LDBC Social Network Benchmark

Interactive Workload

Arnau Prat
DAMA – UPC

*Sparsity

Summary of SNB-Interactive

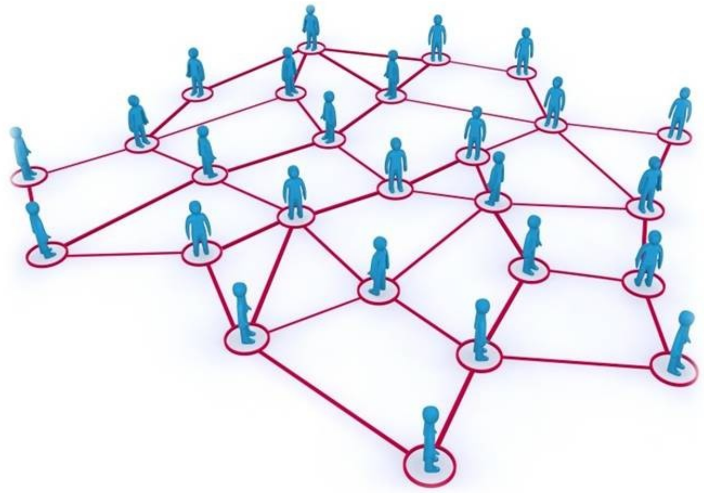
- Simple but challenging interactive queries on top of a social network site
  - Interactive queries
  - Flexible: Declarative and API based systems
  - Latency and throughput are both important
  - Easy to use
- All software and docs at https://github.com/ldbc

Table of Contents

- LDBC SNB Datage
- LDBC SNB Interactive Queries
- LDBC Workload Driver
- Conclusions

