<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>PASCO: Parallel SimRank Computation at Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Li, Z; Fang, Y; Liu, Q; Cheng, J; Cheng, RCK; Lui, JCS</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>The 2015 ACM Symposium on Cloud Computing (SoCC 2015), Kohala Coast, HI., 27-29 August 2015.</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2015</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/214756">http://hdl.handle.net/10722/214756</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Creative Commons: Attribution 3.0 Hong Kong License</td>
</tr>
</tbody>
</table>
SimRank [1]

- Graph data grows rapidly
  1. Internet of Things
  2. World Wide Web

- Similarity is fundamental
  1. Information retrieval
  2. Recommender system
  3. Churn prediction

- SimRank - two objects are similar if referenced by similar objects

\[
s(i, j) = \frac{c}{m(i) m(j)} \sum_{k \in \text{in}(i), l \in \text{in}(j)} s(k, l)
\]

Similarity Propagation

- To compute \(a_w\), we obtain \(P^e\), using Monte Carlo Simulation
  1. Place \(R\) random walkers on node \(i\)
  2. Each walker walks \(t\) steps along in-links
  3. Count the distribution of walkers

- Online queries
  + MCSP: Monte Carlo simulation for single-query
    - constant time complexity: \(O(TR)\)
  + MCSS: Monte Carlo simulation for single-source query
    - constant time complexity: \(O(T^2R \log d)\)
  + MCAP: Monte Carlo simulation for all-query
    - use MCSS repeatedly; time complexity: \(O(nT^2R \log d)\)

Implementation on Spark

- Why Spark?
  1. General-purpose in-memory cluster computing
  2. Easy-to-use operations for distributed applications

Two implementation models
- Broadcasting: Graph stored in each machine
- RDD (Resilient Distributed Dataset): Graph stored in an RDD

Experiments

- Setup: cluster, datasets, and default parameters - 10 nodes (each with 16 cores, 377GB RAM, 20TB disk)
- CloudWalker – Big SimRank, instant response
- Contribution
  1. Enable parallel SimRank computation
  2. Test on the largest graph, clue-web(|V|=1B, |E|=43B)

- Problem
  1. SimRank Decomposition \(S = cP^D + D\)
  2. Transition matrix on graph
  3. Diagonal correction matrix to be estimated
  \(S = D + cP^D + c^2P^2D + \ldots\)
  1. how to compute \(D\) for big graph?
  2. how to query efficiently given \(D\)?

- Offline indexing \(x = \{D_{11}, D_{22}, \ldots, D_{nn}\}\)
- Key observation: self-similarity is 1.0
- Indexing linear system \(a_x x = \{1, 2, \ldots, n\}\)
  1. Given \(a_{x_{-1}}\), \(x_{-1}\)
  2. Generate \(a_x x\) by Monte Carlo simulation, in parallel
  3. Solve the linear system via Jacobi method, in parallel

- CloudWalker outperforms state of the art

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Nodes</th>
<th>Edges</th>
<th>Size</th>
<th>Parameter</th>
<th>Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wiki-vote</td>
<td>7</td>
<td>103K</td>
<td>476.8K</td>
<td>c</td>
<td>0.6</td>
<td>decay factor</td>
</tr>
<tr>
<td>wiki-talk</td>
<td>2.4M</td>
<td>45.6M</td>
<td></td>
<td>T</td>
<td>10</td>
<td># of walk steps</td>
</tr>
<tr>
<td>twitter-2010</td>
<td>42M</td>
<td>11.4G</td>
<td></td>
<td>L</td>
<td>3</td>
<td># of iterations</td>
</tr>
<tr>
<td>uk-union</td>
<td>131M</td>
<td>48.3G</td>
<td></td>
<td>R</td>
<td>100</td>
<td># of walkers in simulating (a)</td>
</tr>
<tr>
<td>clue-web</td>
<td>18</td>
<td>40.1G</td>
<td></td>
<td>P</td>
<td>50,000</td>
<td># of walkers in MCSP and MCSS</td>
</tr>
</tbody>
</table>

CloudWalker outperforms state of the art

- Preprocessing, single-query and single-source queries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>wiki-vote</td>
<td>43.4s</td>
<td>42.5s</td>
<td>42ms</td>
</tr>
<tr>
<td>wiki-talk</td>
<td>N/A</td>
<td>N/A</td>
<td>18ms</td>
</tr>
<tr>
<td>twitter-2010</td>
<td>-</td>
<td>-</td>
<td>14376s</td>
</tr>
<tr>
<td>uk-union</td>
<td>-</td>
<td>-</td>
<td>8291s</td>
</tr>
<tr>
<td>clue-web</td>
<td>-</td>
<td>-</td>
<td>235m</td>
</tr>
</tbody>
</table>

10x larger than the largest graph reported on SimRank

Effectiveness: CloudWalker converges quickly

Broadcasting is more efficient, but RDD is more scalable

CloudWalker outperforms state of the art

2. D. Fogaras and B. Racz. Scaling link-based similarity search. WWW’05.