An Example of Visualization in Data Mining

by

Bruce L. Golden

R. H. Smith School of Business

University of Maryland

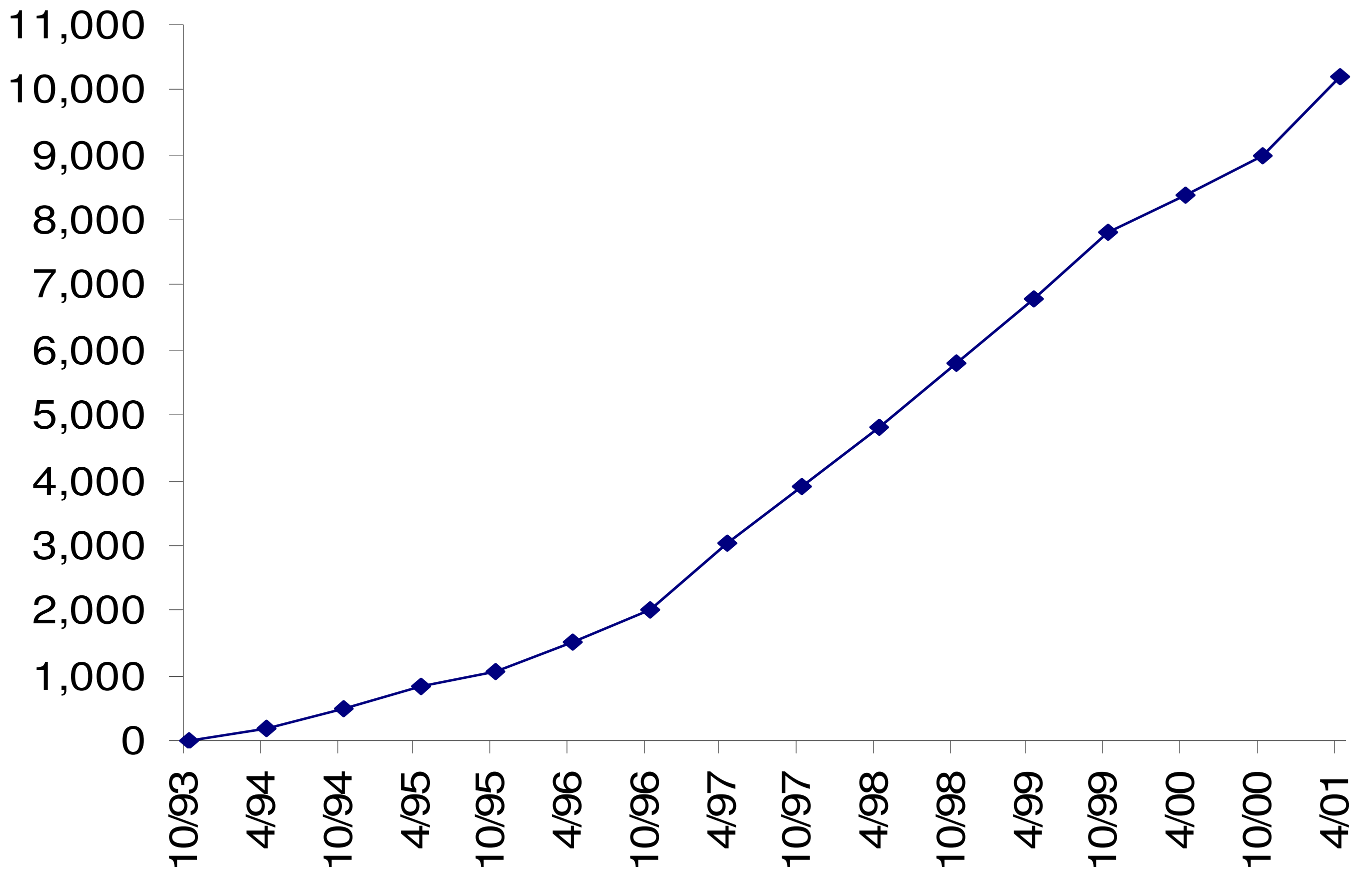
College Park, MD 20742

Presented at Netcentricity Symposium – 3/30/01

Data Mining Overview

- Data mining involves the exploration and analysis of large amounts of data in order to discover meaningful patterns
- The field dates back to a 1989 workshop
- The field has grown dramatically since 1989
  - Data mining software tools (> 200)
  - KDnuggets News, the major e-newsletter in the field, has > 10,000 subscribers
  - Many conferences, courses, and successful applications

Data Mining Overview -- continued



KDnuggets News Subscribers over Time

Data Mining Overview -- continued

- Sample applications

  -- Direct marketing

  -- Telecom

  -- E-commerce

  -- Fraud detection

  -- Customer Relationship Management (CRM)

  -- Text mining

  -- Bioinformatics

- What is the size of the data mining industry ?

Customer Relationship Management

- Powerful new marketing tool
  - Mine data for information about customers
  - Use information to sell more efficiently and design new products
  - Mimic the old days when all shopping was local and shopkeepers knew your name and needs
  - Convert phone calls and web visits to sales

Customer Relationship Management -- continued

- North American market for CRM software will grow from \$3.9B in 2000 to \$11.9B by 2005 (Datamonitor)

- Worldwide spending on CRM will grow from \$23B in 2000 to \$ 40B by the end of 2001 to \$76.3B in 2005 (The Gartner Group)

Focus of Paper

- The focus of this paper will be on a visualization project based on adjacency data (Fiske data)
- The paper illustrates the power of visualization
- Visualization generates insights and impact
- My co-authors on this project are E. Condon, S. Lele, S. Raghavan, and E. Wasil

