Network-Aware Workload Scheduling for Scalable Linked Data Stream Processing

**Problem**

**Query**

```query
REGISTER QUERY PublicSpend AS
SELECT { ?company name ?agency name ?contract id ?min price ?max price ?factor }
FROM STREAM crsp/ticker/trdf [RANGE 20 DAY STEP 1 DAY]
FROM STREAM usaspending.org/contracts/trdf [RANGE 20 DAY STEP 1 DAY]
WHERE { 
  GRAPH < crsp/ticker/trdf > {
    ?ticker id wC:P R C ?ticker price ;
    wC:COMNAM ?company name ;
    wC:TICKER ?ticker symbol .
  }
  UNION GRAPH < usaspending.org/contracts/trdf > {
    ?contract id us:agencyid ?agency name ;
    us:obligatedamount ?contract amount ;
    us:vendorname ?company name .
  }
}
AGGREGATE { (?min price , MIN , {?ticker price}) }
AGGREGATE { (?max price , MAX , {?ticker price}) }
BIND (?max price / ?min price AS ?factor)
FILTER(?factor > 2)
```

**Approach**

1: Record the “Communication Graph”
2: Apply a Graph Partitioning Algorithm
3: Use Partitioning for Scheduling
4: Repeat

**Results**

SRBench Q3

Network Messages (%) In

<table>
<thead>
<tr>
<th>R</th>
<th>A</th>
<th>F</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRBench Q3 Aug 8-13, 2004</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OpenGov 2001

Network Messages (%) Out

<table>
<thead>
<tr>
<th>T</th>
<th>A</th>
<th>B</th>
<th>F</th>
<th>J</th>
<th>O</th>
</tr>
</thead>
</table>

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec