Programming Assignment 2

Small-World Networks: Analyzing Large-scale Social Networks Using MapReduce

Due: October 14, 2021 5:00PM
Submission: via Canvas, individual submission

Objectives
The goal of this programming assignment is to enable you to gain experience in:

- Counting the number of edges and vertices of the Google+ "friend" networks
- Measuring small-world network properties using MapReduce
- Analyzing the results of your measurements

1. Introduction

Small-world networks are a class of networks that are "highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs.[1]". This property of networks is closely related to the efficient information transfer between and within regional clusters. Social networks are intuitive examples of the Small-world networks. Cliques or clusters of friends are often interconnected while each member in a cluster is only several friend relationships away from anyone else.

In this assignment, you should profile the Google+ ego networks using the "friend" network (gfn) that you have created in the assignment 1. For students who could not extract the friend network correctly, we provide the pre-processed friend network as a part of assignment 2. To analyze the Small-world networks property, you should calculate the following measurements using Hadoop MapReduce:

(1) Requirement 1: Counting the numbers of edges and vertices of the Google+ friend networks
(2) Requirement 2: The average geodesic path length
(3) Requirement 3: The global cluster coefficient

[1] Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small-world'networks. *nature*, *393*(6684), pp.440-442.

2. Background

In the 1950s, Ithiel de Sola Pool and Manfred Kochen wrote a manuscript called "Contacts and Influence"[2] including a discussion of quantifying the distance between people through chains of connections. To test for the existence of short paths of social connections between people, experimental psychologist Stanley Milgram conducted landmark studies in the 1960s on the Small-world phenomenon in human social networks[3] . In this research, Milgram quantified the typical distance between nodes in a social network and to show that one should expect it to be small.

In one of his experiments, Milgram sent 96 packages to people living in Omaha, NE, USA who he chose "randomly" from a telephone directory. Each package contained an official-looking booklet with an instruction that recipients should attempt to get this booklet to a specific target individual (a friend of Milgram's who lived in Boston, MA, USA). The only information supplied about Milgram's friend was his name, his address, and the fact that he was a stockbroker. Each recipient was instructed to send the package to somebody that they knew on a first-name basis who they felt would be socially "closer" to Milgram's friend. Each person who subsequently received one of the packages was also asked to follow the same instructions.

As the result, the target individual received 18 of the 96 packages. This success rate was higher than expected, the mean number of hops of completed paths was about 5.9. This led to the popularization of the idea that there are no more than about 6 steps between each pair of people in the world, which is encapsulated by the phrase "6 degrees of separation."

3. Identifying Small-World Networks

Small-world networks tend to contain cliques, and near-cliques. This can be quantified with the clustering coefficient. Also, most pairs of nodes will be connected by at least one geodesic path (shortest path). This follows from the defining property that the mean-shortest path length be small. Several other properties are often associated with small-world networks. Often there is a super hub in the network with a high number of connections. These hubs serve as the common connections mediating the short pathlengths between other edges.

In this assignment, we will focus on the two properties, the average characteristic path length (*L-gfn*) and the cluster coefficient (*C-gfn*). You must measure the average geodesic path and the cluster coefficient of the Google+ friends network using MapReduce.

3.1. The Average Geodesic Path Length (Requirement 1 and 2)

Small-world networks have short pathlengths as is commonly observed in random networks. A network (or graph) consists of nodes (i.e., vertices) connected by edges. A path in a network is a sequence of alternating nodes and

[2] I. de Sola Pool, M. Kochen. Contacts and influence. Social Networks 1, 5–51 (1978–1979).

[3] J. Travers, S. Milgram. An experimental study of the small world problem. *Sociometry* 32, 425–443 (1969).

edges that starts with a node and ends with a node. Nodes or edges can appear multiple times in the same path, and the number of edges in a path is the length of the path. If a graph is connected, then any node can be reached via a finite-length path starting from any other node. The shortest path between a pair of nodes is called a *geodesic path* and there can be more than one such path. In a friend graph, the distance from a person to her friends is 1, the distance from her to a friend of her friend who is not also her friend is 2, and so on.

