Finding Reliable Subgraphs from Large Probabilistic Graphs*

Petteri Hintsanen and Hannu Toivonen

HIIT Basic Research Unit, Department of Computer Science,
PO Box 68, FI-00014 University of Helsinki, Finland
{petteri.hintsanen,hannu.toivonen}@cs.helsinki.fi

Consider information search or discovery in a large graph of concepts and their weighted relationships. The user initiates a query by specifying some concepts (“search terms”), and wishes to obtain other concepts and relationships that connect the search concepts. An application example is in analysis of biological information, conveniently represented as a graph of biological concepts and their relations. A search engine we envision would allow a life scientist to query for connections between a gene and a phenotype, for instance, to find information supporting a hypothesis, or to help discover new hypothesis.

We propose two new methods for this problem, formalized as the *most reliable subgraph problem*. Such a problem is specified by a probabilistic graph G subject to random edge failures, a set of terminal vertices, and an integer K . The objective is to remove K edges from G such that the probability of connecting the terminals in the remaining subgraph is maximized. Extracting a subgraph of maximal reliability is a fundamental task. Besides the search problem described above, subgraph extraction is useful for visualization of large graphs and for preprocessing large graphs for other analysis methods.

The proposed methods, BPI and SPA, are based on greedy strategies for incrementally constructing a reliable subgraph of the desired size. BPI is based on simple use of best paths to span a subgraph, whereas SPA involves a more elaborate construction of series-parallel graphs. Unlike previous methods that prune the original graph until it reaches the given size, these incremental methods are relatively insensitive to the size of the original graph. We provide some technical details and a rough analysis of the algorithms. The practical performance of the methods is evaluated on real probabilistic graphs from the biological domain. The results indicate that the proposed methods scale much better to large input graphs, both computationally and in terms of the quality of the result. We use subgraph extraction as a component in Biomine¹, a search engine prototype for information discovery in biological databases.

References

1. Hintsanen, P., Toivonen, H.: Finding reliable subgraphs from large probabilistic graphs. *Data Mining and Knowledge Discovery* 17(1), 3–23 (August 2008)

* This is an extended abstract of an article published in the *Data Mining and Knowledge Discovery* journal [1].

¹ <http://biomine.cs.helsinki.fi>