# Referee Report

Paper Number:
Reviewer's Name: Jesse Bannon
Name of Paper: A Case Study of Complex Graph Analysis in Distributed Memory: Implementation and Optimization
Author(s): George Slota, Sivasankaran Rajamanickam, Kamesh Madduri

## Section I. Overview

**A. Reader Interest**

1. Which category describes this manuscript?

(x) Practice / Application / Case Study / Experience Report
( ) Research / Technology
( ) Survey / Tutorial / How-To

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under IIIA.

( ) Very Relevant
(x) Relevant
( ) Interesting - but not very relevant
( ) Irrelevant

**B. Content**

1. Please explain how this manuscript advances this field of research and / or contributes something new to the literature. Please explain your answer under IIIA. Public Comments.

2. Is the manuscript technically sound? Please explain your answer under IIIA. Public Comments.

(x) Yes
( ) Appears to be - but didn't check completely
( ) Partially
( ) No

**C. Presentation**

1. Are the title, abstract, and keywords appropriate? Please explain your answer under IIIA. Public Comments.

(x) Yes
( ) No

2. Does the manuscript contain sufficient and appropriate references? Please explain your answer under IIIA.

(x) References are sufficient and appropriate
( ) Important references are missing; more references are needed
( ) Number of references are excessive

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under IIIA. Public Comments.

(x) Yes
( ) Could be improved
( ) No

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain your answer under IIIA. Public Comments.

( ) Satisfactory
( ) Could be improved
(x) Poor

5. Please rate and comment on the readability of this manuscript. Please explain your answer under IIIA.

( ) Easy to read
(x) Readable - but requires some effort to understand
( ) Difficult to read and understand
( ) Unreadable

## Section II. Summary and Recommendation

**A. Evaluation**

Please rate the manuscript. Please explain your answer under IIIA. Public Comments.

( ) Award Quality
( ) Excellent
( ) Good
(x) Fair
( ) Poor

**B. Recommendation**

Please make your recommendation.  Please explain your answer under IIIA. Public Comments.

( ) Accept with no changes
( ) Author should prepare a minor revision
(x) Author should prepare a major revision for a second review
( ) Reject

## Section III. Detailed Comments

### A. Public Comments (these will be made available to the author)

**Explanation for the Recommendation**
---

The algorithms presented are simplistic alternatives to other implementations that offer a significant speedup. However, the presentation of the algorithms severely lack. There is a missing gap between reading the text and the pseudo-code. Too many variables and ideas are spread amongst the text, which cause the reader to have to flip back and forth between pages and the code to understand. A better alternative to enhance organization would be to present pseudo-code between condensed paragraphs of information. This would give the reader a stronger understanding of the algorithms in a single pass of reading the text.

The performance results give insight to the internals of the algorithms, and describe why they perform as shown. The main contribution of the paper is to show how these implementations are magnitudes faster than other implementations. Only one small graphic showcases this. Fig. 4 could be split up into two larger graphics with subsections describing in more detail why the other implementations performed as they did.

Lastly, section VI does not contribute anything to this work. While the analysis is somewhat interesting, it does not relate to performance and is useless to anyone seeking faster implementations of graph problems. This entire section should be removed in exchange for a more detailed related works and room for better organization of the algorithms.

**Summary of the Paper and Assessment**

---

The paper presents two novel distributed implementations of commonly structured algorithms in graph analysis with end-to-end performance results, and compares these results against other popular software frameworks that support graph analysis. The novelty comes from the simple data structures used and the minimal amount of code required for full functionality of the algorithms. The importance of performance analytics on real world data sets is mentioned throughout the paper to emphasize their work is not optimized for synthetic data.

Their experiment is performed on the NCSA Blue Waters supercomputer. One of the main reasons for choosing this supercomputer is for the high performance file system. Its I/O bandwidth is 960 GB/s. They also list several datasets that they use in their performance analysis. They do not have labels for their dataset variables, so the reader has to infer what these variables are.

They first describe the data ingestion portion of their algorithm. They use two 32-bit unsigned integers to represent a directed edge, and stripe the input file across multiple storage units to store it into a one-dimensional graph representation in parallel on each task. It is unclear when they say “to store incoming edges, the order of edges is reversed and another exchange is performed”.

The subsequent sections include the pseudo-code of their algorithms to help assist in their explanation. While their decomposition is sound, the way they describe the algorithms is hard to read: variable definitions are spread throughout the text and the pseudo-code has no comments. Thankfully the algorithms presented are simple because had they not, it would have been especially hard to fully understand them.

Their performance results deliver, but lack comparison to other implementations. The only graph which does compare to other versions is small, and compares just two of their six algorithms (one of each class) to other implementations. Comparing to the size of the graphics in Section VI, this should have received more space in the paper.

Section VI describes some insights found while working with the web crawl dataset. The information presented does not contribute anything to their findings.