Programming Assignment 1

Explorative Analytics of an Evolving Citation Network using Apache Spark

Due: Feb. 14 Tuesday 5:00PM
Submission: via Canvas

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

- Installing and using analytics tools such as HDFS and Apache Spark
- Performing analytics over a large-scale temporal network

1. Overview

In this assignment, you will perform an analysis of a continuously evolving temporal network. Large-scale networks are observed in many different sociological and scientific settings such as computer networks, networks of social media, academic/technical citation networks and hyperlink networks. To understand such networks, there have been several properties of interest based primarily on two key measurements: the degrees of nodes and the shortest distances between pairs of nodes. The node-to-node distances often infer the graph's diameter, which is the maximum shortest distance among all the connected pairs of nodes.

Most of the large networks evolve over time by adding new members/items and relationships between them or removing some of them. To investigate network evolution, there have been two major hypotheses. (a) the average node degree in the network remains constant over time. (Or the number of edges grows linearly in the number of nodes.). (b) the diameter is a slowly growing function of the network size. How are these hypotheses (a) and (b) reflected in real-world data?

In this assignment, you measure fundamental network properties with a citation network and investigate how they evolve. You will perform the following computations to investigate growth patterns of these networks using Apache Spark.

2. Empirical Observation 1: Density of the graph

You should measure the number of nodes $n(t)$ and the number of edges $e(t)$ at each timestamp t. Here, $n(t)$ and $e(t)$ are the final number of nodes and edges at the target time t. As part of the results, you must generate a set of log-log plots of the number of edges $e(t)$ versus the number of nodes $n(t)$. Observations on a yearly basis must be provided as part of your results.

The dataset for this assignment is the arXiv citation graph[1] that covers papers published in the period from January 1993 to April 2003 (11 years). For each year y ($1993 \leq y \leq 2003$), you should calculate $e(t)$ and $n(t)$. Consider only the out-degree of a node.

[1] J. Gehrke, P. Ginsparg, and J. M. Kleinberg. Overview of the 2003 kdd cup. SIGKDD Explorations, 5(2):149–151, 2003 (Available in Canvas)

Then record the results in a table (See Table 1) and plot them on a logarithmic scale with the corresponding year values (See Figure 1). Your measurements should be a snapshot of the graph at the end of the given year y. If you are calculating the values for the year y, your measurements must include all the nodes and edges observed at the end of the year y. For the data in 2003, consider data recorded until April, 2003. Similarly, for the data in 1992, consider data available at the end of 1992 in this dataset.

