Random Node Pair Sampling-Based Estimation of Average Path Lengths in Networks

Luis E. Castro, Department of Industrial Engineering, University of Miami, Coral Gables, USA
Nazrul I. Shaikh, Department of Industrial Engineering, University of Miami, Coral Gables, USA

ABSTRACT

This article describes how the average path length (APL) of a network is an important metric that provides insights on the interconnectivity in a network and how much time and effort would be required for search and navigation on that network. However, the estimation of APL is time-consuming as its computational complexity scales nonlinearly with the network size. In this article, the authors develop a computationally efficient random node pair sampling algorithm that enables the estimation of APL with a specified precision and confidence. The proposed sampling algorithms provide a speed-up factor ranging from 240-750 for networks with more than 100,000 nodes. The authors also find that the computational time required for estimation APL does not necessarily increase with the network size; it shows an inverted U shape instead.

KEYWORDS

Average Path Length, Network Sampling, Shortest Path Length, Social Networks

1. INTRODUCTION

The average of the shortest paths between any two nodes of a network is a global metric of high relevance. Popularly called as the average path length (APL), it provides useful insights on the level of interconnectivity in a network and the time it would take for information/goods to flow between any two randomly selected points on the network. APL has been shown to be an important metric for tasks such as the designing of real-life transportation networks (Balmer et al., 2004; Klunder and Post, 2006; Ziliaskopoulos et al., 1997), design of routing networks (Costa et al., 2007; Dabek et al., 2004), design of web-based networks (Backstrom et al., 2012; Fu et al., 2008; Kleinberg, 2000; Newman, 2000), studying propagation of diseases (Dekker, 2013), diffusion of information and opinion dynamics (Yildiz et al., 2011). Researchers have also shown that search and navigation are easier when APL is small (Zhang et al., 2008). However, estimating APL takes a lot of time and is sometimes infeasible on account of lack of computational resources (Wang, 2006).

Researchers have shown that the computational time required to estimate APL scale as $V(V + E')$ where $V$ is the number of vertices and $E'$ a function of the number of edges in the network (Madduri et al., 2007). $E'$ scales as $O(V^{\gamma})$ where $1 \leq \gamma \leq 2$. This quadratic to cubic scaling of the computational time with network size makes the estimation of APL impractical as the network sizes increase (Wang, 2006). As an example, it takes approximately 9.6 hours to estimate the APL of a synthetic network created using the preferential attachment algorithm with about 100,000 nodes on

DOI: 10.4018/IJORIS.2018070102

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
a 16GB RAM PC and the time requirement increases to more than five days for a network with 1 Million nodes. Such a long wait time is generally impractical, especially in simulation and emulation studies that require generation of 100s of synthetic networks and the estimation of APL for each scenario.

In this paper, we develop a random node pair sampling based strategy to estimate the APL for mid-sized networks when both computational time and capacity are limited. Though sampling introduces some uncertainty in the reliability of the parameter estimates, the precision and confidence can be controlled using a combination of the right sampling strategy and sample size. We therefore propose and demonstrate the efficacy (regarding computational time, confidence level, and precision) of the proposed node pair sampling algorithm. We compare the proposed algorithm with random node sampling algorithm and algorithms wherein the node sampling is non-uniform. The random node pair sampling algorithm yields a speed up factor of more than 411 when compared to the algorithm that uses random node sampling and speed up of 750 when compared to the algorithm that measures APL using the population information. The proposed algorithm uses the central limit theorem approximation to determine the sample size for a given precision and confidence level.

This paper is organized into six sections. We present a brief literature review in Section 2 where we focus on the algorithms used for estimating the shortest path lengths (SPL) and network sampling. APL estimates use SPL. The proposed sampling based APL estimation algorithms are presented in Section 3, and their performance on simulated networks is discussed in Section 4. The algorithms are also applied to real networks and the performance of the algorithms on real networks in presented in Section 5. Finally, the conclusions are presented in Section 6. We limit our focus on mid-sized network i.e., networks of size up to 5 million nodes as these can be processed on a single core on most modern-day PCs. We do this so that the focus of the paper stays on speed-up and scale-up due to sampling without the need for discussing graph partitioning.

2. BACKGROUND

Estimation of APL is intricately linked to the distribution of the shortest paths between any two randomly selected nodes.

2.1. The SPL Problem

The SPL, also called the geodesic distance between a pair of nodes, is defined as the minimum number of nodes that need to be traversed to reach the desired destination node from a given source. SPL is used in several areas such as transportation systems and route planning (Arney et al., 2013; Balmer et al., 2004; Klunder and Post, 2006; Li et al., 2012; Ziliaskopoulos et al., 1997), server selection and data queries (Costa et al., 2007; Dabek et al., 2004; Rètvári et al., 2007), and path finding in social networks (Boyles and Rambha, 2016; Kleinberg, 2000; Leskovec and Faloutsos, 2006; Leskovec et al., 2005).

Based on the domain and the purpose of estimating SPL, the research on SPL can be classified into four. The first class of research on SPL attempt to find the shortest path between a specified source $i$ and destination $j$. The second class of research focuses on finding the SPL from source $i$ to all other nodes while the third class focuses on finding the SPL between all combinations of sources $i$ and destinations $j$. The fourth class relies on the creation of component hierarchies for each node based on spanning trees and subsequently estimate the SPL.

Researchers usually rely on the Floyd-Warshall algorithm (Floyd, 1962) and its derivatives for finding the shortest path and SPL between a specified source $i$ and destination $j$. This algorithm compares all the paths in a network from a node $i$ to $j$, starting with a comparison based only on their neighbors, and continuously increasing until the optimal value of the shortest path is reached. This recurrence may be seen as a dynamic programming sequence. The algorithm scales as $O(V^3)$.
[www.igi-global.com/chapter/a-cognitive-analytics-management-framework-cam-part-1/121485?camid=4v1a](www.igi-global.com/chapter/a-cognitive-analytics-management-framework-cam-part-1/121485?camid=4v1a)

Evaluation of BPS and Its Impact: Quantitative Approach
[www.igi-global.com/chapter/evaluation-of-bps-and-its-impact/121933?camid=4v1a](www.igi-global.com/chapter/evaluation-of-bps-and-its-impact/121933?camid=4v1a)