UNIVERSITY of HOUSTON

Vector-Matrix Multiplication to Solve Graph Problems with SQL

Wellington M. Cabrera
Advisor: C.Ordonez

MOTIVATION

- Graph Problems are among the toughest problems in data analytics.
- Some problems are really **BIG** (social networks, transportation networks, WWW).
- Special graph databases /systems (as Giraph) are required to analyze big graphs (Sure???)

Why Relational Databases?

- Large amount of data stored in Relational Databases.
- Relationships between people, documents, places can be modeled as graphs.
- Database Management systems has been studied and optimized for years.

The challenge: Compute graph algorithms with regular SQL Queries

Bellman-Ford
Reachability
Topological Sort
Page Rank

These algorithms can be solved as a sequence of vector-matrix multiplications

DEFINITIONS

Let $G = (V, E)$, E is stored in a table $E(i,j,v)$. Note:
a) E is equivalent to the adjacency matrix of E, omitting zeroes
b) weights/distances are represented by v
c) If $O(V) = O(E)$, we say that E is sparse
Let S a vector of graph vertices, S is stored in a table $S(j,v)$. Note:
a) v can represent distance, reachability, order, probability
b) We omit values with no information
(like ∞ for distances, 0 for probabilities)

S

j	v
1	1/7
2	1/7
3	1/7
4	1/7
5	1/7
6	1/7
7	1/7

$\times$

E

i	j	v
1	2	2
2	3	3
2	5	2
3	1	2
3	4	2
4	2	3
4	6	1
5	1	2
5	4	1
5	7	3
6	3	2

Algorithms

Bellman-Ford:	Shortest path from a single source s.
Reachability:	Set of all vertices reachable from a single source s.
Topological Sort:	Linear order of vertices, keeping precedence.
PageRank:	Probability to reach each vertex, after a random walk.

ALGORITHM

The four algorithms follow the following pattern:
1) Initialize S_0 (depending on problem)
2) While $\Delta < \varepsilon$
compute $S_i = S_{i-1} \times E$, i++, compute Δ
3) Return S_k (or $\sum S_i$)

Vector-matrix multiplication under some semiring

$$c_j = \bigoplus_{k=1} (v_k \otimes a_{k,j})$$

Algorithm	$\oplus$	$\otimes$	Meaning of S.v	$\lvert S_0 \rvert$	Out put	Time Complexity
Bellman-Ford	min	+	distance	1	S_k	$O(knlog(n))$
Reachability	$\sum$	x	Reachability	1	$\sum S_i$	$O(kn), O(knlog(n))$
Topological Sort	$\wedge$	$\Rightarrow$	Order	|{deg(i)=0}|	$\sum S_i$	$O(knlog(n))$
PageRank	$\sum$	×	Probability	b	S_k	$O(knlog(n))$

ONE ALGORITHM FOR FOUR PROBLEMS

Input: E, S_0, s, $\oplus$, $\otimes$, ε ,f, *sumFlag*
Output: R_d

$R_0 \leftarrow \sigma_{sumFlag}(S_0)$;
$d = 0$;
$\Delta = 1$;
while $\Delta > \epsilon$ **do**
  $d = d + 1$;
  $S_d \leftarrow \pi_{i:\oplus(E.v \otimes S.v)}(E \bowtie_{j=j} S_{d-1})$;
  if *sumFlag* **then**
    $R_d \leftarrow \pi_{j:sum(v)}(S_d \cup R_{d-1})$;
  else
    $R_d = S_d$;
  end
  $\Delta = f(R_d, R_{d-1}, S_d, S_{d-1})$;
end

Vector-matrix multiplication with SQL

```
Select i, ⊕agg ( S.v ⊗ E.v)
  from S_{d-1} as S join E on S.j=E.j
  group by i
```

RESEARCH QUESTIONS

Can this algorithm solve more graph problems ?
Can SQL defeat Giraph or Spark GraphX in same hardware?