Year	1992	1993	1994	1995	1996	…	2001	2002	2003 (Apr)
$e(t)$									
$n(t)$									

Table 1: A template of the submission requirement 1-1

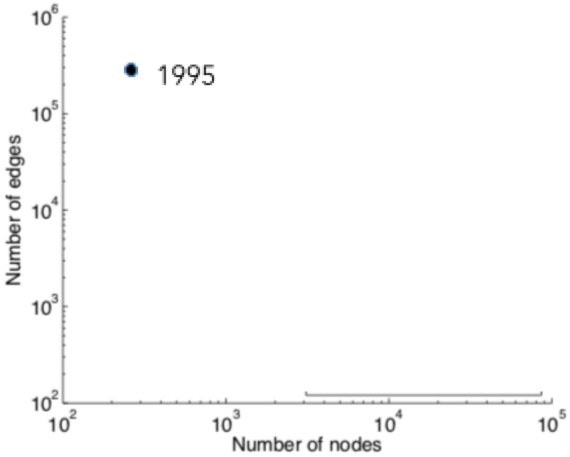


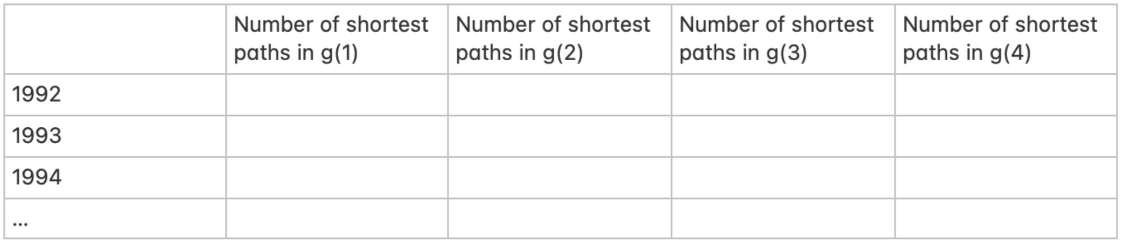
Figure 1: A template of the submission requirement 1-2

3. Empirical Observation 2: Diameter of the graph

In this phase, you should observe the change of diameter of the arXiv dataset. We assume that two nodes in an undirected network are connected if there is a path between them. Let g(d) denote the fraction of connected node pairs whose shortest path has length at most d. For example, for a graph G, g(1)is a set of nodes that has edge between them.
Note: For this experiment, please consider the graph as an undirected graph.

To represent the path between nodes, the hop-plot for a network is defined as a set of pairs (d, g(d)).This provides the cumulative distribution of distances between connected node pairs. To observe the impact of newly added nodes and their citations, you should calculate g(1), g(2), g(3), and g(4) for each year and count the number of shortest paths per subgraph.

Table 2: A template of the submission requirement 2



	Number of shortest paths in g(1)	Number of shortest paths in g(2)	Number of shortest paths in g(3)	Number of shortest paths in g(4)
1992				
1993				
1994				
...				

4. Analysis of the temporal citation network

As a part of the programming assignment 1, you should submit a short analysis report answering the following questions 1 and 2. Your analysis should not exceed 1,000 words.

Analysis question. 1

There are a set of fundamental network properties that vary with time. These properties are often assessed based on growth patterns. In the context of temporal network evolution, the densification power law is a concept that posits that the number of edges grows in a power law over the number of nodes over time. This concept is often contrasted to the general assumption of a linear trend.

Based on your observation that was plotted following the template depicted in Figure 1, does your plot correspond to the exponent in the densification law? If it does, explain why you conclude that your growth pattern matches the densification law. Otherwise, explain why your growth pattern shows the specific trend that you have observed.

Analysis question. 2

Based on your observation that was calculated in Table 2, what can you conclude about the growth pattern of connectivity? Also, explain why it
occurs with the given dataset.

5. Requirements of Programing Assignment 1

Your submission of this assignment should provide:

A. Your source codes. Please do not submit your dataset.

B. Results of the empirical observation 1 (result table and graph: please see Table 1 and Figure 1)

C. Results of the empirical observation 2 (result table: please see Table 2)

D. Short report of the analysis (see section 4)

<u>You are not allowed to use existing implementations. A 100% deduction will be assessed in that case.</u>

Demonstration of your software should be online. Demonstration will include an interview discussing implementation and design details. Your submission should include your source codes and results. Your submission will be via Canvas.

6. Programming Environment

A. Requirements

All software demonstrations of this assignment will be using the cluster in CSB120. You should make sure your software works on the machines in CSB120.

B. Setup your Hadoop and Spark cluster

This assignment includes a requirement of setting up your own HDFS and Spark cluster on 10 nodes. For writing your intermediate and output data, you should use your own cluster setup. We now have an automated script for setting up the files in your hadoopConf/sparkConf and built-in modules for setting the paths of your .bashrc file. You can find instructions for these in the Hadoop_Spark_Setup.pdf located on the Programming Assignment 1 page in Canvas.

7. Grading

Your submissions will be graded based on the online demonstration to the GTA using the computing cluster in CSB120. All of the items described in the section 5 will be evaluated. This assignment will account for 15% of your final course grade.

- 5 points: Empirical observation 1
- 5 points: Empirical observation 2
- 3 points: Analysis
- 2 points: Set up of the cluster

8. Late Policy

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