

8th Technical User Community meeting - LDDBC

Peter Boncz

boncz@cwi.nl



ldbcouncil.org

The screenshot shows a web browser window with the address bar displaying 'ldbcouncil.org'. The website header features the LDBC logo (a green hexagon with a white cube inside) and the tagline 'The graph & RDF benchmark reference'. Navigation links include 'BENCHMARKS', 'INDUSTRY', 'PUBLIC', and 'DEVELOPER'. The main content area has a blue background with a white graph pattern and a central 'BENCHMARKS' heading. Below the heading is a paragraph of text and a 'READ MORE' button. At the bottom, there are three green boxes with white text and LDBC logos, each containing a different section title and a list of sub-questions.

ldbcouncil.org

Apps D2.5 Maps

LDBC The graph & RDF benchmark reference

BENCHMARKS » INDUSTRY » PUBLIC » DEVELOPER »

BENCHMARKS

Here you may find the results for different benchmarks, i.e. the Social Network Benchmark (SNB) and the Semantic Publishing Benchmark (SPB), their definitions and best practices, the repositories where to find the data generators and the query implementations, an access to the intranet for the LDBC Industry partners and a list of the LDBC member vendors.

READ MORE

LDBC official benchmarks for industry

- Semantic Publishing Benchmark (SPB)

Why LDBC?

- What are Graph Database systems?
- What are RDF Database systems?
- Why is benchmarking valuable?
- What is the mission of LDBC?

The benchmarking community

- Test the SPB and/or contribute to it
- Test the SNB and/or contribute to it
- Provide Feedback on the Forum

LDBC Organization (non-profit)



“sponsors”



- + non-profit members (FORTH, STI2) & personal members
- + **Task Forces**, volunteers developing benchmarks
- + **TUC**: Technical User Community (8 workshops, ~40 graph and RDF user case studies, 18 vendor presentations)

What does a benchmark consist of?

- Four main elements:
 - *data & schema*: defines the structure of the data
 - *workloads*: defines the set of operations to perform
 - *performance metrics*: used to measure (quantitatively) the performance of the systems
 - *execution & auditing rules*: defined to assure that the results from different executions of the benchmark are valid and comparable
- Software as Open Source (GitHub)
 - data generator, query drivers, validation tools, ...

Audience

- For **developers** facing graph processing tasks
 - recognizable scenario to compare merits of different products and technologies
- For **vendors** of graph database technology
 - checklist of features and performance characteristics
- For **researchers**, both industrial and academic
 - challenges in multiple choke-point areas such as graph query optimization and (distributed) graph analysis

SPB scope

- The scenario involves a media/ publisher organization that maintains semantic metadata about its Journalistic assets (articles, photos, videos, papers, books, etc), also called Creative Works
- The Semantic Publishing Benchmark simulates:
 - Consumption of RDF metadata (Creative Works)
 - Updates of RDF metadata, related to Annotations
- Aims to be an industrially mature RDF database benchmark (SPARQL1.1, some reasoning, text and GIS queries, backup&restore)

LDBC Task Forces

- Semantic Publishing Benchmark Task Force
 - Develops industry-grade RDF benchmark
- Social Network Benchmark Task Force
 - Develops benchmark for graph data management systems
 - Broad coverage: three workloads
- Graph Analytics Task Force
 - Spin-off from the SNB task force
- Graph Query Language Task Force
 - Not strictly about benchmarking
 - Studies features of graph database query languages

Semantic Publishing Benchmark (SPB)



Home | Football | Formula 1 | Cricket | Rugby U | Rugby L | Tennis | Golf | London 2012 | More Sports ▾

Countries ▾ Bulgaria Athletes | Schedule & Results | Medals | Olympic Sports ▾

Share f t


Information on this page will not be updated. Facts were accurate as of August 13, 2012.

Bulgaria

Key Facts


- Top medal sports (pre-2012)
Wrestling
- Capital
Sofia
- Population
7,500,000
- Size
110,994km²
- Languages



Team GB's Campbell secures medal

Luke Campbell is guaranteed an Olympic medal after beating Bulgaria's Detelin Dalakiev in his bantamweight semi-final.

5 Aug 12



Bulgaria beat GB volleyball men

MEN'S VOLLEYBALL
29 Jul 12

Great Britain's men produce a battling display on their Olympic debut but are beaten in straight sets by Bulgaria at Earls Court.

