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

Zhuo Chen[1], Yang Chen[2], Cong Ding[3], Beixing Deng[1] and Xing Li[1]
[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
[2] Department of Computer Science, Duke University, Durham, USA
[3] Institute of Computer Science, University of Goettingen, Goettingen, Germany
E-mail: chenzhuo08@mails.tsinghua.edu.cn, ychen@cs.duke.edu,
cong@cs.uni-goettingen.de, dengbx@mail.tsinghua.edu.cn, xing@cernet.edu.cn

ABSTRACT

Computing the shortest-path distances between nodes is a key problem in analyzing social graphs. Traditional methods like breadth-first search (BFS) do not scale well with graph size. Recently, a Graph Coordinate System, called Orion, has been proposed to estimate shortest-path distances in a scalable way. Orion uses a landmark-based approach, which does not take account of the shortest-path distances between non-landmark nodes in coordinate calculation. Such biased input for the coordinate system cannot characterize the graph structure well. In this paper, we propose Pomelo, which calculates the graph coordinates in a decentralized manner. Every node in Pomelo computes its shortest-path distances to both nearby neighbors and some random distant neighbors. By introducing the novel *partial BFS*, the computational overhead of Pomelo is tunable. Our experimental results from different representative social graphs show that Pomelo greatly outperforms Orion in estimation accuracy while maintaining the same computational overhead.

Categories and Subject Descriptors

J.4 [**Computer Application**]: Social and behaviorial sciences

General Terms

Algorithms, Human Factors, Measurement

Keywords

Online social network, Graph Coordinate System

1. INTRODUCTION

Recently, online social networks (OSN), such as Facebook and Flickr, have gained significant popularity among users of the Internet. Computing the shortest-path distances between nodes in the social graph is one of the essential problems in analyzing the graph properties, such as computing centrality, detecting mutual friends and detecting community. Many widely used applications also benefit from the knowledge of node distances in the social graph. For instance, users of e-commerce sites may select more trustworthy sellers according to the shortest-path distances to the sellers. One can also filter out query results using shortest-path distances in websites like LinkedIn [6].

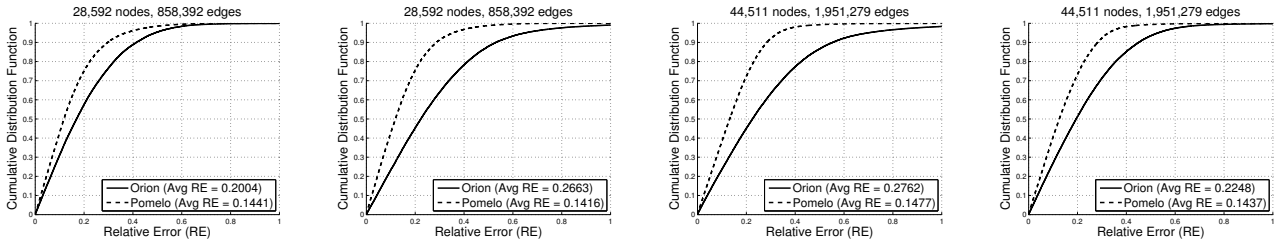
However, traditional shortest-path computation algorithms like breadth-first search and Dijkstra do not scale with graph size. They calculate all pairs shortest-paths in $\Theta(N^3)$, where N is the number of nodes. This is not tolerable for nowadays real-world massive networks, especially when the shortest-path distance must be provided in the order of several milliseconds in online applications. In [6], Zhao et al. try to embed the nodes of a social graph into a Euclidean space by assigning each node a set of low-dimensional coordinates. They propose Orion, a Graph Coordinate System, which simply uses the Euclidean distance between two nodes to estimate the actual shortest-path distance. However, similar to landmark-based Network Coordinate (NC) systems for Internet latency estimation such as GNP [5], Orion does not take account of the shortest-path distances between non-landmark nodes when calculating coordinates. Such biased input cannot characterize the graph structure well and will lead to an inaccurate estimation.

In this paper, we propose Pomelo, an accurate and decentralized Graph Coordinate System. Our contributions are three-folds: (1) The coordinate calculation of every Pomelo node refers to both the nearby neighbors and some distant neighbors. This hybrid neighbor selection policy is in accordance with the intuition in [3], thus achieving the better overall estimation accuracy. (2) Instead of conducting a complete BFS for every node as the starting node, which is $\Theta(N^3)$ in computational overhead, we propose the novel *partial BFS* for every node bounded by a pre-defined budget. This tunable budget balances the tradeoff between computational cost and estimation accuracy. In Section 3, we adjust this budget to ensure fair comparison between Orion and Pomelo. (3) In evaluating Pomelo on representative social graphs, we show that with the same computational overhead, Pomelo estimates the shortest-path distances much more accurately than Orion does.

2. SYSTEM DESIGN OF POMELO

Since we want to choose a decentralized architecture for Pomelo, the representative distributed NC system, called Vivaldi [3], is a natural choice for coordinate calculation framework. Pomelo utilizes the spring model proposed in Vivaldi to characterize the social graph. In each updating round, every node adapts its coordinates by referring to one of its neighbors. Some rounds later, Pomelo converges to minimize the sum of squares of the absolute errors, i.e., the

SIGCOMM'11, August 15–19, 2011, Toronto, Ontario, Canada.
ACM 978-1-4503-0797-0/11/08.



