Electronics and Computer Science
Faculty of Physical Sciences and Engineering
University of Southampton

Vlad Sebastian Velici
December 6, 2014

Similar nodes in large graphs

Project Supervisor: Dr. Adam Prügel-Bennett
Second Examiner: Dr. Sasan Mahmoodi

A progress report submitted for the award of
MEng Computer Science with Artificial Intelligence

Contents

Introduction

Plenty of datasets are or can be represented as graphs where vertices represent entities and edges represent relationships between entities. A problem of interest is to find entities that are similarly connected. Example instances of this problem are finding *people you may know* in a social network, people with common interests from research publications repositories or identifying possible duplicates in a dataset.

It is easy to find exact similarities between vertices in small graphs by performing pairwise comparisons. Such an algorithm is too slow for large datasets of millions of vertices. This project investigates a method to compute an approximation of similarities between nodes and attempt to evaluate its performance on different datasets.

The algorithm

Limitations of current implementation

The current implementation of this algorithm has various limitations which are discussed along with possible improvements.

Only undirected graphs Directed graphs are not currently supported. In practice, datasets have meaningful unidirectional relationships (e.g. in a social network person A follows person B, but B does not follow A), and often datasets are represented as directed graphs rather than undirected graphs. The algorithm can be adapted to support both directed and undirected graphs but it will have the disadvantage of requiring to compute V^{-1} (for undirected graphs, $V^{-1} = V^T$).

Not distributed What if the dataset is too large to fit into main memory? The algorithm is currently only designed to run on one machine. Methods of distributing the algorithm on more than one machine over a network will be investigated in the future.

REFERENCES

Anand, U., 2010. The Elusive Free Radicals, *The Clinical Chemist,* [e-journal] Available at:<http://www.clinchem.org/content/56/10/1649.full.pdf> [Accessed 2 November 2013]

Biology Forums, 2012. *Normal glomerulus. Acute glomerulonephritis.* [online] Available at: <http://biology-forums.com/index.php?action=gallery;sa=view;id=9284> [Accessed 23 October 2013].