark

- R1 Target platforms and systems
- R2 Diverse, representative benchmark elements: Algorithms, Datasets, etc.
- R3 Diverse, representative process: Performance, Scalability and Robustness
- R4 Include a renewal process
- R5 Modern software engineering

[1] Alexandru Losup, LDBC Graphalytics: A Benchmark for Large-Scale Graph Analysis on Parallel and Distributed Platforms, VLDB 2016

Why revisit the benchmark: Algorithm

- Selected algorithms are **representative** but not **diverse**

1 Algorithms: BFS, PR, WCC, CDLP, LCC, SSSP

2 Similar Computing Patterns: **ISVP (iterative, single-phased and value-propagation-based**)

3 The **appearance-dominated** selection procedure is biased

[2] V Kalavri, V Vlassov, S Haridi, High-level programming abstractions for distributed graph processing, TKDE 2017

Our Proposal: Categorization

- **Centrality:** PageRank、 Personalized PageRank、 Degree Centrality、 Betweenness Centrality、 Closeness Centrality
- **Clustering/Community Detection:** Local Clustering Coefficient、 Louvain、 Label Propagation、 Minimum Cut Algorithm
- **Similarity:** Cosine、 Jaccard、 SimRank
- **Community Search:** Core Decomposition、 K-Truss、 Clique、 K-ECC、 Biclique
- **Pattern Matching:** Triangle Counting、 Subgraph Matching
- **Traversal/Path:** BFS、 DFS、 Single Source Shortest Path、 Topological Sort、 Minimum Spanning Tree、 Max Flow、 Cycle Detection
- **Other:** Strongly Connected Components、 Weakly Connected Components、 Maximum Independent Set、 Color

Selection of LDBC

Our Proposal: Multi-dimensional

Algorithms	Number of Papers	DBLP	Google Scholar	Web of Science	Time Complexity
Label Propagation	39	771	130000	699	$k * m$
Single Source Shortest Path	33	584	17800	282	$m + n * \log n$
K-Clique	31	352	39500	73	$k * m * a^{k-2}$
Core Decomposition	29	179	107000	454	$m + n$
PageRank	28	1012	21700	753	$k * m$
Triangle Counting	27	252	21700	210	$m^{1.5}$
Betweenness Centrality	20	304	32100	283	n^3
Louvain	8	299	181000	127	$n * \log n$

⇩ Categories

⇩ Appearances

Academic Search Engines

⇩ Textbook Complexity

Why revisit the benchmark: Datasets

- Selected datasets are narrow in

Characteristics

Real	Gen	Model
Social (Gaming)	SNB	Small-world
Knowledge	Graph500	Power-law

Graphs in real life are more diverse:

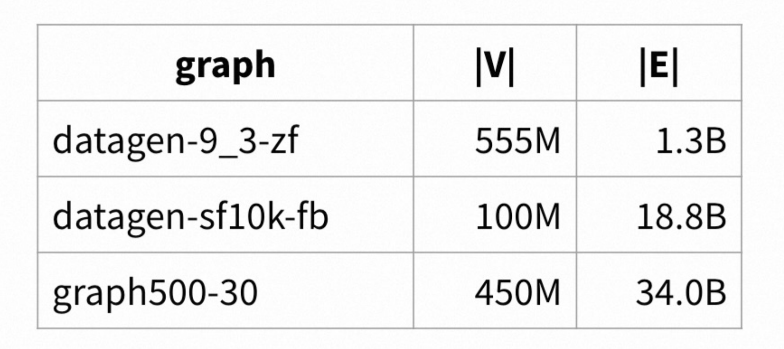
❖ Road/route networks are sparse

❖ Product-customer graphs are bi-partite

❖ etc.

Sizes

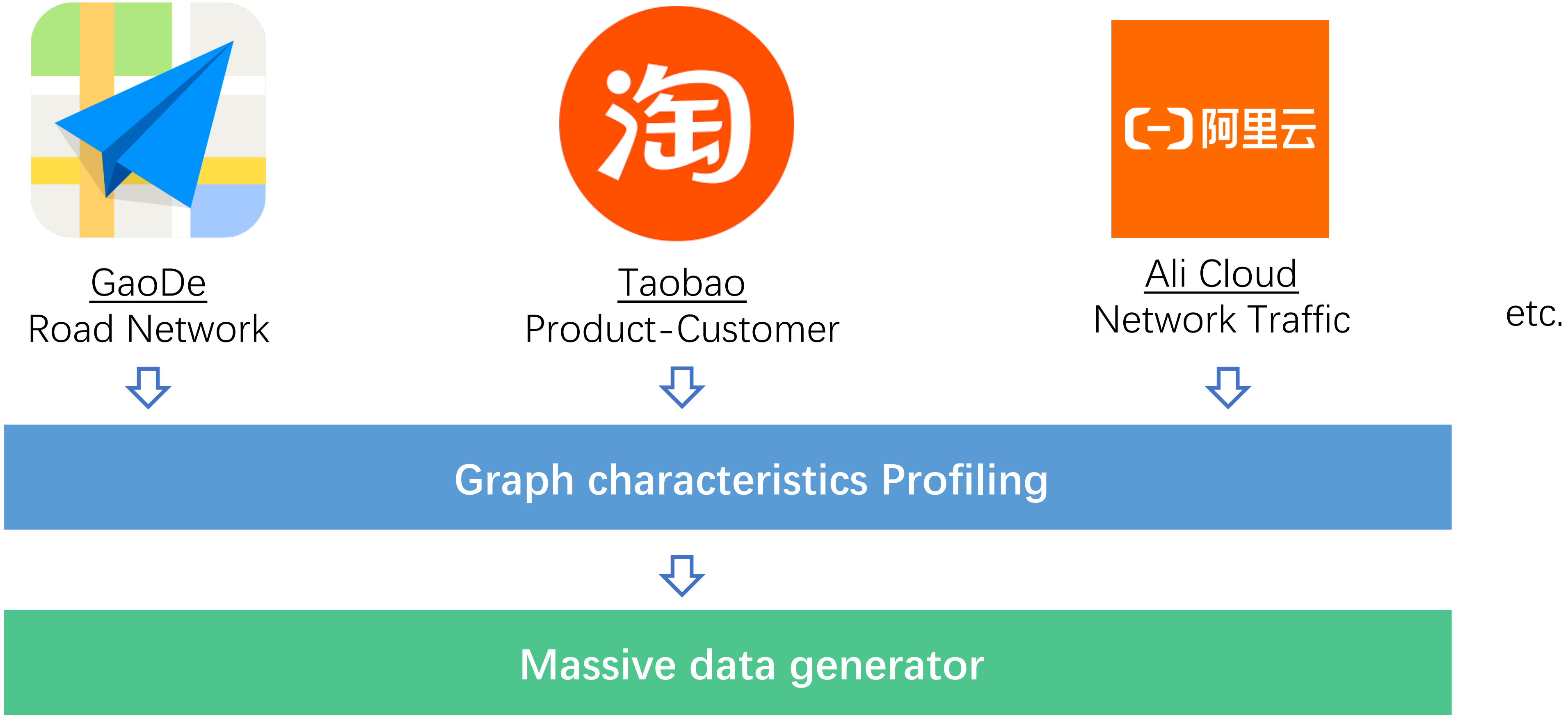
The largest real-life dataset (twitter-mpi) has only ~2B edges