Motivation

- Typically, data are provided in multidimensional format
  - A large table where the rows represent countries and the columns represent socio-economic variables
- Alternatively, data may be provided in adjacency format
  - Consumers who buy item *a* are likely to buy or consider buying items *b*, *c*, and *d* also
  - Students who apply to college *a* are likely to apply to colleges *b*, *c*, and *d* also

Motivation -- continued

- More on adjacency
  - If the purchase of item i results in the recommendation of item j, then item j is adjacent to item i
  - Adjacency data for n alternatives can be summarized in an n x n adjacency matrix, $A = (a_{ij})$, where

$$a_{ij} = \begin{cases} 1 & \text{if item } j \text{ is adjacent to item } i\text{, and} \\ 0 & \text{otherwise} \end{cases}$$

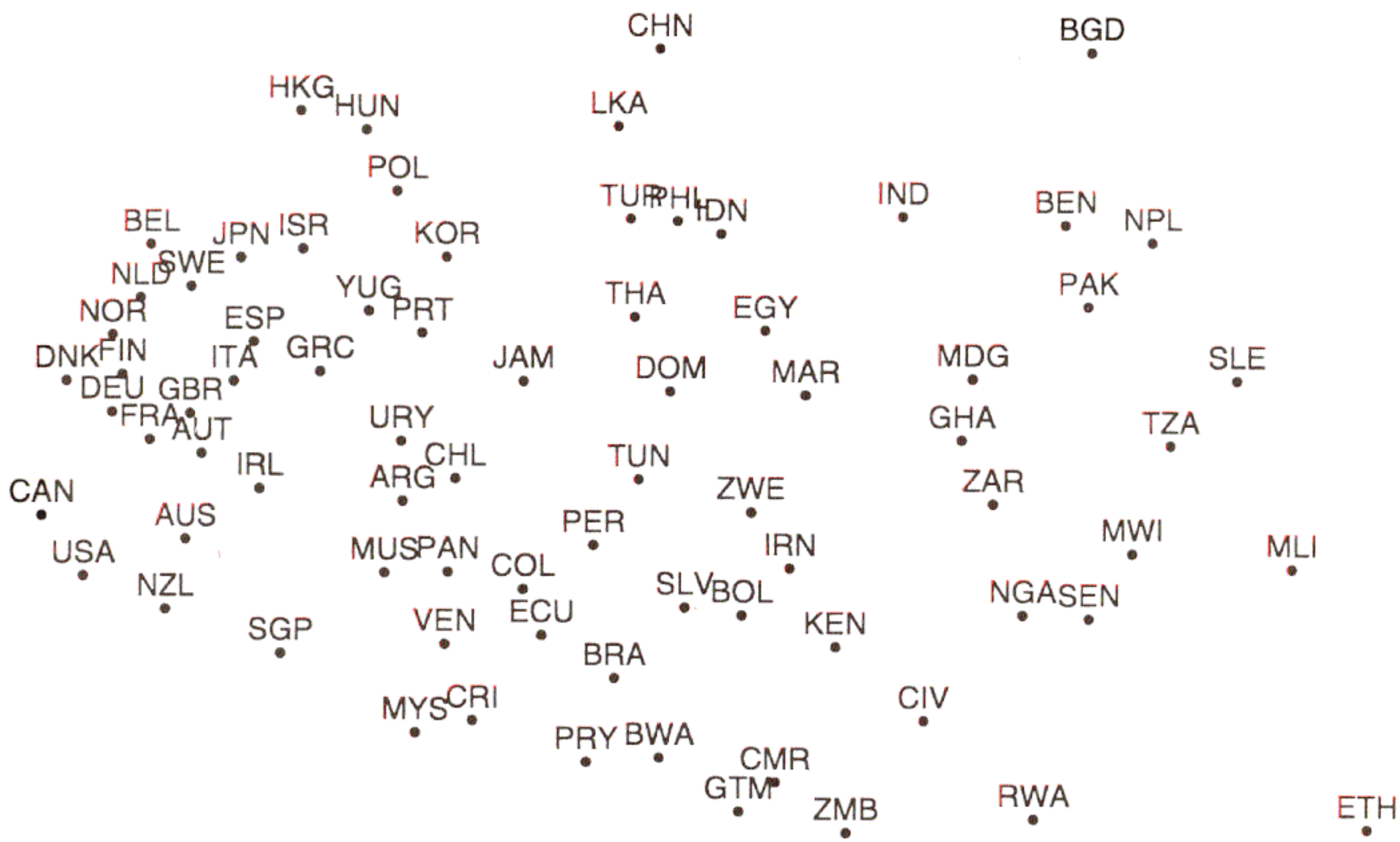
  - Adjacency is not necessarily symmetric

Motivation -- continued

- Adjacency indicates a notion of similarity
- Given adjacency data w.r.t. n items or alternatives, can we display the items in a two-dimensional map?
- Traditional tools such as multidimensional scaling and Sammon maps work well with data in multidimensional format
- Can these tools work well with adjacency data?

