GRAPHALYTICS
A Big Data Benchmark for Graph-Processing Platforms

Mihai Capotâ, Yong Guo, Ana Lucia Varbanescu,
Tim Hegeman, Jorai Rijsdijk,
Alexandru Iosup,
Jose Larriba Pey, Arnaud Prat, Peter Boncz, Hassan Chafi

GRAPHALYTICS was made possible by a generous contribution from Oracle.
TU Delft - the Netherlands - Europe

Delft
- Founded: 13th century
- Population: 100,000

Barcelona
- Population: 16.5 million

Delft
- Founded: 1842
- Population: 13,000
The Parallel and Distributed Systems Group at TU Delft

Home page
- www.pds.ewi.tudelft.nl

Publications
- see PDS publication database at publications.st.ewi.tudelft.nl

Winners IEEE TCSC Scale Challenge 2014
Graphs at the Core of Our Society: The LinkedIn Example
The State of LinkedIn

A very good resource for matchmaking workforce and prospective employers

Vital for your company’s life, as your Head of HR would tell you

Vital for the prospective employees

100M Mar 2011, 69M May 2010

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
Graphs at the Core of Our Society: The LinkedIn Example

The State of LinkedIn

Sources: Vincenzo Cosenza, The State of LinkedIn, [http://vincos.it/the-state-of-linkedin/](http://vincos.it/the-state-of-linkedin/)

We now have 300 million LinkedIn members, more than half of whom live outside of the U.S. That’s enough to make LinkedIn the fourth largest country in the world. In celebration, we took a look back to see how much our membership has grown and diversified over the past five years. It’s a helpful reminder of not only where we’ve been, but also where we’re headed as we work to create economic opportunity for every professional in the world.
The data deluge: large-scale graphs

LinkedIn
300M users
??? edges

Twitter
270M MAU
200+ avg followers
>54B edges

Yahoo!

Friendster

Facebook
1.2B MAU 0.8B DAU
200+ avg followers
>240B edges
The data deluge: large-scale graphs

Linkedin
Oracle 1.2M followers, 132k employees
company/day: 40-60 posts, 500-700 comments

Twitter
270M MAU
200+ avg followers
>54B edges

Facebook
1.2B MAU 0.8B DAU
200+ avg followers
>240B edges

Friendster

Oracle

TU Delft
Delft University of Technology
The data deluge: large-scale graphs

Linkedin
Oracle 1.2M followers

Twitter
270M MAU

Yahoo!

Friendster

Data-intensive workload
10x graph size ➔ 100x—1,000x slower

Apple

1.2B MAU 0.8B DAU
200+ avg followers

>240B edges

TU Delft
The data deluge: large-scale graphs

Linkedin

Oracle 1.2M followers, 132k employees

Data-intensive workload
10x graph size $\Rightarrow$ 100x–1,000x slower

Compute-intensive workload
more complex analysis $\Rightarrow$ ?x slower

friendster

Twitter

270M MAU

$>240B$ edges

$1.2B$ MAU, $0.8B$ DAU

$>200+$ avg followers

$200+$ avg followers
The data deluge: large-scale graphs

Linkedin
Oracle 1.2M followers

Data-intensive workload
10x graph size $\Rightarrow$ 100x—1,000x slower

Compute-intensive workload
more complex analysis $\Rightarrow$ ?x slower

Dataset-dependent workload
unfriendly graphs $\Rightarrow$ ??x slower
Graphs at the Core of Our Society: The LinkedIn Example

The State of LinkedIn

3-4 new users every second

but fewer visitors (and page views)

Great, if you can process this graph: opinion mining, hub detection, etc.

100+ million questions of customer retention, of (lost) customer influence, of ...

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
Graphs at the Core of Our Society: The LinkedIn Example

The State of LinkedIn

3-4 new users every second

but fewer visitors (and page views)

Periodic and/or continuous analytics at full scale

Sources: Vincenzo Cosenza, The State of LinkedIn, http://vincos.it/the-state-of-linkedin/
The “sorry, but...” moment

Supporting multiple users

10x number of users ➞ ????x slower
Graph Processing @large

A Graph Processing Platform

Interactive processing not considered in this presentation.
Streaming not considered in this presentation.
Interactive processing not considered in this presentation.
Streaming not considered in this presentation.
Graph-Processing Platforms

- Platform: the combined hardware, software, and programming system that is being used to complete a graph processing task

Which to choose? What to tune?
What is the performance of graph-processing platforms?

- **Metrics Diversity**
- **Graph Diversity**
- **Algorithm Diversity**

Graphalytics = comprehensive benchmarking suite for graph processing across all platforms
Graphalytics = A Challenging Benchmarking Process

- Methodological challenges
  - Challenge 1. Evaluation process
  - Challenge 2. Selection and design of performance metrics
  - Challenge 3. Dataset selection and analysis of coverage
  - Challenge 4. Algorithm selection and analysis of coverage

- Practical challenges
  - Challenge 5. Scalability of evaluation, selection processes
  - Challenge 6. Portability
  - Challenge 7. Result reporting

Graphalytics = Many Classes of Algorithms

- Literature survey of metrics, datasets, and algorithms
  - 10 top research conferences: SIGMOD, VLDB, HPDC …
  - Key word: graph processing, social network
  - 2009–2013, 124 articles

