Technische Universiteit **Eindhoven** University of Technology

J.A.J. Hellings MSc

Jelle is currently a PhD student at Hasselt University, Belgium. He graduated cum laude for the Master program in Computer Science and Engineering and also obtained a certificate in Philosophy; both at the TU/e.
His graduation topic was external memory bisimulation; and graduation was under daily supervision of George Fletcher.

Bisimulation partitioning and partition maintenance

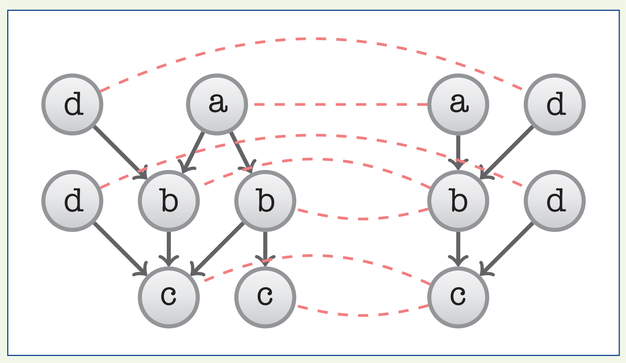

Figure 1. An example of a graph (left) and its structural description based on bisimulation (right). Bisimilar equivalent nodes in the two graphs are connected by a red, dotted line.

Networks-or, in mathematical terms, graphs-are fundamental structures that arise in numerous areas. Road networks or railway networks are obvious examples, but graphs are also used extensively to model relational information. The nodes of the graph are then the objects of interest, and the edges indicate which pairs of objects are related. For example, in social networks the nodes are people and the edges denote friendships between people.

Representing data by graphs gives access to a broad range of graph-based analysis technologies. The cost of graph-based analysis technologies heavily depends on the size of the graph: the bigger the graphs are, the longer computations on the graph will take. Because graphs are typically huge, compression techniques are necessary in order to support efficient analysis of the graph data.

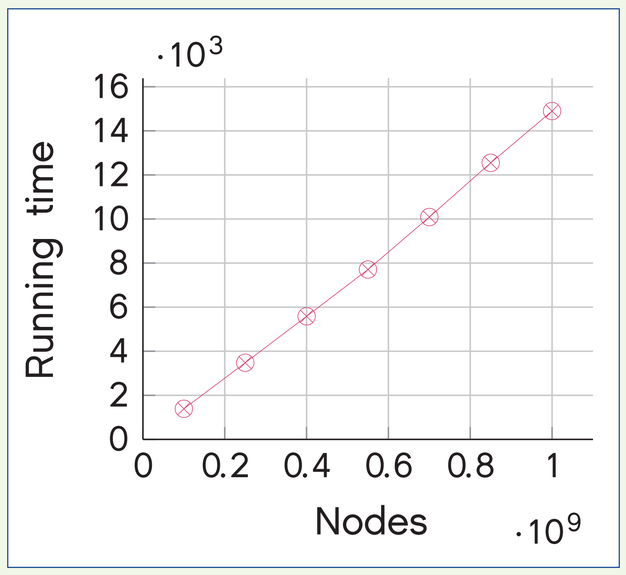

Figure 2. The running time of the implementation of our algorithm on very large graphs; showing good scalability for very large inputs.

One popular method to perform compression uses the concept of bisimulation. Intuitively, two nodes are bisimilar if they ‘reflect the same behaviour’. Often one can compress a graph significantly by bisimulation partitioning, that is, by grouping its nodes into clusters of bisimilar nodes and replacing each cluster by a single node. Performing bisimulation partitioning has been very time-consuming when graph data sets are so large that they do not fit in the computer's main memory but must be stored on disk. Existing algorithms spend most of their time transferring data back and forth between main memory and disk, which is extremely costly. Thus the application of bisimulation partitioning was essentially limited to relatively small graphs.

In our research we have discovered the first practical and efficient approach for bisimulation partitioning massive graphs residing on disk. This sets the stage for progress in the wide variety of practical applications of bisimulation. We have specialized our methods to practical XML technologies and we have provided open-source implementations of our algorithms and ran performance tests which confirm the practical efficiency of our algorithms.

Lastly our results will be presented at SIGMOD 2012; the premier international conference on databases. It is extremely selective-in 2012, only 16% of submissions were accepted-and publishing a paper in the highly cited SIGMOD formal proceedings (published by ACM Press) is very prestigious.

Nominated for the TU/e Final Project Award 2012