Powerful Visualization Techniques

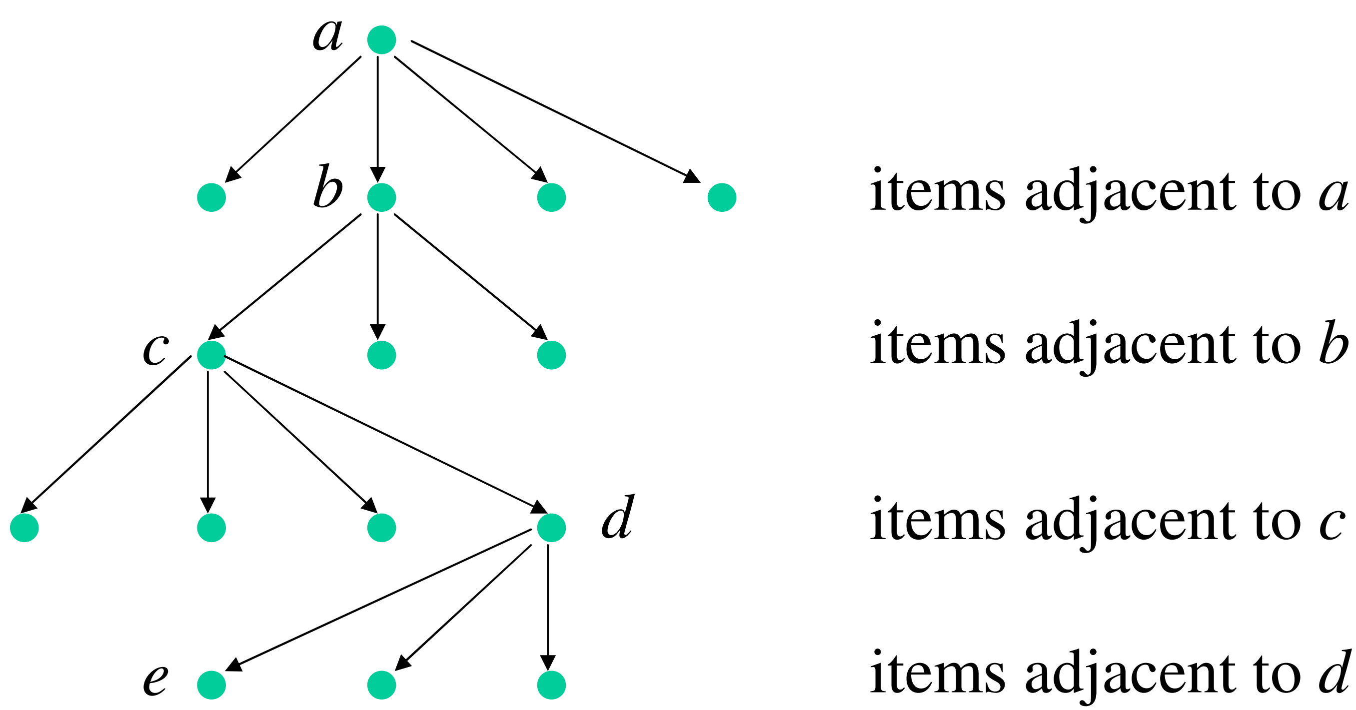
- Multidimensional scaling (MDS)
- Sammon maps
- Both use Euclidean distance (more or less) as a similarity measure
- Euclidean distances typically come from multidimensional format data
- How can we obtain distances from adjacency format data ?



Sammon Map of World Poverty Data Set (World Bank, 1992)

Obtaining Distances from Adjacency Data

- How can we use linkage information to determine distances ?



Obtaining Distances from Adjacency Data -- continued

1. Start with the n x n 0-1 asymmetric adjacency matrix

2. Convert the adjacency matrix to a directed graph
   - Create a node for each item (n nodes)
   - Create a directed arc from node i to node j if $a_{ij} = 1$

3. Compute distance measures
   - Each arc has a length of 1
   - Compute the all-pairs shortest path distance matrix D
   - The distance from node i to node j is d_{ij}

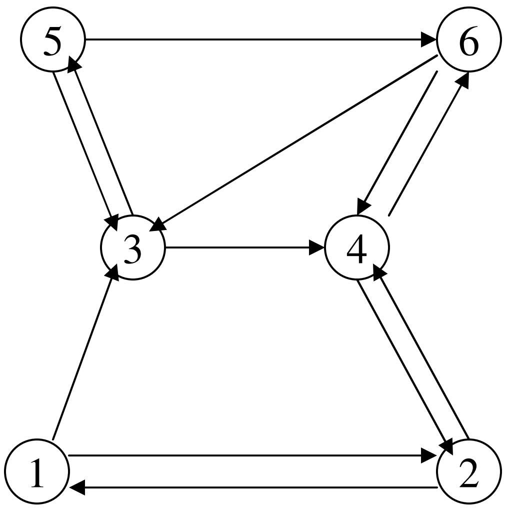
Obtaining Distances from Adjacency Data -- continued

4. Modify the distance matrix *D*, to obtain a final distance matrix *X*
   - Symmetry
   - Disconnected components

- Example 1

$A =$

	1	2	3	4	5	6
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	0	0	0	1	1	0
4	0	1	0	0	0	1
5	0	0	1	0	0	1
6	0	0	1	1	0	0



