A Simplification Algorithm for Visualizing the Structure of Complex Graphs

Dan Hennessey, Dan Brooks, Alex Fridman, Dr. David Breen
Department of Computer Science

Introduction

Graphs are a popular construct for representing many types of data. Graphs frequently have hundreds or thousands of nodes and edges connecting them. Displaying all of these nodes and edges to a user provides no real benefit, since the density and complexity of the graph overpowers and hides standard displays due to resolution limitations. Graph simplification provides one way of making these graphs easily comprehensible while not removing the actual meaning or structure of the graph.

In our approach a complex, cluttered graph is progressively thinned by the user until the underlying structure of the graph becomes apparent. This is achieved through the use of weighting functions that quantify the importance of the nodes to be removed. The actual meaning or structure of the graphs and edges to a user provides no real benefit, since the density and complexity of the graph overpowers and hides standard displays due to resolution limitations.

Number of N-Ring Neighbors: The node weight is based on the number of neighboring nodes exactly N rings away from the original node. The more neighbors a node has the better connected it is within the graph.

Number of Shortest Paths: Weight is calculated from the number of shortest paths that pass through the node. For a shortest path between nodes (u, v), each node that belongs to the path has its weight incremented by one. The endpoints of the path do not have their weights incremented. When there are two or more shortest paths of the same length between two nodes one of these is selected at random during the node increment process.

Eccentricity: A node's eccentricity is defined as the maximum distance within the set of shortest paths from the node to all other nodes in the graph. A node's weight is set equal to its eccentricity value minus the maximum eccentricity value for all nodes. Higher values in this metric mean that the given node is a significant distance away from the farthest component in the entire graph, or in other words it is closer to the periphery of the graph than the center.

Distance to Center Metric

Distance to Leaf Metric

Shortest Distance to a Center Node: This metric is a variation on the eccentricity metric mentioned previously. All nodes with the maximum eccentricity value are defined as center nodes of the graph. An importance metric is computed based on a node's shortest distance to one of the center nodes (there can be more than one) minus the maximum distance for all nodes. Center nodes are important because they have a maximum distance to the periphery of the graph, meaning they are centrally located in the graph.

Shortest Distance to a Leaf Node: This metric assumes leaf nodes exist within the graph. It is assumed that the further a node is from a leaf node, the more centrally located within the graph it is, making that node more important.

Node Pruning

The pruning process visits all nodes and removes those nodes, and their associated edges, with importance values below a user-specified threshold, T. In order to maintain the semantic structure of the graph we do not prune nodes whose removal would split the graph into multiple components and destroy its connectivity. Producing a simplified graph that has been split into multiple components would convey incorrect structural information about the graph.

The pruning process can be done incrementally. When pruning a weighted graph no new increment value along with a larger threshold value. The initial threshold is set to 0. The increment is added to the current threshold and the pruning is performed again on the previously pruned graph. This continues until the current threshold becomes greater than or equal to the larger threshold. Incrementally removing nodes will typically provide better results than a non-incremental removal.

Results

Inet Graph

This graph has been generated by the INET Internet Topology Generator. This graph represents an Autonomous System level Internet topology. The original form of this graph contains 4500 nodes.

Citation Graph

This graph is derived from a set of papers and their references are interconnected. Each node represents a paper and edge indicates a reference from one paper to another. The original form of this graph contains 1025 nodes and 15430 edges.

Future Work

- Development of further weighting metrics.
- Further methods of pruning, such as node or edge collapsing.

Acknowledgements

We would like to thank Dr. Chaomei Chen of Drexel's iSchool for providing the citation graph data set. This work is funded by the U.S. Army Communications-Electronics Research, Development and Engineering Center (CERDEC), under contract #DAAB07-07-1-9-L304.

Metric Performance

Figures 1-8. Each set of figures contains a series of images generated by simplifying the graphs based on the specified metrics. The T value is the threshold value used to generate the graph. The N values are the number of nodes left in the simplified graph.