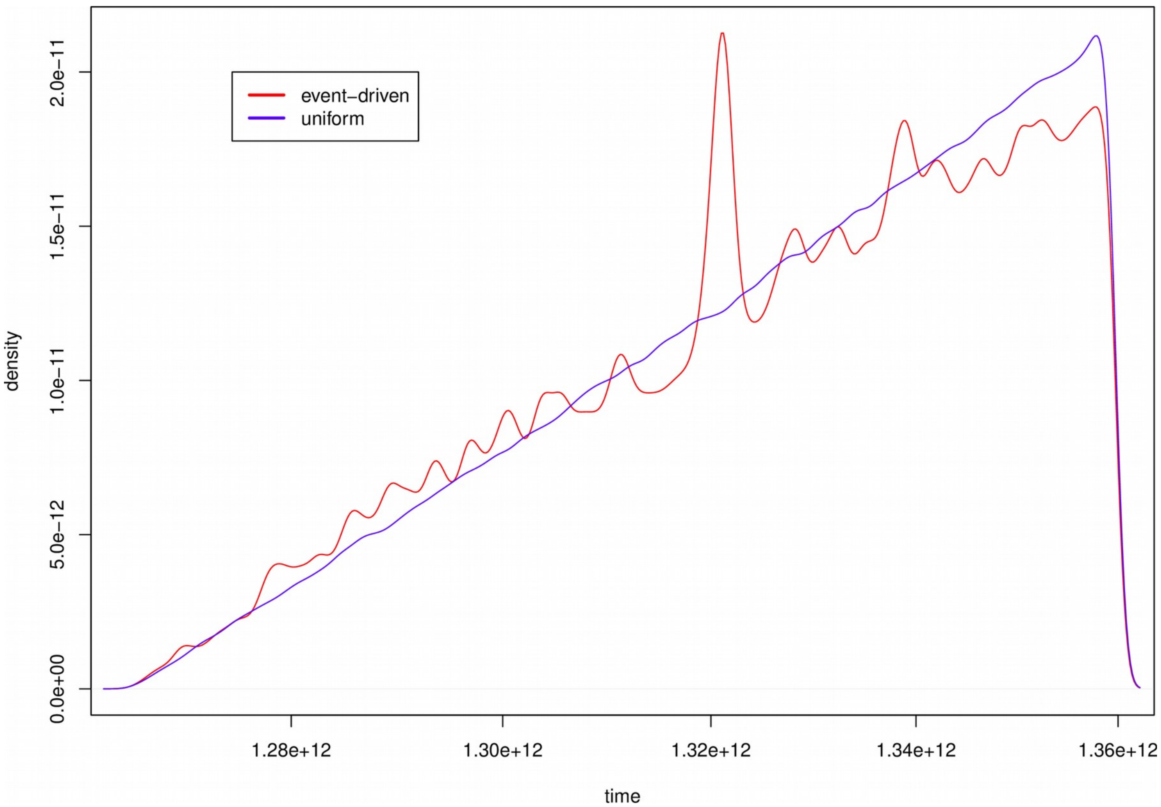
LDBC SNB Datagen

- Generates a realistic social network with the Facebook degree distribution (persons, groups, posts, likes, etc.)
  - Correlated graph → Similar people have a larger probability to connected, correlated attributes, etc.



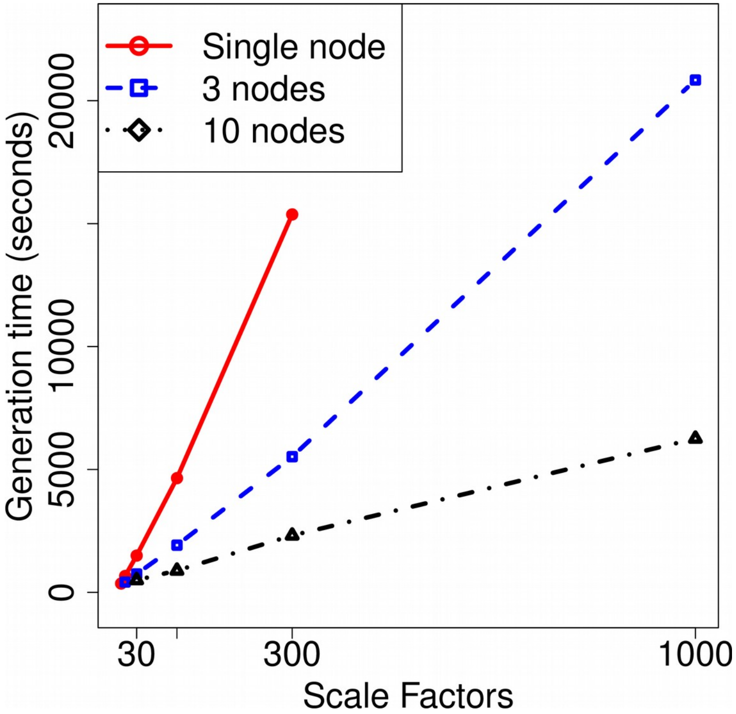
LDBC SNB Datagen

- Generates a realistic social network with the Facebook degree distribution (persons, groups, posts, likes, etc.)
  - Correlated graph → Similar people have a larger probability to connected, correlated attributes, etc.
  - Event driven activity volume



LDBC SNB Datagen

- Generates a realistic social network with the Facebook degree distribution (persons, groups, posts, likes, etc.)
  - Correlated graph → Similar people have a larger probability to connected, correlated attributes, etc.
  - Event driven activity volume
  - Scalable



LDBC SNB Datagen

- Generates a realistic social network with the Facebook degree distribution (persons, groups, posts, likes, etc.)
  - Correlated graph → Similar people have a larger probability to connected, correlated attributes, etc.
  - Event driven activity volume
  - Scalable
  - Deterministic → Allows a fair comparison between SUTs and reproducibility of benchmark executions

LDBC SNB Datagen

- Scale Factors
  - 1,3,10,30,100,300,1000
  - Based on the size of the dataset on dist in CSV format

SF	Relations	Persons	Messages	Activity	Size
SF1	20M	11K	3M	3 years	1GB
SF10	200M	73K	30M	3 years	10GB
SF100	2000M	499K	300M	3 years	100GB
SF1000	20000M	3600K	3000M	3 years	1000GB

* approximated numbers

LDBC SNB Datagen

- 90% of the network is output as CSV to be bulk loaded
- The rest 10% is output as update streams
- Substitution parameters for each complex read query type
  - Parameter binding to reduce variability between queries

LDBC SNB Interactive queries

- 14 Complex reads (interactive yet complex, target choke-points):
  - Query 6: Given a **start Person** and some Tag, find the other Tags that occur together with this Tag on Posts that were created by start Person’s friends and friends of friends
  - Query 14: Given **two Persons**, find all (unweighted) shortest paths between these two Persons, in the subgraph induced by the Knows relationship. Then, for each path calculate a weight. The nodes in the path are Persons, and the weight of a path is the sum of weights between every pair of consecutive Person nodes in the path. The weight for a pair of Persons is calculated such that every reply (by one of the Persons) to a Post (by the other Person) contributes 1.0, and every reply (by ones of the Persons) to a Comment (by the other Person) contributes 0.5.