Medal Table

Show me: Countries GO

Rank	Country				Total
1	United States	46	29	29	104
2	China	38	27	23	88
3	Great Britain & N. Ireland	29	17	19	65
63	Bulgaria	0	1	1	2

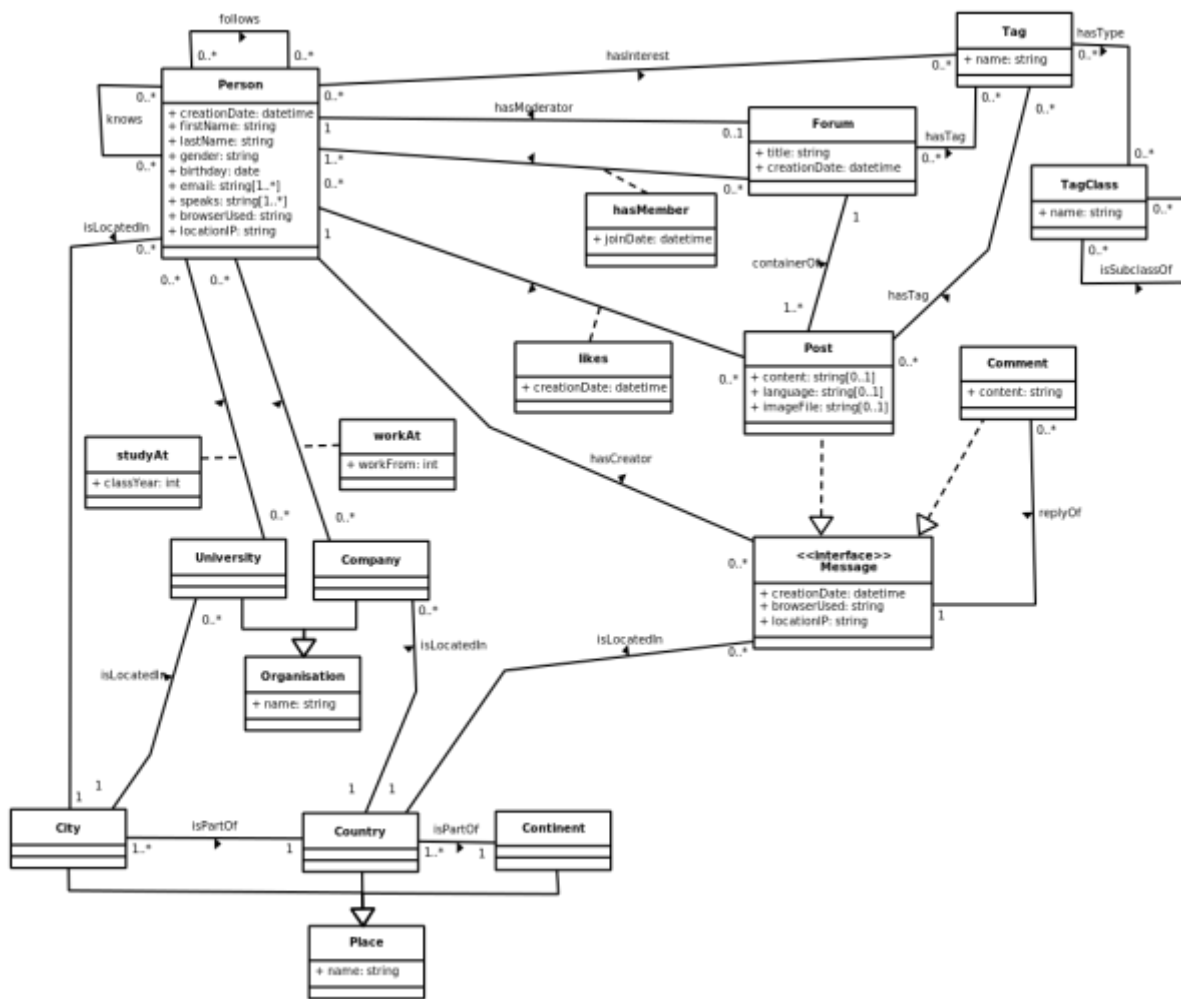
[View full Bulgaria table](#)