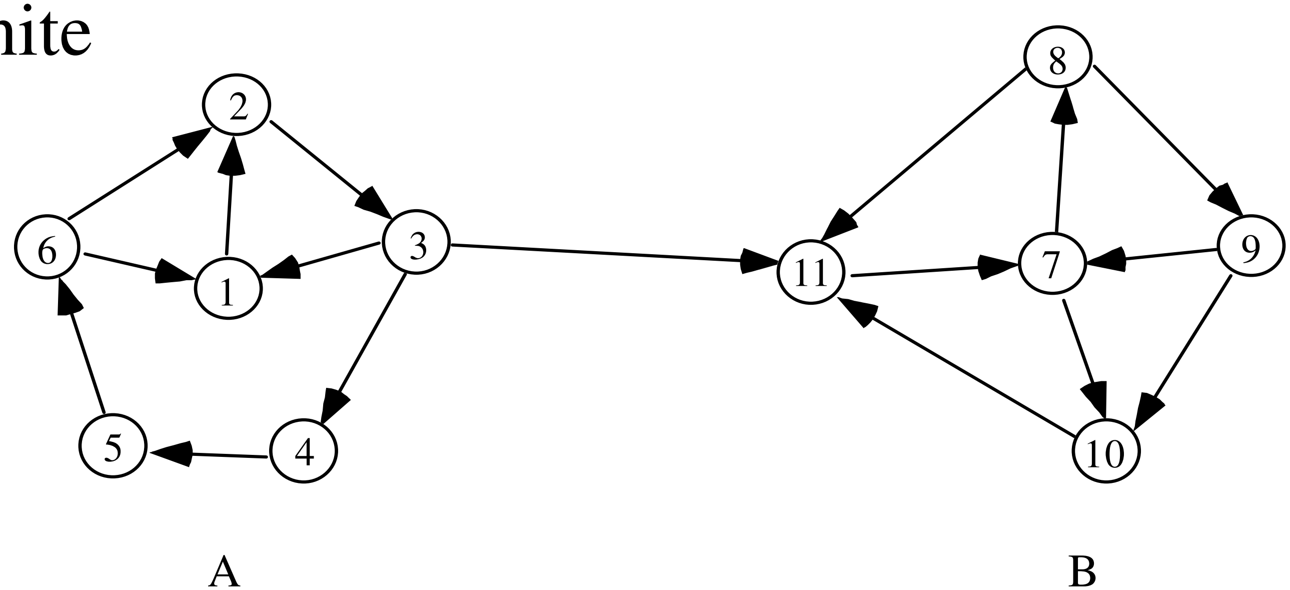
Example 1 -- continued

- Find shortest paths between all pairs of nodes to obtain D
- Average d_{ij} and d_{ji} to arrive at a symmetric distance matrix X

$$D = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 1 & 2 & 2 & 3 \\ 2 & 1 & 0 & 2 & 1 & 3 & 2 \\ 3 & 3 & 2 & 0 & 1 & 1 & 2 \\ 4 & 2 & 1 & 2 & 0 & 3 & 1 \\ 5 & 4 & 3 & 1 & 2 & 0 & 1 \\ 6 & 3 & 2 & 1 & 1 & 2 & 0 \end{array} \qquad X = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 2 & 1 & 0 & 2 & 1 & 3 & 2 \\ 3 & 2 & 2 & 0 & 1.5 & 1 & 1.5 \\ 4 & 2 & 1 & 1.5 & 0 & 2.5 & 1 \\ 5 & 3 & 3 & 1 & 2.5 & 0 & 1.5 \\ 6 & 3 & 2 & 1.5 & 1 & 1.5 & 0 \end{array}$$

Example 2

- A and B are strongly connected components
- The graph below is weakly connected
- There are paths from A to B, but none from B to A
- MDS and Sammon maps require that distances be finite



Ensuring Finite and Symmetric Distances

- Basic idea: simply replace all infinite distances with a large finite value, say R
- If R is too large
  - The points within each strongly connected component will be pushed together in the map
  - Within-component relationships will be difficult to see
- If R is too small
  - Distinct components (e.g., A and B) may blend together in the map

Ensuring Finite and Symmetric Distances -- continued

- R must be chosen carefully (see Technical Report)
- This leads to a finite distance matrix D
- Next, we obtain the final distance matrix *X* where

$$x_{ij} = x_{ji} = \left(d_{ij} + d_{ji}\right) / 2$$

- X becomes input to a Sammon map or MDS procedure

Application: College Selection

- Data source: The Fiske Guide to Colleges, 2000 edition
  - Contains information on 300 colleges
  - Approx. 750 pages
  - Loaded with statistics and ratings
  - For each school, its biggest overlaps are listed
- Overlaps: “the colleges and universities to which its applicants are also applying in greatest numbers and which thus represent its major competitors”

Overlaps and the Adjacency Matrix

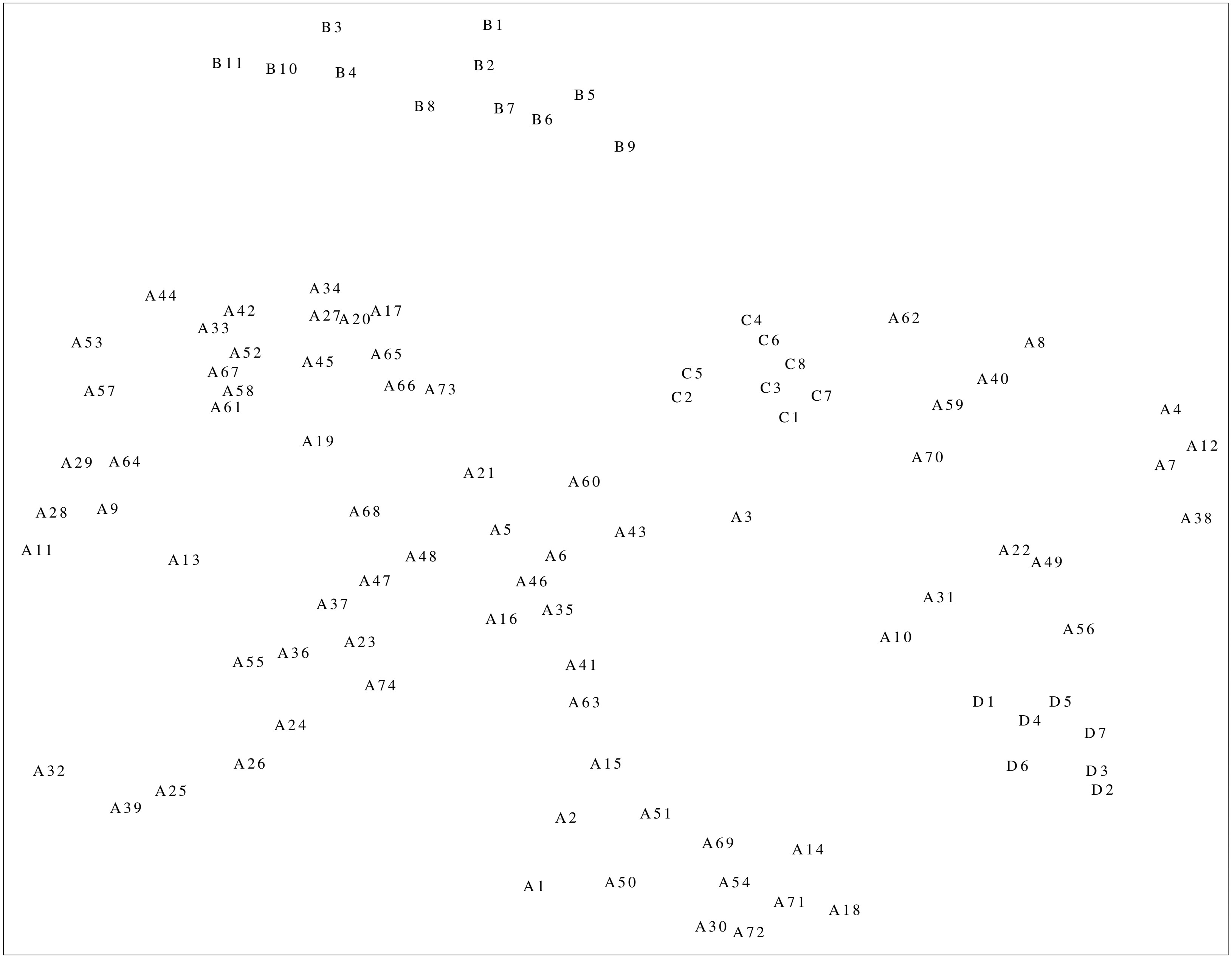
- Penn’s overlaps are Harvard, Princeton, Yale, Cornell, and Brown
- Harvard’s overlaps are Princeton, Yale, Stanford, M.I.T., and Brown
- Note the lack of symmetry
  - Harvard is adjacent to Penn, but not vice versa
- Some clean-up of the overlap data was required
- An illustration of the adjacency matrix follows

