

Multilinear Algebra for Analyzing Data with Multiple Linkages

Daniel M. Dunlavy*, Tamara G. Kolda†, and W. Philip Kegelmeyer†

Multi-link graphs, i.e., graphs with multiple link types, are challenging to analyze, yet such data is ubiquitous. For example, social networks clearly have many types of links—familial, communication (phone, email, etc.), organizational, geographical, etc. Our overarching goal is to analyze data with multiple link types, and to derive feature vectors for each individual node (or data object). As a motivating example, we use journal publication data—specifically considering several of the many ways that two papers may be linked. The analysis is applied to five years of journal publication data from eleven journals and a set of conference proceedings published by the Society for Industrial and Applied Mathematics (SIAM). The nodes represent published papers. Explicit, directed links exist whenever one paper cites another. Undirected similarity links are derived based on title, abstract, keyword, and authorship. Historically, bibliometric researchers have focused solely on citation analysis or text analysis, but not both simultaneously. Though this work focuses on the analysis of publication data, the techniques are applicable to a wide range of data analysis tasks.

Link analysis typically focuses on a single link type. For example, two established web analysis techniques, PageRank and HITS, consider the structure of the web and decompose the adjacency matrix of a graph representing the hyperlink structure. Instead of decomposing an adjacency matrix that represents a single link type, our approach is to decompose an adjacency *tensor* that represents multiple link types.

A tensor is a multidimensional, or N -way, array. For multiple linkages, a three-way array can be used where each two-dimensional *frontal slice* represents the adjacency matrix for a single link type. If there are N nodes and K link types, then the data can be represented as a three-way tensor of size $N \times N \times K$ where the (i, j, k) entry is nonzero if node i is connected to node j by link type k . For the SIAM bibliometric data mentioned above, the five different link types correspond to (frontal) slices in a three-way tensor.

The CANDECOMP/PARAFAC (CP) tensor decomposition is a higher-order analogue of the matrix singular value decomposition. In this work, we explore the use of the CP decomposition applied to the adjacency tensor of a multi-link graph for performing data analysis and prediction as follows:

- Revealing “communities” within the data and how they are connected.
- Generating feature vectors for nodes in the graph, which are then compared directly to get a similarity score that combines the multiple linkage types.
- Analyzing a set of feature vectors that represents a *body of work*, e.g., by a given author, to find the most similar papers in the larger collection.
- Using feature vectors for author name disambiguation and resolution.
- Using feature vectors as input to supervised learning methods (decision trees and ensembles) to predict journals that would be good candidates for submitting a manuscript for publication.

*Computer Science and Informatics Department, Sandia National Laboratories, Albuquerque, NM 87185. Research supported in part by the Applied Mathematics Research program of the Office of Advanced Scientific Computing Research of DOE’s Office of Science through a John von Neumann Fellowship at Sandia National Laboratories.

†Informatics and Decision Sciences, Sandia National Laboratories, Livermore, CA 94551.