Bulgaria Medallists

Bronze
Tsvetelina Puleva
Men's Heavyweight (91kg)

Silver
Stanka Zlateva Hristova
Women's Freestyle 72kg

Social Network Benchmark: schema



Benchmark Workloads

- **Interactive:** tests throughput running short queries while consistently handling concurrent updates
 - *Show all photos posted by my friends that I was tagged in*
- **Business Intelligence:** consists of complex structured queries for analyzing online behavior
 - *Influential people the topic of open source development?*
- **Graph Analytics:** tests the functionality and scalability on most of the data as a single operation
 - *PageRank, Shortest Path(s), Community Detection*

Systems

- **Graph database systems**
 - e.g. Neo4j, InfiniteGraph, DEX, Titan
- **Graph programming frameworks**
 - e.g. Giraph, Signal/Collect, Graphlab, Green Marl, Grappa
- **RDF database systems**
 - e.g. OWLIM, Virtuoso, BigData, Jena TDB, Stardog, Allegrograph
- **Relational database systems**
 - e.g. Postgres, MySQL, Oracle, DB2, SQLServer, Virtuoso, MonetDB, Vectorwise, Vertica
- **noSQL database systems**
 - e.g. HBase, REDIS, MongoDB, CouchDB, or even MapReduce systems like Hadoop and Pig

Workloads by system

System	Interactive	Business Intelligence	Graph Analytics
Graph databases	Yes	Yes	Maybe
Graph programming frameworks	-	Yes	Yes
RDF databases	Yes	Yes	-
Relational databases	Yes	Yes	Maybe, by keeping state in temporary tables, and using the functional features of PL-SQL
NoSQL Key-value	Maybe	Maybe	-
NoSQL MapReduce	-	Maybe	Yes

Interactive (On-line) Workload

- Test online ACID features and scalability
- The system under test is expected to run in a steady state, providing durable storage
- Updates are typically small
- Updates will conflict a small percentage of the time
- Queries typically touch a small fraction of the database

SIGMOD 2015 paper

The LDBC Social Network Benchmark: Interactive Workload

Orri Erling
OpenLink Software, UK
oerling@openlinksw.com

Alex Averbuch
Neo Technology, Sweden
alex.averbuch@
neotechnology.com

Josep Larriba-Pey
Sparsity Technologies, Spain
larri@sparsity-
technologies.com

Hassan Chafi
Oracle Labs, USA
hassan.chafi@oracle.com

Andrey Gubichev
TU Munich, Germany
gubichev@in.tum.de

Arnau Prat^{*}
Universitat Politècnica de
Catalunya, Spain
aprat@ac.upc.edu

Minh-Duc Pham
VU University Amsterdam,
The Netherlands
m.d.pham@vu.nl

Peter Boncz
CWI, Amsterdam, The
Netherlands
boncz@cwi.nl

ABSTRACT

The Linked Data Benchmark Council (LDBC) is now two years underway and has gathered strong industrial participation for its mission to establish benchmarks, and benchmarking practices for evaluating graph data management systems. The LDBC introduced a new *choke-point* driven methodology for developing benchmark workloads, which combines user input with input from expert systems architects, which we outline. This paper describes the LDBC Social Network Benchmark (SNB), and presents database benchmarking innovation in terms of graph query function-

tional table, for instance as a table where every row contains an edge, and the start and end vertex of every edge are a foreign key reference (in SQL terms). However, what makes a data management problem a graph problem is that the data analysis is not only about the values of the data items in such a table, but about the *connection patterns* between the various pieces. SQL-based systems were not originally designed for this – though systems have implemented diverse extensions for navigational and recursive query execution.

In recent years, the database industry has seen a proliferation of new graph-oriented data management technologies.