Entries in the Adjacency Matrix for a Sample of Eight Schools

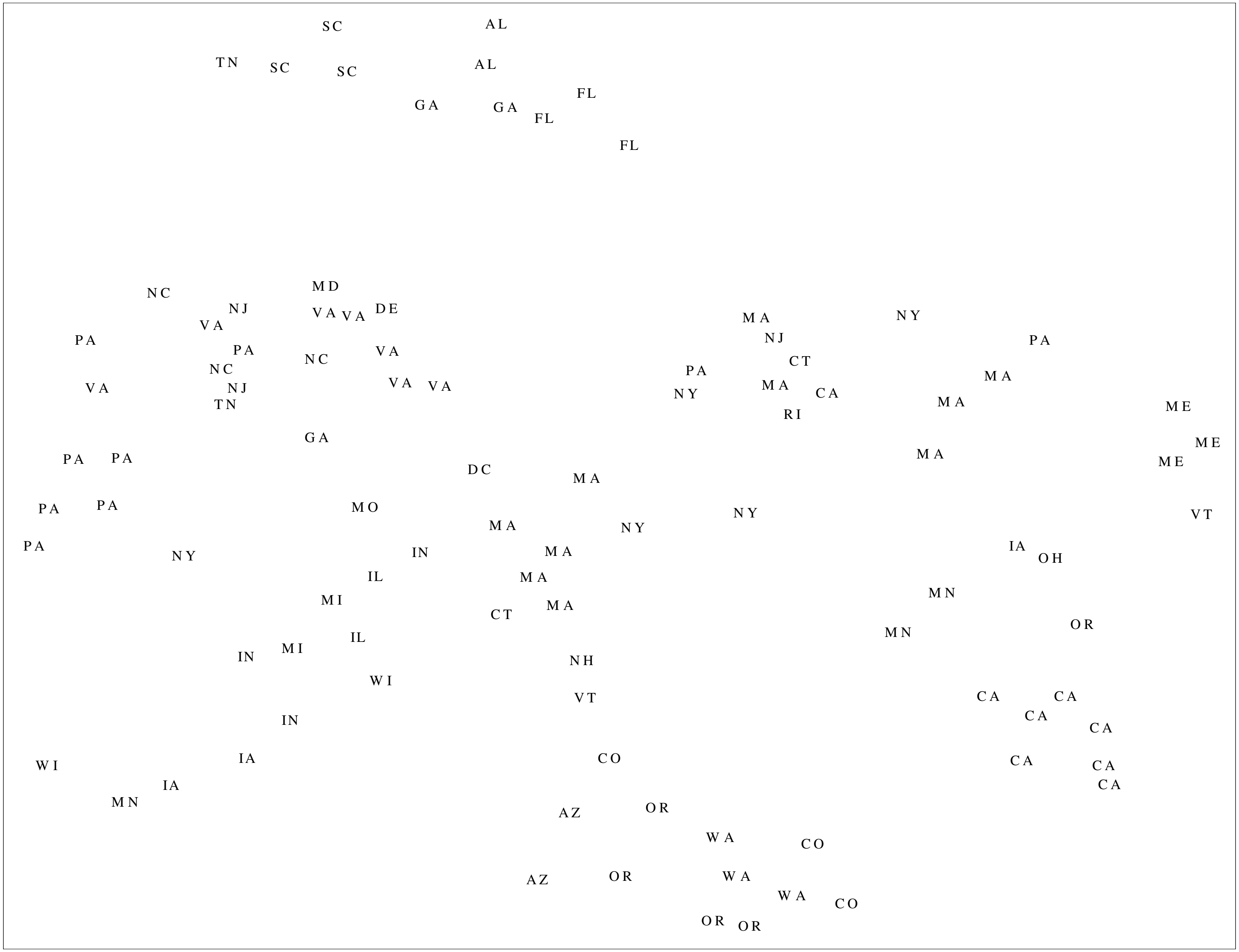
School	Brown	Cornell U.	Harvard	MIT	Penn	Princeton	Stanford	Yale
Brown	0	1	1	0	0	1	1	1
Cornell U.	1	0	1	0	1	1	0	1
Harvard	1	0	0	1	0	1	1	1
MIT	0	1	1	0	0	1	1	1
Penn	1	1	1	0	0	1	0	1
Princeton	0	0	1	1	0	0	1	1
Stanford	1	0	1	1	0	1	0	1
Yale	1	0	1	0	1	1	1	0

