GRAPH LAYOUT TECHNIQUES AND MULTIDIMENSIONAL DATA ANALYSIS

JAN DE LEEUW
UNIVERSITY OF CALIFORNIA, LOS ANGELES

GEORGE MICHAILIDIS
THE UNIVERSITY OF MICHIGAN

ABSTRACT. In this paper we explore the relationship between multivariate data analysis and techniques for graph drawing or graph layout. Although both classes of techniques were created for quite different purposes, we find many common principles and implementations. We start with a discussion of the data analysis techniques, in particular multiple correspondence analysis, multidimensional scaling, parallel coordinate plotting, and seriation. We then discuss parallels in the graph layout literature.

1. DATA AND GRAPHS

The amount of data and information collected and retained by organizations and businesses is constantly increasing, due to advances in data collection, computerization of transactions and breakthroughs in storage technology. Typically, the applications involve large-scale information banks, such as data warehouses ranging in size into terabytes, that contain interrelated data from a number of sources (e.g. customer and product databases). In order to extract useful information from such large datasets, it is necessary to be able to *identify* patterns, trends and relationships in the data and *visualize* their global structure to facilitate decision making. Graph theoretical concepts are capable of capturing complicated structures and relationships in both numerical and categorical data. In this paper we explore the relationship between multivariate data analysis and techniques for graph drawing or graph layout, and examine how coupling ideas from these two fields can lead to new and improved methodology and tools for mining large databases and presentation of large datasets.

1.1. **Multivariables and Coding.**

The data structure we are interested in consists of n observations on m categorical variables, where variable j has k_j categories (possible values). Using categorical variables causes no real loss of generality: so-called *continuous* variables are merely categorical variables with a large number of numerical categories. We use K for the total number of categories over all variables ($K = \sum_{j=1}^{m} k_j$).