In Figure 1, we are given a network encompassing the set of nodes, $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$. The geodesic path length between v_7 and v_4 is 1 and the path is (v_4, v_7). Similarly, the geodesic path length between v_2 and v_4 is 3 and the path is (v_2, v_1, v_7, v_4). In this case, there is another geodesic path: (v_2, v_1, v_5, v_4).
Small values of shortest path length ensure that information or resources easily spreads throughout the network. This property makes distributed information processing possible on technological networks and supports the six degrees of separation often reported in social networks.

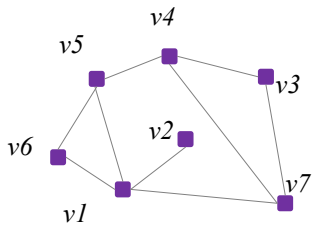

Figure 1. A simple example network

You must calculate all of the geodesic paths between pairs of vertices in the friend network and return the average geodesic distance using MapReduce. To calculate the average length, you should use the following formula:

$$L = \frac{2}{v(v-1)} \sum_{if\ n, i \neq j} d_{i,j}$$

, where v is the number of nodes in the Google+ friend network and d-i,j is the pathlength between the node i and j. If there is a pair of nodes without existing path, please ignore those path lengths. This computation may involve multiple MapReduce jobs. Consider only unique pairs of nodes. Since this graph is non-directed graph, the pair **(*i, j*)** and **(*j, i*)** are considered as the same pair. Also, if there are multiple shortest paths between the nodes, consider only one shortest path for measuring the average geodesic pathlengths.

To calculate the total shortest path lengths of this graph ($\sum_{if\ n, i \neq j} d_{i,j}$), you can count unique path with each path length and add them up. For example,

$$\sum_{if\ n, i \neq j} d_{i,j} = \sum_{for\ d_{i,j}=1} d_{i,j} + \sum_{for\ d_{i,j}=2} d_{i,j} + \sum_{for\ d_{i,j}=3} d_{i,j} + \cdots + \sum_{for\ d_{i,j}=\max\_length} d_{i,j}$$

Your software must include all of the paths with the pathlength of 1 to 8 (max_length). Please note that these paths must include the **"shortest" paths** between given pair of nodes. If there are three paths, a-b, a-c-b, and a-f-d-s-g-b, your calculation must consider only a-b. Also, do not count the duplicate path. The path a-b and b-a should be considered as the same path.

Your submission must include the measurements as depicted in the following table.

Category	$\sum_{for\ d_{i,j}=1} d_{i,j}$	$\sum_{for\ d_{i,j}=2} d_{i,j}$	$\sum_{for\ d_{i,j}=3} d_{i,j}$	…	$\sum_{for\ d_{i,j}=08} d_{i,j}$	Total shortest pathlength	Total number of vertices	The Average Geodesic Path Length

Measurement								

Table 1. The submission requirement 1

3.2. Cluster Coefficient: Large degree of Transitivity (Requirement 3)

The overall clustering in a network can be determined by averaging the clustering across all individual nodes. High clustering supports specialization as local collections of strongly interconnected nodes readily share information or resources. Conceptually, clustering is quite straightforward to comprehend. In a real-world analogy, clustering represents the probability that one's friends are also friends of each other.

A cluster coefficient is a measure of the degree to which nodes in a graph tend to cluster together. A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected edges. Therefore, a triangle (e.g. v_1, v_5, v_6 in the Figure 1) involves three different triplets, $((v_1, v_5),(v_5,v_6))$, $((v_1, v_5),(v_1,v_6))$ and $((v_1, v_6),(v_5,v_6))$. In this assignment, you should use the following measure proposed by Luce and Perry[4]. A global clustering coefficient is defined as the following:

$$C = \frac{3 \times number\ of\ triangles}{number\ of\ connected\ triples}$$

, where a triangle consists of 3 nodes that are completely connected to each other (i.e., a 3-clique) and a connected triple consists of three nodes $\{i, j, k\}$ such that node i is connected to node j and node j is connected to node k. The factor of 3 arises because each triangle gets counted 3 times in a connected triple. The clustering coefficient C indicates how many triples are in fact triangles. A complete graph, in which every pair of nodes is connected by an edge, with $N \geq 3$ nodes yields the maximum possible value of $C = 1$, as all triples are also triangles. The minimum value of the clustering coefficient is $C = 0$.