Proof of Concept

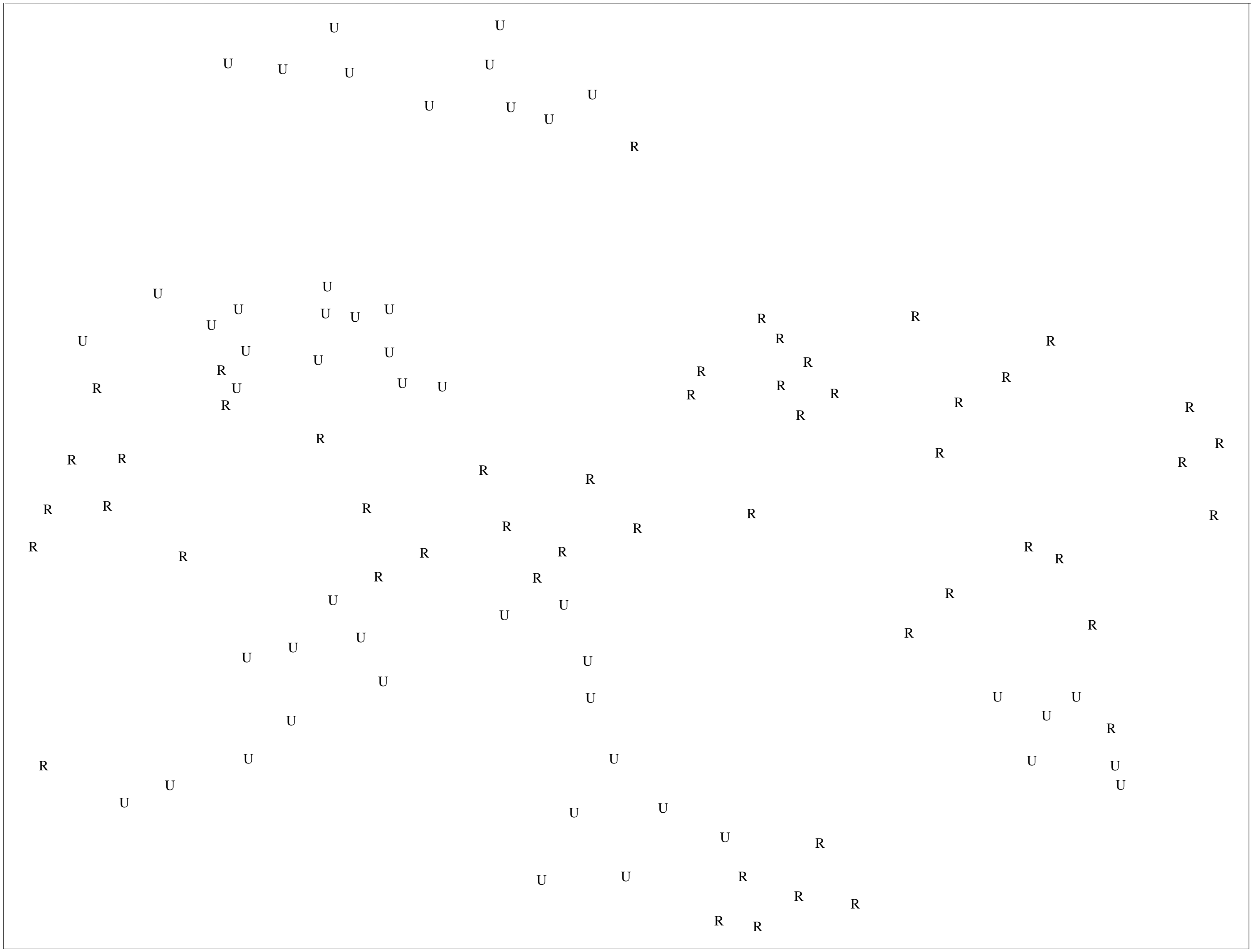
- Start with 300 colleges and the associated adjacency matrix
- From the directed graph, several strongly connected components emerge
- We focus on the four largest to test the concept (100 schools)
  - Component A has 74 schools
  - Component B has 11 southern colleges
  - Component C has 8 mainly Ivy League colleges
  - Component D has 7 California universities