LDBC SNB Interactive queries

- 7 Short reads (balance read/write ratio of workload. mimic user behavior):
  - Given a start Person, retrieve their first name, last name, birthday, IP address, browser, and city of residence
  - Given a start Person, retrieve all of their friends, and the date at which they became friends
  - Given a Message (Post or Comment), retrieve the (1-hop) Comments that reply to it. In addition, return a boolean flag indicating if the author of the reply knows the author of the original message. If author is same as original author, return false for "knows" flag

LDBC SNB Interactive queries

- 8 Updates:
  - Add Person
  - Add Knows
  - Add Post
  - Add Post Like
  - Add Comment
  - Add Comment Like
  - Add Group
  - Add Group Membership

LDBC Workload Driver

- Responsible of generating the Workload = Stream of operations
  - scheduled start time (real time)
  - type (e.g. ComplexQuery1)
  - parameters (e.g. Person ID)

LDBC Workload Driver

- Updates
  - substitution parameters read from datagen update streams
  - time stamps ("simulation time") read from datagen update streams

LDBC Workload Driver

- Complex Reads
  - substitution parameters read from datagen files
  - scheduled start times assigned by driver as multiples of update frequency
    - e.g. for every 132 Updates the driver generates 1 ComplexQuery1

LDBC Workload Driver

- two groups of Short Reads: "person centric" & "message centric"
- after each Complex Read a sequence of Short Reads is executed
  - sequence appoximates walk through network
  - at each step there is a probability of taking another step, which decreases at each step
  - steps consist of either all "person centric" or all "message centric" operations
    - e.g., (person centric operations)->(flip coin)->(message centric operations)->(flip coin)...
  - mimics user "following links"/Facebook-stalking :-)
  - substitution parameters read taken from results of recent Complex Reads and Short reads

LDBC Workload Driver - Execution

- Driver schedules operations as close to their scheduled start times as possible
- "Time Compression Ratio" used to configure target throughput
- Vendor provides callbacks that driver use to execute operations
- Number of worker threads configurable
- For every executed operation, driver logs the following (used for auditing)
  - operation type
  - scheduled start time
  - actual start time
  - runtime

LDBC Workload Driver – Validation Mode

- Given a vendor implementation & workload, driver generates validation datasets
- Stream of operations + their results
- Validation datasets can then be used to validate other vendor implementations (e.g. compare results)
- Official validation datasets are provided by the LDBC SNB