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph Statistics</td>
<td>Diameter, PageRank</td>
<td>16.1</td>
</tr>
<tr>
<td>Graph Traversal</td>
<td>BFS, SSSP, DFS</td>
<td>46.3</td>
</tr>
<tr>
<td>Connected Component</td>
<td>Reachability, BiCC</td>
<td>13.4</td>
</tr>
<tr>
<td>Community Detection</td>
<td>Clustering, Nearest Neighbor</td>
<td>5.4</td>
</tr>
<tr>
<td>Graph Evolution</td>
<td>Forest Fire Model, PAM</td>
<td>4.0</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Future work

Graphalytics = Real & Synthetic Datasets

<table>
<thead>
<tr>
<th>Graphs</th>
<th>#V</th>
<th>#E</th>
<th>d</th>
<th>̅D</th>
<th>Directivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 Amazon</td>
<td>262,111</td>
<td>1,234,877</td>
<td>1.8</td>
<td>4.7</td>
<td>directed</td>
</tr>
<tr>
<td>G2 WikiTalk</td>
<td>2,388,953</td>
<td>5,018,445</td>
<td>0.1</td>
<td>2.1</td>
<td>directed</td>
</tr>
<tr>
<td>G3 KGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>undirected</td>
</tr>
<tr>
<td>G4 Citation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>directed</td>
</tr>
<tr>
<td>G5 DotaLeague</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>undirected</td>
</tr>
<tr>
<td>G6 Synth</td>
<td>2,594,350</td>
<td>64,132,045</td>
<td>2.2</td>
<td>55.0</td>
<td>undirected</td>
</tr>
<tr>
<td>G7 Friendster</td>
<td>65,608,366</td>
<td>1,806,067,135</td>
<td>0.1</td>
<td>55.1</td>
<td>undirected</td>
</tr>
</tbody>
</table>

**Interaction graphs (possible work)**

---


---

http://www.graph500.org/

http://gta.st.ewi.tudelft.nl/

---

https://snap.stanford.edu/
Graphalytics = Advanced Harness

Cloud support technically feasible, methodologically difficult
Graphalytics = Enhanced LDBC Datagen

- A battery of graphs covering a rich set of configurations
- Datagen extensions to
  - More diverse degree distributions
  - Clustering coefficient and assortativity

Ongoing work

LDBC D3.3.34 [http://ldbcouncil.org/sites/default/files/LDBC_D3.3.34.pdf](http://ldbcouncil.org/sites/default/files/LDBC_D3.3.34.pdf) and Orri Erling et al. The LDBC Social Network Benchmark: Interactive Workload, SIGMOD'15
Graphalytics = Advanced Monitoring & Logging System

- Automatic analysis matching the programming model

A. Iosup et al., Towards Benchmarking IaaS and PaaS Clouds for Graph Analytics. WBDB 2014
Graphalytics = Choke-Point Analysis

- Choke points are crucial technological challenges that platforms are struggling with

- Examples
  - Network traffic
  - Access locality
  - Skewed execution

- Challenge: Select benchmark workload based on real-world scenarios, but make sure they cover the important choke points

near-future work
Graphalytics = Advanced Software Engineering Process

https://github.com/mihaic/graphalytics/

- All significant modifications to Graphalytics are peer-reviewed by developers
  - Internal release to LDBC partners (Feb 2015)
  - Public release, announced first through LDBC (Apr 2015*)
- Jenkins continuous integration server
- SonarQube software quality analyzer
Graphalytics in Practice

Data ingestion not included here!

- Many more metrics supported
- 10 platforms tested w prototype implementation
- 5 classes of algorithms
- Missing results = failures of the respective systems

6 real-world datasets +
2 synthetic generators
Key Findings So Far

- Performance is function of (Dataset, Algorithm, Platform, Deployment)
  - Previous performance studies lead to tunnel vision

- Platforms have their specific drawbacks (crashes, long execution time, tuning, etc.)
  - Best-performing system depends on stakeholder needs

- Some platforms can scale up reasonably with cluster size (horizontally) or number of cores (vertically)
  - Strong vs weak scaling still a challenge—workload scaling tricky
  - Single-algorithm is not workflow/multi-tenancy

Thank you for your attention!
Comments? Questions? Suggestions?

http://graphalytics.ewi.tudelft.nl
https://github.com/mihaic/graphalytics/

PELGA 2015, May 15
http://sites.google.com/site/pelga2015/

Alexandru Iosup
A.Iosup@tudelft.nl

GRAPHALYTICS was made possible by a generous contribution from Oracle.
A few extra slides
Discussion

- How much preprocessing should we allow in the ETL phase?
- How to choose a metric that captures the preprocessing?

http://graphalytics.ewi.tudelft.nl
Discussion

• How should we assess the correctness of algorithms that produce approximate results?
• Are sampling algorithms acceptable as trade-off time to benchmark vs benchmarking result?

http://graphalytics.ewi.tudelft.nl
Discussion

• How to setup the platforms? Should we allow algorithm-specific platform setups or should we require only one setup to be used for all algorithms?

http://graphalytics.ewi.tudelft.nl
Discussion

- Towards full use cases, full workflows, and inter-operation of big data processing systems
- How to benchmark the entire chain needed to produce useful results, perhaps even the human in the loop?

http://graphalytics.ewi.tudelft.nl