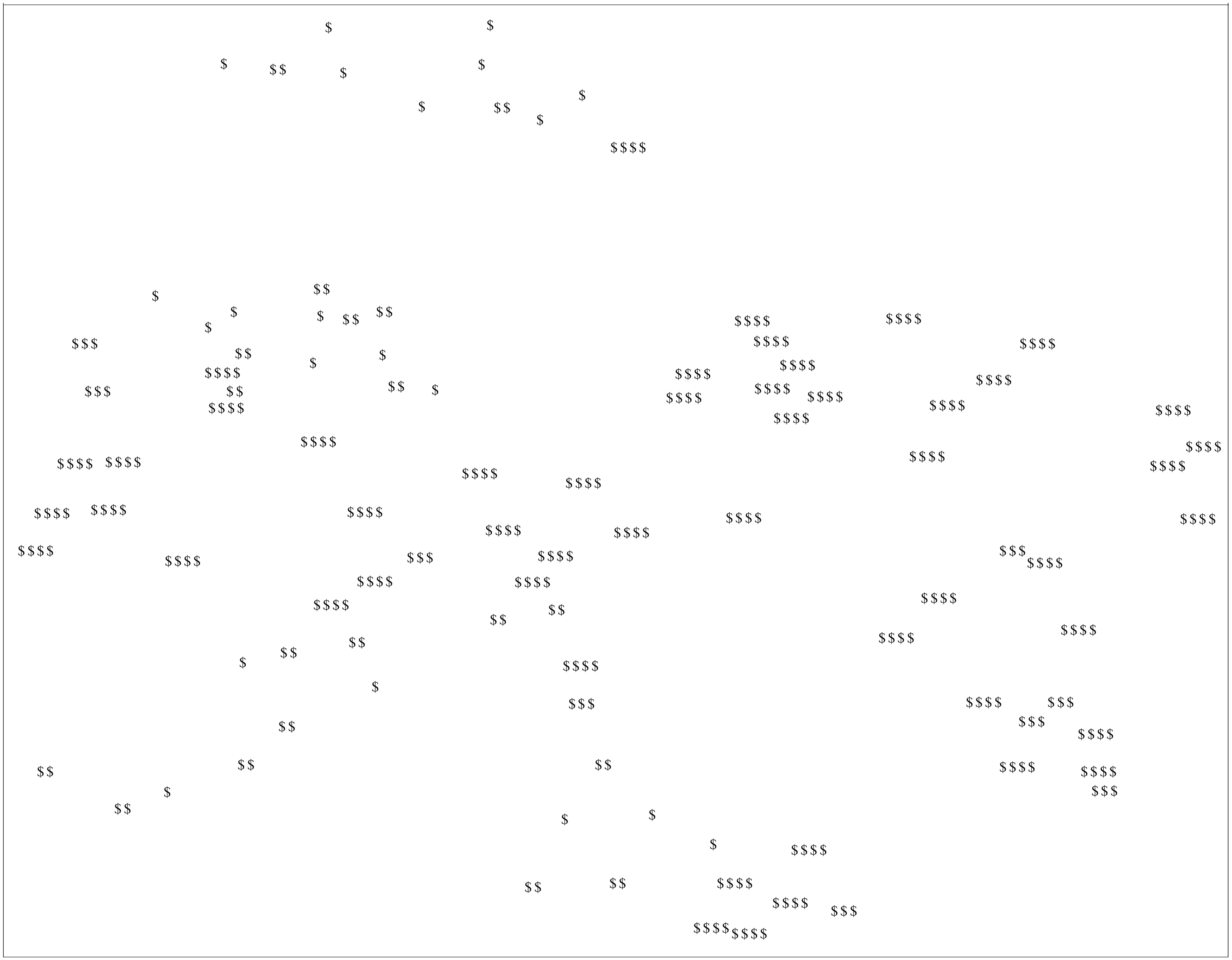
Sammon Map with Each School Labeled by its Component Identifier



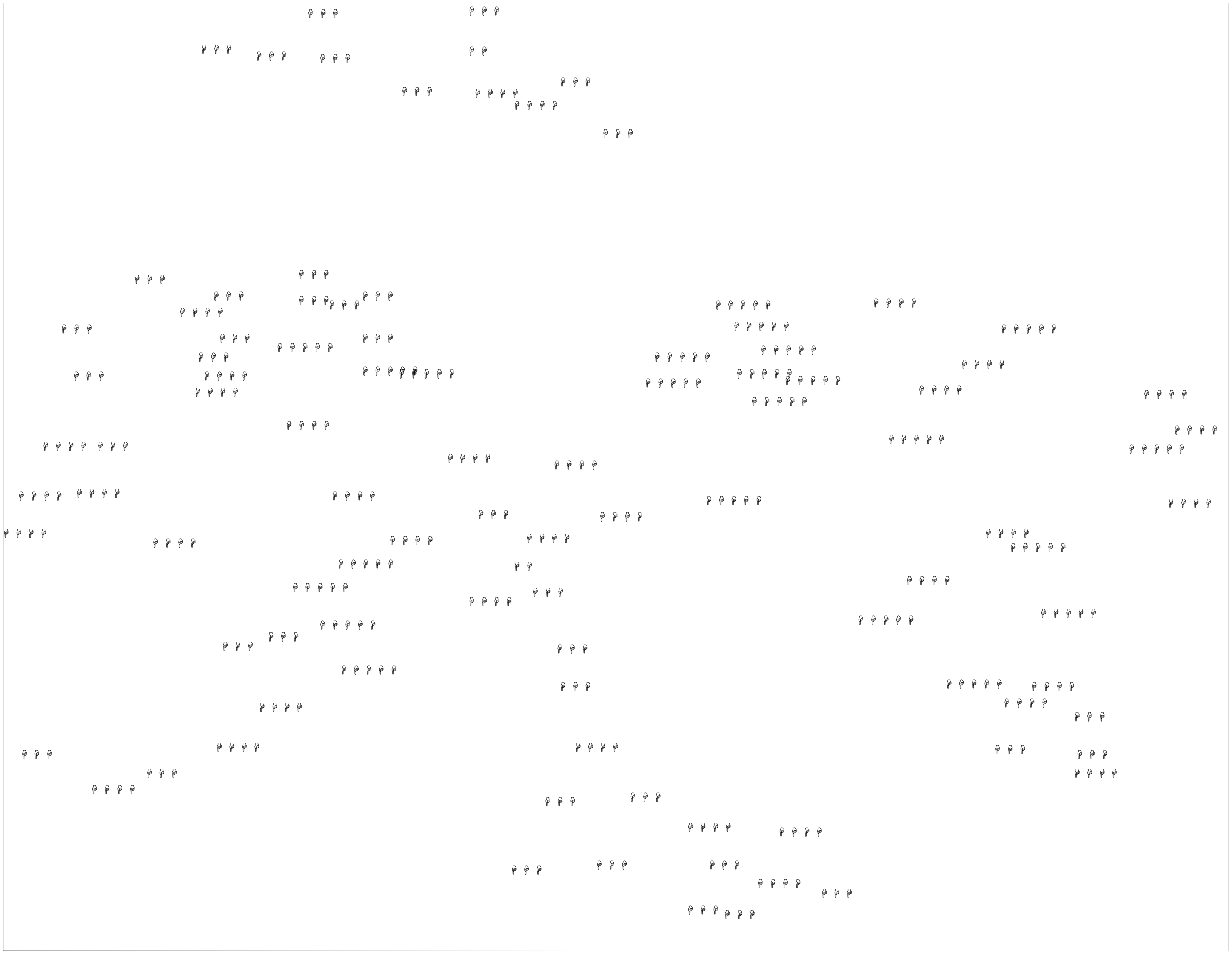
Sammon Map with Each School Labeled by its Geographical Location