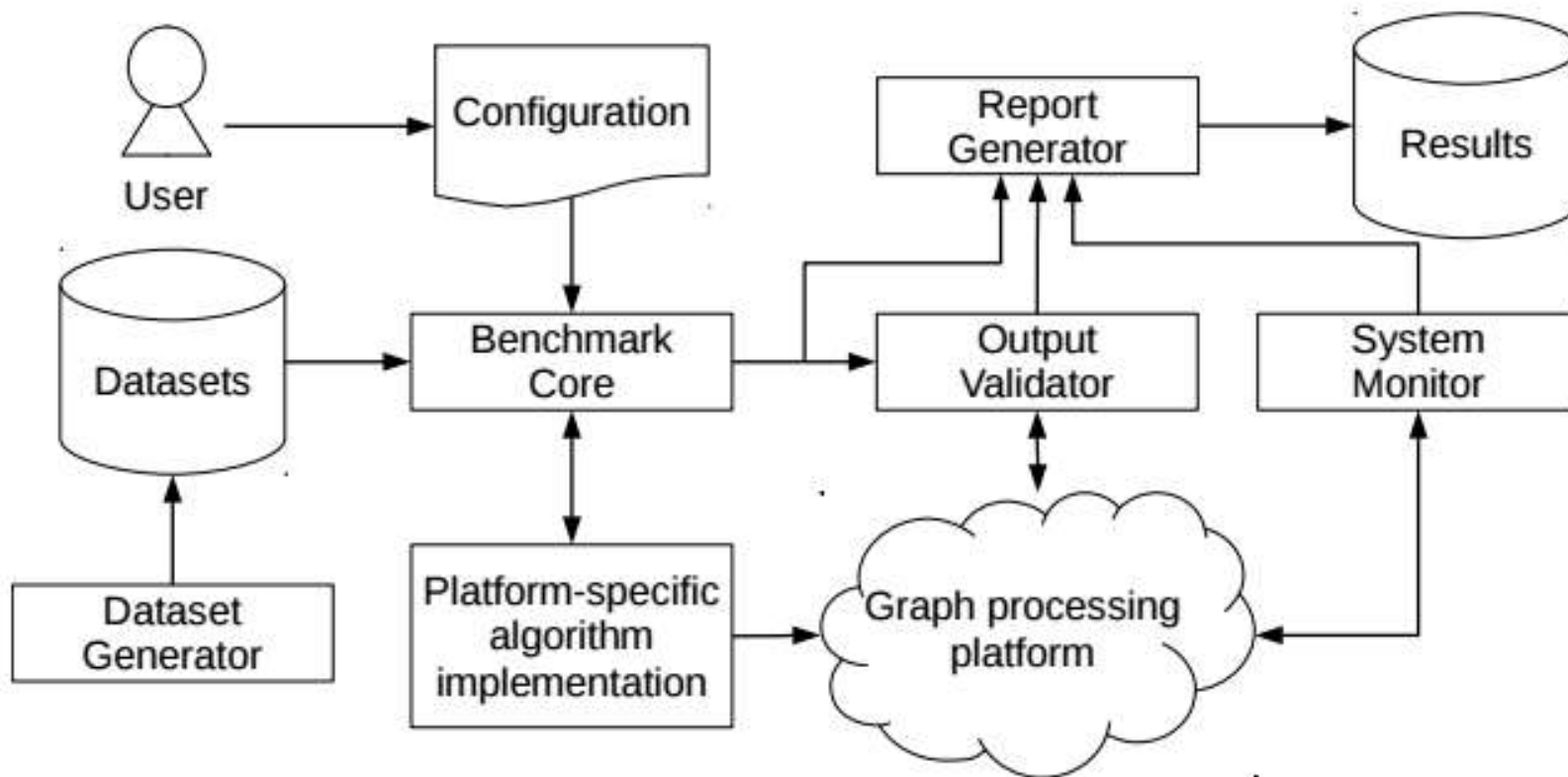
Business Intelligence Workload

- The workload stresses query execution and optimization
- Queries typically touch a large fraction of the data
- The queries are concurrent with trickle load
- The queries touch more data as the database grows

Graph Analytics Workload (Graphalytics)

- The analytics is done on most of the data in the graph as a single operation
- The analysis itself produces large intermediate results
- The analysis transactional: no need for isolation from possible concurrent updates

Graphalytics Architecture



Graphalytics Algorithms

- general statistics (STATS)
 - counts the numbers of vertices and edges in the graph and computes the mean local clustering coefficients
- breadth-first search (BFS)
 - traverses the graph starting from a seed vertex, visiting first all the neighbors of a vertex before moving to the neighbors of the neighbors.
- connected components (CONN) algorithm
 - determines for each vertex the connected component it belongs to.
- community detection (CD) algorithm
 - detects groups of nodes that are connected to each other stronger than they are connected to the rest of the graph
- graph evolution (EVO)
 - predicts the evolution of the graph according to the “forest fire” model