You must calculate the global cluster coefficient of the Google+ friend networks using MapReduce. Your submission must contain the total number of connected triples and number of triangles. This computation may involve multiple MapReduce jobs.

Your submission must include the measurements as depicted in the following table.

Category	Total number of triples	Total number of triangles	Global Clustering Coefficient
Measurement			

Table 2. The submission requirement 2

[4] R. D. Luce and A. D. Perry (1949). "A method of matrix analysis of group structure". Psychometrika. 14 (1): 95–116. doi:10.1007/BF02289146

3.3. Analysis of the Small-World Networkness for the Google+ "Friend" networks (Requirement 1, 2, and 3)

The short average path length is a characteristic of random graphs, while high clustering is a property of lattice networks. Watts and Strogatz's model shows that the characteristic pathlength L and clustering coefficient C of small-world network are closely related to L and C of random networks[5]. To apply this observation, you must calculate L and C of random networks for given number of edges and vertices in the Google+ friend networks.

Assume that the e as the total number of edges and v as the total number of vertices in the graph and you have counted as a part of the requirement 1. Pathlength L of random networks (*L-random*) is estimated as $\ln(v)\ /\ln\ (k)$, where $k\ (=e/v)$ is the average number of edges per node. Also, the clustering coefficient, C, of random networks (*C-random*) is estimated as k/v. Compare these numbers with your *L-gfn* and *C-gfn* calculated in the section 3.

Your submission must include the measurements as depicted in the following table.

	L-gfn	*L-random*	*C-gfn*	*C-random*
Google+ Friend Network				

Table 3. The submission requirement 3

Your analysis of the above table will be discussed as a part of the software demonstration with GTA. To prepare for the discussion, please read Watts and Strogatz's paper listed below.

- Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small-world'networks. *nature*, *393*(6684), pp.440-442

4. Programming Requirements

4.1. Programming Language

You are required to use Java (Version 1.8 or higher) for this assignment.

4.2. Installing and configuring Hadoop

For this assignment, you will be working on your own Hadoop Cluster, which should have finished this as a part of PA0. The walkthrough guidelines are available at: [Link]

5. Submission

This is an individual assignment. The result of your computation should be stored as file(s). You should generate the results using the Hadoop's MapReduce programing framework and **not a standalone program** that executes only on one machine [standalone programs will result in an automatic zero on the assignment].

[5] Watts, D.J. and Strogatz, S.H., 1998. Collective dynamics of 'small-world'networks. *nature*, *393*(6684), pp.440-442.

Please submit a tar ball of your source files and output data (From the full dataset) via Canvas. The source files should include your java code of the Mapper/Reducer functions and any script file that you will use for demo. You will download your source files/script from Canvas for the demo. Do not miss any files. For your demo, you are not allowed to use any file outside of your Canvas submission.

6. Grading

Each of the submissions will be graded based on the demonstration of your software to GTA. During the demonstration, you should present capabilities to accomplish the following requirements:

1. Counting the number of edges and distinct vertices (2 points)
2. The Average Geodesic Path Length (3 points)
3. Cluster Coefficient (3 points)
4. Analysis (2 points)

The demo includes a short interview of your software design and implementation details. Your responses and demonstration of capabilities will count towards your score. This assignment will account for 10% of your final grade. The grading will be done on a 10 point scale.

You are required to work alone on this assignment.

7. Late policy

Please check the late policy posted on the course web page.