graph	|V|	|E|
datagen-9_3-zf	555M	1.3B
datagen-sf10k-fb	100M	18.8B
graph500-30	450M	34.0B

The latest graphalytics challenge includes much larger generated data

Our Proposal: Gen with real-life characteristics



GaoDe
Road Network

Taobao
Product-Customer

Ali Cloud
Network Traffic

etc.

Graph characteristics Profiling

Massive data generator

Why revisit the benchmark: Process

- Platform-oriented Process
    - Performance: Makespan, Processing time
    - Scalability: Speedup
    - Robustness: Stress-test, Performance variability
- Our proposal
    - Platform-oriented + User-oriented
    - User-oriented
        - Expressiveness: **can** user implement certain algorithm
        - Productivity: **how (easy)** can user implement certain algorithm

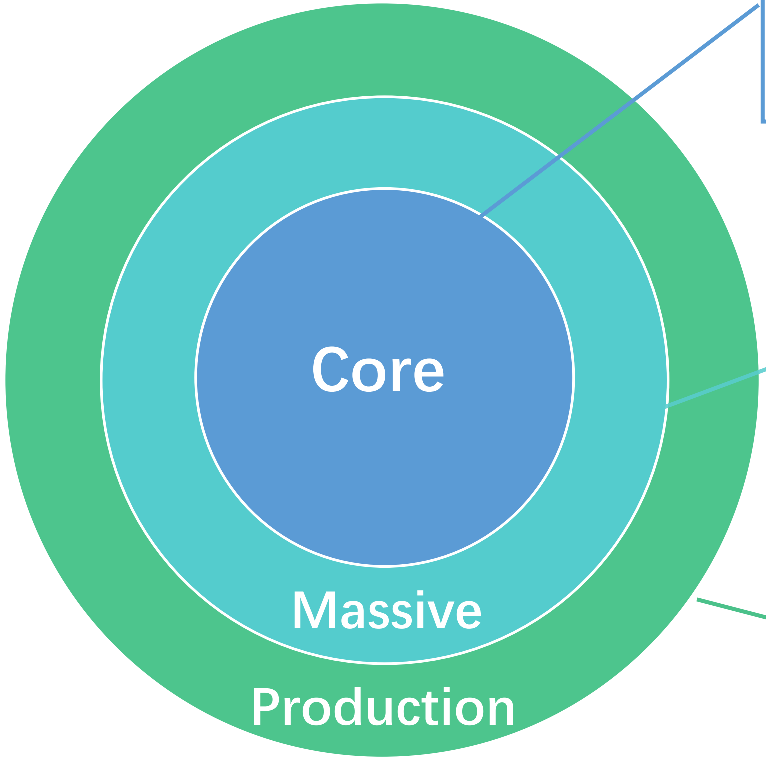
Why revisit the benchmark: Software

- Modern but not **golden**
  - Software dependency issues
  - Repeated generation of some data
  - Hard to deploy in a cluster for large-scale
- Our Proposal:
  - Go cloud-native
    - Docker image: resolve software dependency issues
    - Cloud storage: for archiving the data (without repeatedly generating)
    - K8s for easy deployment in a cluster
    - etc

Wait, will this complicate the benchmark?

- More algorithms
- More/Larger datasets
- More metrics to evaluate

Our Proposal: Benchmark Hierarchies



- 8 core algorithms
- Others remain as LDBC Graphalytics
- **Suitable for competition/challenge**

- 8 core algorithms
- Massive datasets + Docker/Cloud Services, Perf metrics remain as LDBC Graphalytics
- **Suitable for PoC & HPC**

- 20+ algorithms
- Massive datasets + Docker/Cloud Services + User-oriented perf metrics
- **Suitable for new-product promotion**

THANKS