LDBC Workload Driver - Example

- SF10, tcr 0.5, 1000k

Query	Count
Q1	36
Q2	25
Q3	10
Q4	26
Q5	14
Q6	4
Q7	16

Query	Count
Q8	63
Q9	3
Q10	27
Q11	49
Q12	21
Q13	49
Q14	19

Query	Count
S1	451
S2	451
S3	451
S4	444
S5	444
S6	444
S7	444

Query	Count
U1	1
U2	124
U3	153
U4	3

Query	Count
U5	244
U6	18
U7	74
U8	22

LDBC Workload Driver - Example

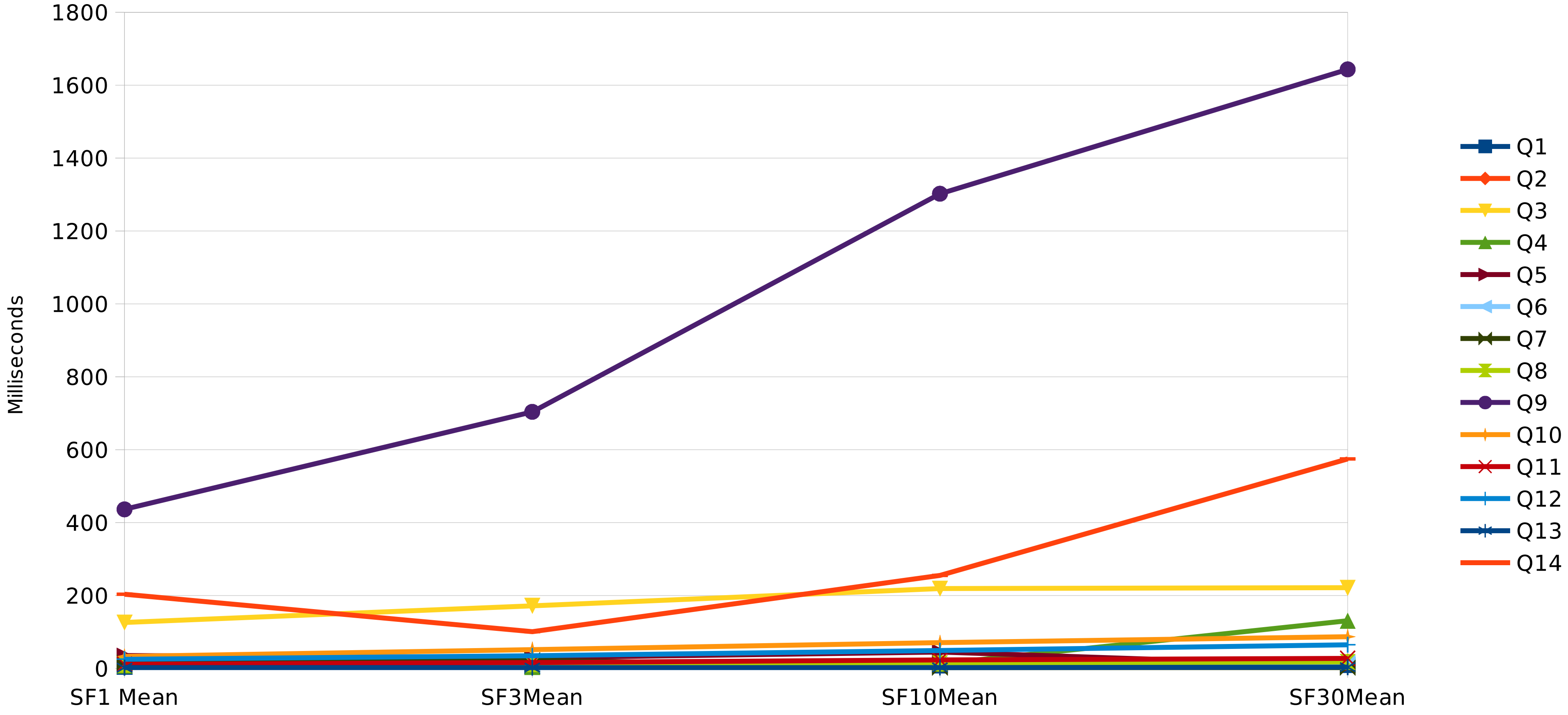
- **Query mix for SF10**

Query	Frequency
Q1	26
Q2	37
Q3	106
Q4	36
Q5	72
Q6	316
Q7	48
Q8	9
Q9	384
Q10	37
Q11	20
Q12	44
Q13	19
Q14	49

- **Query mix for SF300**

Query	Frequency
Q1	26
Q2	37
Q3	142
Q4	46
Q5	84
Q6	580
Q7	32
Q8	3
Q9	705
Q10	44
Q11	24
Q12	44
Q13	19
Q14	49

LDBC Workload Driver - Example



LDBC Workload Driver - Rules

- Benchmark executions must meet the following rules to be valid:
  - queries must pass validation datasets
  - at most 5% of the queries actual start time can be one second greater than scheduled start time
  - must comprise at least 2 hours of simulation time
  - at any point, the test machine is disconnected and those commited must be persistent
- Performance metrics are:
  - latencies for each query
  - throughput

Conclusions

- SNB Interactive on top of synthetic Social Network data
- 3 Types of queries:
  - Complex Reads
  - Short Reads
  - Updates
- The driver builds a query wich mimics a user behavior
- Both latency and throughput are important. Persistence is mandatory
- All software is open source. We are open for contributions!

Thank you