eXtra Large Joins

Ali Mohammadi Shanghooshabad
Department of Computer Science
University of Warwick
Coventry, UK

EPSRC
Engineering and Physical Sciences Research Council

Skip the full join generation

XL Joins Challenges

- Running time for big tables
- Storage cost for Huge join result
- inefficient analytics over the huge join result
- Money cost in clouds

Random join sampling problem

T1

A	B	C
a1	b1	c1
a1	b2	c1
a2	b2	c2
a3	b2	c2

$\bowtie_C$

T2

C	D	E
c1	d1	e1
c2	d1	e1
c2	d1	e2
c3	d2	e2

Join Result

A	B	C	D	E	Prob
a1	b1	c1	d1	e1	1/6
a1	b2	c1	d1	e1	1/6
a2	b2	c2	d1	e1	1/6
a2	b2	c2	d1	e2	1/6
a3	b2	c2	d1	e1	1/6
a3	b2	c2	d1	e2	1/6

Uniform Sample

A	B	C	D	E
a1	b2	c1	d1	e1
a2	b2	c2	d1	e1
a2	b2	c2	d1	e2

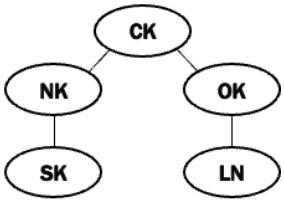
- **Knowledge Discovery**
  - Classification over the join sample
  - Clustering over the sample
  - Regression over the sample
  - etc.
- **Model Join**
  - Join the existing models
  - Build the models over exisitng data
- **AQP**
  - Approximate query processing on a sample of the join result

PGMs as core idea

Schema and an example join query

nation (nationkey NK)
supplier (nationkey NK, suppkey SK)
customer (custkey CK, nationkey NK)
orders (orderkey OK, custkey CK)
lineitem (orderkey OK, linenumber LN)

```
SELECT n.nationkey as NK, s.suppkey as SK, c.custkey as CK,
       o.orderkey as OK, l.linenumber as LN
FROM   nation n, supplier s, customer c, orders o, lineitem l
WHERE  n.nationkey = s.nationkey
  AND s.nationkey = c.nationkey
  AND c.custkey = o.custkey
  AND o.orderkey = l.orderkey;
```



$$P(x) = \frac{1}{Z}\prod_{c\in C}\psi_c(x_c)$$

$$Z = \sum_x \prod_{c\in C}\psi_c(x_c)$$

Sum-Product Message Passing Alg.

Ancestral sampling

Uniform Sample

CK	NK	OK	SK	LN
a1	b2	c1	d1	e1
a2	b2	c2	d1	e1
a2	b2	c2	d1	e2

Model Join

The result of a *model join* should be "similar" to the result we would have obtained if we joined the underlying tables.

Why data is absent sometimes?

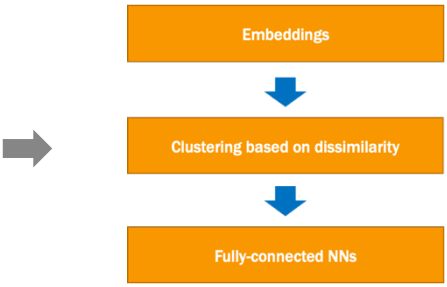
- Storage problem
- Privacy preserving
- Accurate, light and fast in-memory models (DBEst, DeepDB etc.)

Solution?

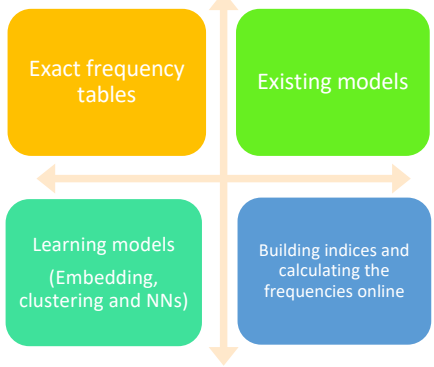
- The idea is that any edge of the PGM graph could be a model providing needed information.

Learning Challenges in tabular data

- Higher Number of distinct values
- Categorical attributes

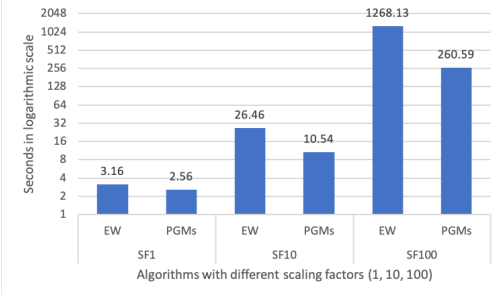


Strategies in the edges of the PGMs



Preliminary results

5X faster on QX SF=100 to generate a sample of one million



EW algorithm is the SOTA approach introduced in Zhao et al, SIGMOD 2018