Pomelo: Accurate and Decentralized Shortest-path Distance Estimation in Social Graphs

Zhuo Chen¹, Yang Chen², Cong Ding³, Beixing Deng¹ and Xing Li¹
¹ Tsinghua University ² Duke University ³ University of Goettingen

Background

Shortest-path distances
- Social graph analysis
- Closeness centrality
- Graph diameter
- Friend recommendation
- Community detection

Existing problem
- Breadth First Search (BFS) is time consuming for large social graphs
- $\Theta(N^3)$ to compute all pairs shortest-path distances

Graph Coordinate (GC) System
- Embed all nodes into a d-dimensional Euclidean space in a scalable way (Fig. 1)
- Use geometric distances to estimate shortest-path distances in social graphs
- Landmark-based GCS: Orion [Zhao et al, WOSN'10]: biased input
- Pomelo: Accurate and Decentralized Shortest-path Distance Estimation in Social Graphs

Principal Component Analysis
- Data sets: Flickr (44,511 nodes & 1,951,279 edges) and BlogCatalog (28,592 nodes & 858,392 edges)
- Use PCA to analyze the shortest-path distance matrix (Table 1)
- The first 10 dimensions account for more than 70% of the information (Feasible!)

System Design of Pomelo

Motivation
- Decentralized GCS: utilizes the spring model proposed in Vivaldi [Dabek et al, SIGCOMM'04] to characterize the social graph
- Challenge: the knowledge of all pairs shortest-path distances is $\Theta(N^3)$

Our Solution
- Hybrid neighbor selection (nearby neighbors + distant neighbors)
- Select local neighbors by computing shortest-path distances to nearby nodes: partial BFS bounded by a tunable budget $T$
  - Insight: BFS is more efficient when budget is small (Fig. 2)
  - A small fix set of distant neighbors shared by every node
  - Give the nodes more of a global sense of their place in the graph

Preliminary Results

Compare Pomelo and Orion
- Main features: Table 2
- Overall prediction accuracy: Fig. 3
- RE as a measure: $RE = \frac{|EstimatedDist - ComputedDist|}{ComputedDist}$

Table 2 Features of GC systems

<table>
<thead>
<tr>
<th>Orion</th>
<th>Pomelo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centralized</td>
<td>Decentralized</td>
</tr>
<tr>
<td>Does NOT take account of the shortest-path distances between non-landmark nodes when calculating coordinates</td>
<td>Leverages the shortest-path distance information from EVERY node to its neighbors to avoid the biased input</td>
</tr>
<tr>
<td>Complete BFS</td>
<td>Partial BFS</td>
</tr>
<tr>
<td>Does a complete BFS computation ONLY for landmark node as the starting point</td>
<td>Does a partial BFS for EVERY node as the starting point</td>
</tr>
</tbody>
</table>

Fig. 1 An example of GC System

Fig. 2 Partial BFS

Fig. 3 Estimation Accuracy Evaluation