VLDB2016 paper

LDBC Graphalytics: A Benchmark for Large-Scale Graph Analysis on Parallel and Distributed Platforms

Alexandru Iosup[△] Tim Hegeman[△] Wing Lung Ngai[△] Stijn Heldens[△] Arnau Prat Pérez[□]
Thomas Manhardt[◇] Hassan Chafi[◇] Mihai Capotă[◇] Narayanan Sundaram[◇]
Michael Anderson[◇] Ilie Gabriel Tănase^ψ Yinglong Xia[⊕] Lifeng Nai[⊕] Peter Boncz[▽]

[◇]Oracle Labs [◇]Intel Labs ^ψIBM Research [⊕]Huawei Research America
[△]Delft University of Technology [□]UPC Barcelona [⊕]Georgia Tech [▽]CWI Amsterdam

ABSTRACT

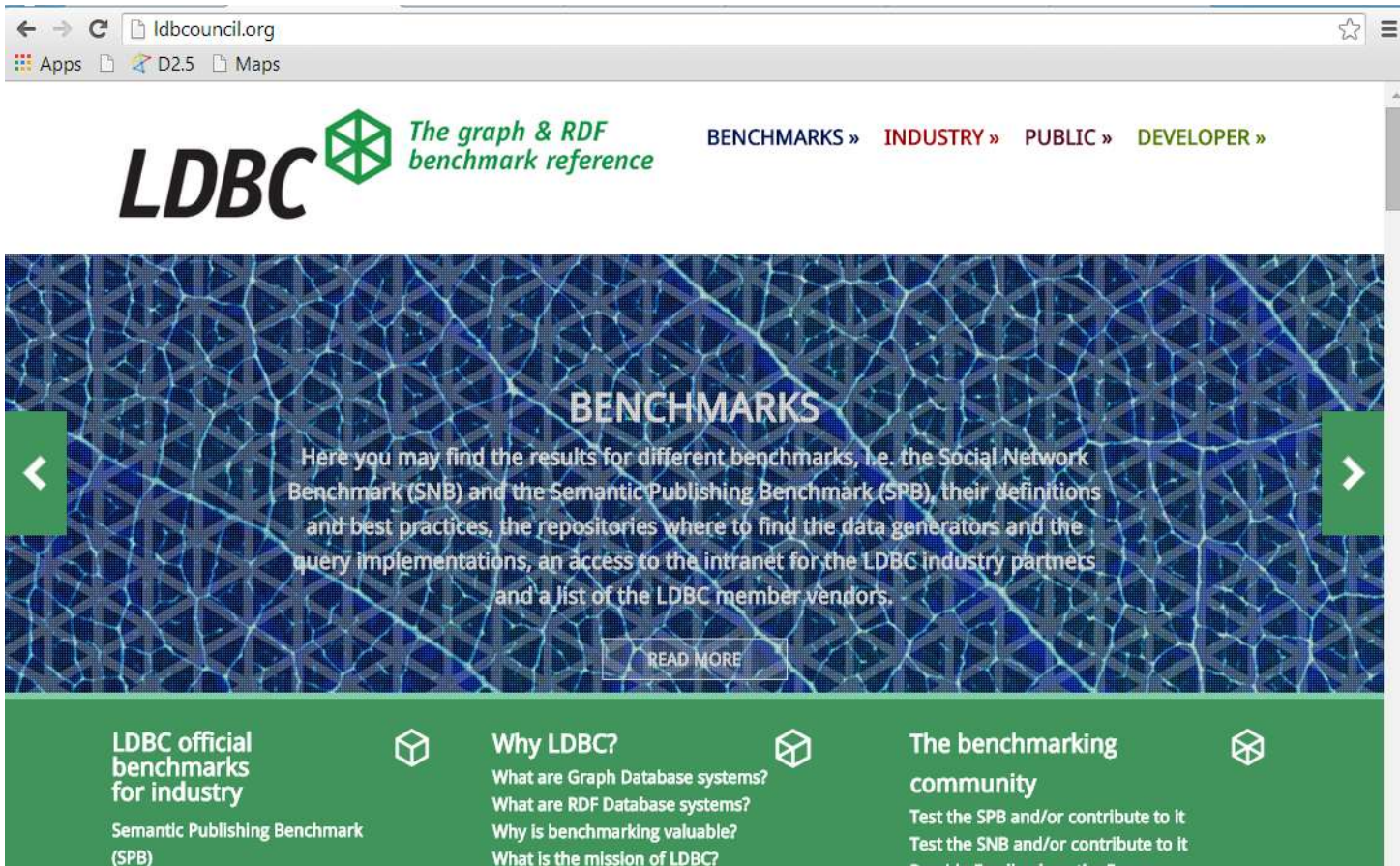
In this paper we introduce LDBC Graphalytics, a new industrial-grade benchmark for graph analysis platforms. It consists of six deterministic algorithms, standard datasets, synthetic dataset generators, and reference output, that enable the objective comparison of graph analysis platforms. Its test harness produces deep metrics that quantify multiple kinds of system scalability, such as horizontal/vertical and weak/strong, and of robustness, such as failures and performance variability. The benchmark comes with open-source software for generating data and monitoring performance. We describe and analyze six implementations of the benchmark (three from the community, three from the industry), providing insights into the strengths and weaknesses of the platforms. Key to our contribution, vendors perform the tuning and benchmarking of their platforms.

