Validity: Motivation

<table>
<thead>
<tr>
<th>Two Major Approaches to Process Aggregate Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Wandering Approach</strong></td>
</tr>
<tr>
<td><strong>Capacity</strong></td>
</tr>
<tr>
<td><strong>Pros</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Validity: Definition

**Example:**

- Max: 3 b: 5 c: 1 d: 7 e: 6
- H(a) = {a, b}
- H(a) = {a, b, c, d, e}
- H(C) = {a, c}

What are the values guaranteed by three types of validity?

- **Snapshot validity:** 5
- **Interval validity:** {5, 6, 7}
- **Single-site validity:** {3, 5, 6, 7}
Single-site Validity: WILDFIRE Protocol

Gossip/Epidemic based algorithm: broadcast and convergecast

Basic procedure

- Host H(q) broadcast the aggregate query and an overestimate of the network diameter to its neighbors
- Hosts get the message, change states to active and send partial aggregates to its neighbors. Also, if there is any new change in partial aggregates, the host send the new partial result again to its neighbors

Theorem: WILDFIRE ensures single-site validity for min and max operations

Example

```
    1   2   3   4
   / \ / \ / \ /
  5   6  7   8
```

Validity: Evaluation of Price

Evaluation (Comparison with Spanning Trees and Directed Acyclic Graph)

- **SPANNINGTREE**: hosts are organized into a spanning tree with root at hq. In the convergecast phase, attributes are selected on hosts and partial aggregate are propagated from leaf hosts to the parent nodes.
- **DIRECTEDACYCLICGRAPH**: provide each host with up to k parents and organize the hosts in a directed acyclic graph.

Communication cost:

- **SPANNINGTREE**: $O(|E| + |H|)$
- **DIRECTEDACYCLICGRAPH**: $O(k|E| + |H|)$
- **WILDFIRE**: $O(D_{estimation}|E|)$

Performance Metrics:

- **Accuracy**: accuracy ratio
- Efficiency: Communication cost, Computation cost, Time cost

Network Topologies: Gnutella, Random, Power-law, Grid

Major experimental results:

- High accuracy with sum and count
- WILDFIRE satisfies single-site validity whereas the performance of the other two protocols deteriorate rapidly with increasing number of leaving hosts.
- WILDFIRE has a tradeoff between validity and performance: it involves much more communication and computation costs
- WILDFIRE pays most of the price in the early phase of convergecast.
- Continuous single-site validity can be implemented within the capture-recapture scheme.

Validity: Accuracy of WILDFIRE

High convergence rate in the accuracy ratio

[Graph showing accuracy ratio over number of repetitions]

Count and Sum

Validity: Verification of Single-Site Validity

Count on Gnutella topology

Sum on Gnutella topology
Validity: Price Communication costs (cmCost) for Count query

- Random, Power-law and Gnutella topologies:
  cmCost(WILDFIRE) = 4 * cmCost(SPANNINGTREE) = 4 * cmCost(DIRECTEDACYCLICGRAPH)
- Grid topology:
  cmCost(WILDFIRE) = 5 * cmCost(SPANNINGTREE) = 5 * cmCost(DIRECTEDACYCLICGRAPH)
- Communication costs for WILDFIRE depends on both data distribution and query type
- The convergecast phase of WILDFIRE may involve less communication costs than that of SPANNINGTREE

Validity: Price Computation costs (cpCost) for Count query

- Power-law topology
  cpCost(WILDFIRE) = 2 * cpCost(SPANNINGTREE)
- Random topology
  cpCost(WILDFIRE) = 4 * cpCost(SPANNINGTREE)
- Grid topology
  cpCost(WILDFIRE) = 44 * cpCost(SPANNINGTREE)
- Why WILDFIRE has a higher computation cost than SPANNINGTREE?
  Increase in communication cost leads to a larger increase in computation cost
- Why does the performance of WILDFIRE deteriorates on Grid?

Validity: Price Time costs

- Latency(WILDFIRE) > latency(DIRECTEDACYCLICGRAPH)
- Overestimate of D has no effect on the communication costs for WILDFIRE

Validity: More on WILDFIRE

- Duplicate incentive processing
  Modified probabilistic counting algorithm
- Continuous single-site validity
  Result reflects the arbitrary value of hosts reachable from hq during a recent interval of width W
- Continuous approximate count queries
  Problem: Count the number of hosts WILDFIRE: OUR messages
  Capture-recapture scheme
  Assumptions
- To improve the efficiency of WILDFIRE
  Adjust the time of convergecast phase
  Leverage domain capabilities
Completeness: Atomicity of Updates

Design goals of DHT:
• Robustness to node failures: replicate data and routing information
• A single-system image to clients: ensure the atomicity of operations on data

Research Objective:
To find an algorithm that yields atomic data access with only one copy of data in the system

Assumptions:
• Asynchronous network
• No-failure servers

Atomicity: Problem Setting

• Three operations in DHT: join, leave and updates
• Aim to find an algorithm that guarantees atomicity, termination and stabilization in a quiescent system with Chord overlay structure

Atomicity: Modified Chord Structure

• Chord:
  A unit ring with logical identifiers;
  Assignment of nodes based on successor relationship;
  Edge information for communication between nodes;
• Additions:
  A table of physical and logical identifiers of in-links and out-links;
  Each node keeps a local copy of data objects it owns;
  Each node keeps a FIFO queue for requests and messages (requests are handled in a serialized manner at each node);
  Each node has a dispatch loop for scheduling procedures;

Atomicity: algorithm

Leave: Step 1

Atomicity: algorithm

Leave: Step 2

Atomicity: algorithm

Leave: Step 3
Atomicity: Fault-Tolerance

- State Machine Replication Algorithm
  1. To have physical nodes form replica groups elsewhere and join the network as a virtual node
  2. Two approaches: virtual nodes; consecutive physical nodes
     - Virtual node:
       Keeps virtual address encoding the replicas;
       Execute the above algorithm;
       Monitor the set of remaining active replicas;

- Consecutive physical nodes:
  A replica group keeps track of the thresholds;
  A new physical node join the replica group by hashing;
  Split a replica group if its size exceeds threshold;
  Merge two consecutive groups if a group size drops below threshold;

Atomicity: Efficiency

- Connect request
  Reduce the number of hops but more cost in constructing and deconstructing edges
- Thresholds are set at O(logn)
- Size of ring regions of replica groups are balanced to within a constant factor with high probability

Questions and Thoughts

- Empirical results show that the cost for max query on WILDFIRE is smaller than that for count. Why?
- Decrease query latency by making estimate of D closer to D. How?
  Heuristic: use WILDFIRE with a large estimate of D initially and find the maximum D among hosts in the network; Use this estimate to construct estimate of D in later queries
- What are the assumptions we have on capture-recapture scheme? How robust is this scheme to reality?
- Can we combine both validity and completeness in one framework?
- Can we extend these algorithms to other types of queries?