(a) BlogCatalog (Random) (b) BlogCatalog (HD) (c) Flickr (Random) (d) Flickr (HD)

Figure 1: Compare Pomelo with Orion using Different Landmark Selection Strategies

differences between the estimated and computed shortest-path distances. In addition, since Vivaldi (with height) has been demonstrated as the most widely used implementation of Vivaldi [4], we also apply a *height* element to the coordinates of each node in Pomelo.

As argued in [6], there is a critical challenge to apply Vivaldi directly. When utilizing the original Vivaldi algorithm, the knowledge of all pairs shortest-path distances is required in order to randomly select neighbors for each node. That said, we have to perform a complete BFS for every node as the starting point. This $\Theta(N^3)$ method is cost prohibitive for processing large social graphs. In contrast, we conduct a partial BFS for each node in Pomelo, where a pre-defined budget T is set. Whenever the total number of nodes and edges visited exceeds this budget, we terminate the BFS procedure and the nodes that have been visited become the qualified nearby neighbors for the starting node. This method is scalable, since it tunes the computational overhead for N nodes to $\Theta(TN)$. Moreover, some extra computational budget is set for conducting complete BFS from a randomly selected set of nodes. In this way, an arbitrary node can also refer to this set of relatively distant nodes, which we refer to as distant neighbors, in the algorithm of Pomelo. That said, the neighbor set of every node in Pomelo is a combination of some nearby neighbors and some distant neighbors. It is shown in [3] that such hybrid neighbor selection policy can actually increase the estimation accuracy in the original Vivaldi system, which can be viewed as a positive indicator for Pomelo.

3. EVALUATION

Firstly, Orion is implemented with a couple of landmark selection strategies, including Random and High-degree (HD) [6]. And then, we calculate BFS_i, which is the total number of nodes and edges visited to complete a BFS procedure starting from node $i(1 \leqslant i \leqslant N)$. Suppose there are K landmarks in Orion, namely $l_1, l_2, ..., l_K$, the total overhead in computing BFS in Orion is then $\sum_{1 \leqslant j \leqslant K} BFS_{l_j}$, which is denoted as O_{sum}. In Pomelo, we let all N nodes share the same set of randomly selected distant neighbors, consisting of $K/2$ nodes, namely $r_1, r_2, ..., r_{K/2}$. Therefore, the total overhead of computing every node's shortest-path distances to these $K/2$ distant neighbors is $\sum_{1 \leqslant j \leqslant K/2} BFS_{r_j}$, which is denoted as $Rand_{sum}$. For the sake of fair comparison, we set the budget for every node when doing the partial BFS in Pomelo as $T = \frac{O_{sum} - Rand_{sum}}{N}$. Since a complete BFS procedure for one node is $\Theta(N^2)$, the above equation shows that T is actually $\Theta(KN)$, so that the partial BFS used in Pomelo is much more efficient than a complete BFS.

We use two representative datasets, named Flickr and BlogCatalog, from [1]. In practice, we randomly sample some nodes in each dataset and includes all the edges between these nodes. The sampled Flickr dataset has 44,511 nodes and 1,951,279 edges and the sampled BlogCatalog dataset has 28,592 nodes and 858,392 edges. The dimension in Orion is 8, and Pomelo uses 7 dimensions with a height element. The number of landmarks is set to be 100, the same as [6]. We use relative error (RE [3, 6]) to compare the accuracy of different systems:

$$RE = \frac{|EstimatedDist - ComputedDist|}{ComputedDist} \quad (1)$$

Figure 1 depicts REs in the format of CDF using different datasets and landmark selection methods. As we see, estimation with Pomelo is much more accurate than that with Orion. Compared with Orion (Random), Pomelo reduces the average RE by between 28.08% (BlogCatalog) and 46.51% (Flickr). Moreover, compared with Orion (HD), Pomelo reduces the average RE by between 36.12% (Flickr) and 46.82% (BlogCatalog).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose Pomelo, a decentralized Graph Coordinate System using a hybrid neighbor selection policy. According to our evaluation, Pomelo achieves much better estimation accuracy than Orion does while maintaining the same computational overhead. For future work, we would like to explore the matrix factorization model [2] for Graph Coordinate System.

5. ACKNOWLEDGMENTS

This work is supported by the National Basic Research Program of China (No.2007CB310806) and National Science Foundation of China (No.60850003) .

6. REFERENCES

[1] Social computing data repository at arizona state university. http://socialcomputing.asu.edu/.

[2] Y. Chen, X. Wang, X. Song, and et al. Phoenix: Towards an accurate, practical and decentralized network coordinate system. In *Proc. of IFIP/TC6 Networking*, 2009.

[3] F. Dabek, R. Cox, and et al. Vivaldi: A decentralized network coordinate system. In *Proc. of ACM SIGCOMM*, 2004.

[4] J. Ledlie, P. Gardner, and M. Seltzer. Network coordinates in the wild. In *Proc. of NSDI*, 2007.

[5] T. Ng and H. Zhang. Predicting internet network distance with coordinates-based approaches. In *Proc. of INFOCOM*, 2002.

[6] X. Zhao, A. Sala, C. Wilson, H. Zheng, and B. Y. Zhao. Orion: shortest path estimation for large social graphs. In *Proc. of The 3rd Workshop on Online Social Networks (WOSN)*, 2010.