sionals on the properties of the various solutions available on the market; to stimulate academic research in graph data storage, indexing, and analysis; and to accelerate the maturing process of the graph data management space as a whole. LDBC organizes a Technical User Community (TUC) that gathers benchmark input and feedback, and as such has investigated graph data management use cases across the fields of marketing, sales, telecommunication, production, publishing, law enforcement and bio-informatics. LDBC previously introduced the Social Network Benchmark [13] (SNB), which models a large social network but targets database systems (graph, SQL or SPARQL) that provide interactive updates and query answers. However, the LDBC scope goes beyond such database workloads: it also includes graph analysis frameworks that facilitate complex and holistic graph computations which may not be easily modeled as database queries, but rather as (iterative) graph algorithms, such as global metrics (e.g. diameter, triangle count) or

More Information


<http://www.ldbcouncil.org>

<http://github.com/ldbc>



← → ↻ ldbcouncil.org ☆ ☰

Apps D2.5 Maps


LDBC  *The graph & RDF benchmark reference*

BENCHMARKS » INDUSTRY » PUBLIC » DEVELOPER »


BENCHMARKS

Here you may find the results for different benchmarks, i.e. the Social Network Benchmark (SNB) and the Semantic Publishing Benchmark (SPB), their definitions and best practices, the repositories where to find the data generators and the query implementations, an access to the intranet for the LDBC industry partners and a list of the LDBC member vendors.


READ MORE

LDBC official benchmarks for industry 

Semantic Publishing Benchmark (SPB)

Why LDBC? 

What are Graph Database systems?
What are RDF Database systems?
Why is benchmarking valuable?
What is the mission of LDBC?

The benchmarking community 

Test the SPB and/or contribute to it
Test the SNB and/or contribute to it
Provide Feedback on the Forum

Blogs
Specifications
Early Result FDRs
Videos of TUC talks
Developer info
Code, Issue Tracking

Graph Query Language Task Force

- Renzo Angles, Universidad de Talca
- **Marcelo Arenas, PUC Chile - task force lead**
- Pablo Barceló, Universidad de Chile
- Peter Boncz, Vrije Universiteit Amsterdam
- George Fletcher, Eindhoven University of Technology
- Irimi Fundulaki, Foundation for Research and Technology - Hellas (FORTH)
- Claudio Gutierrez, Universidad de Chile
- Tobias Lindaaker, Neo Technology
- Marcus Paradies, SAP
- Raquel Pau, UPC
- Arnau Prat, UPC
- Tomer Sagi, HP Labs
- Oskar van Rest, Oracle Labs
- Hannes Voigt, TU Dresden
- Yinglong Xia, Huawei Research merica

GRADES Invitation (Friday)

- Keynote by Jure Leskovec (Stanford + Pinterest)
 - Known for creating the SNAP dataset collection
- First paper on Graph Frames
 - Spark RDD extension for graph data
 - PGQL: Oracle Graph Query Language
 - And more..