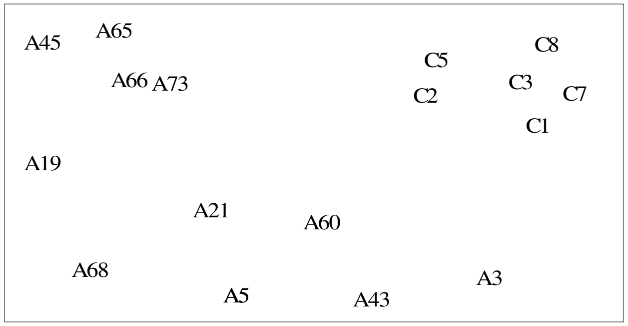
Sammon Map with Each School Labeled by its Designation
(Public (U) or Private (R))



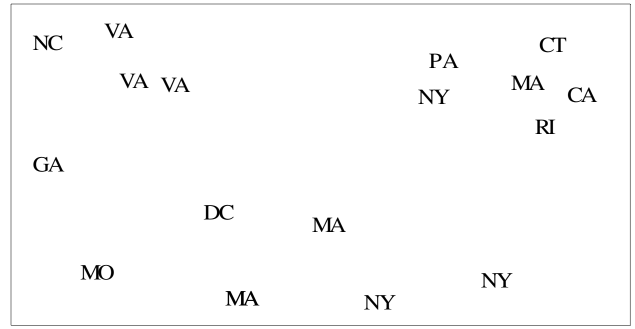
Sammon Map with Each School Labeled by its Cost



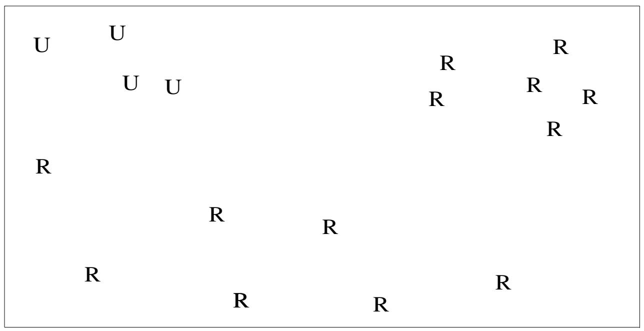
Sammon Map with Each School Labeled by its Academic Quality



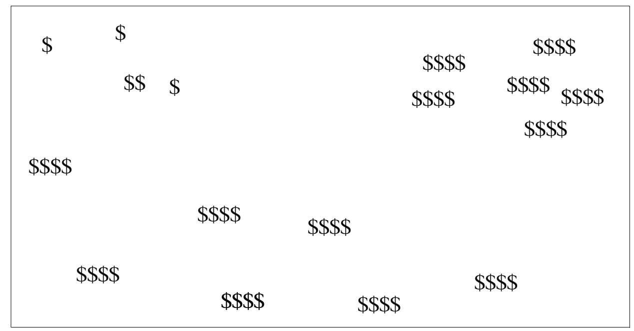
(a) Identifier



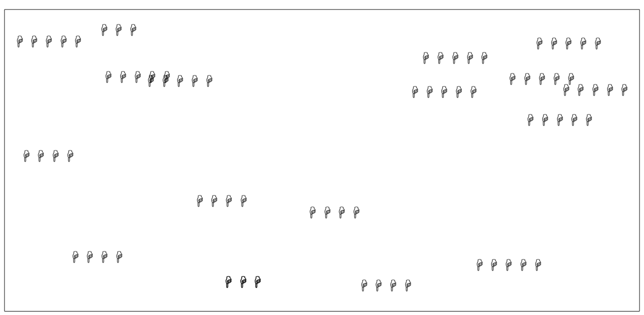
(b) State



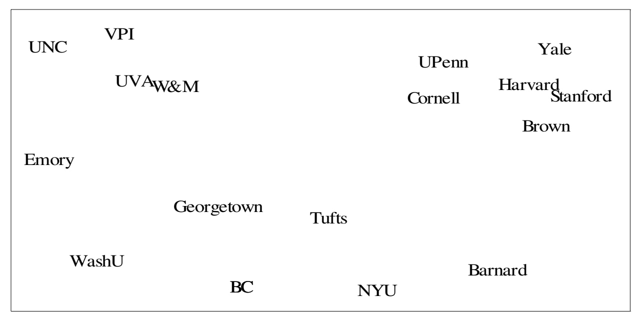
(c) Public or private



(d) Cost



(e) Academics



(f) School name

Six Panels Showing Zoomed Views of Schools that are Neighbors of Tufts University

Benefits of Visualization

- Adjacency (overlap) data provides “local” information only
  - E.g., which schools are Maryland’s overlaps ?
- With visualization, “global” information is more easily conveyed
  - E.g., which schools are similar to Maryland ?

Benefits of Visualization -- continued

- Within group (strongly connected component) and between group relationships are displayed at same time
- A variety of what-if questions can be asked and answered using maps
- Based on this concept, a web-based DSS for college selection is easy to envision

Conclusions

- The approach represents a nice application of shortest paths to data visualization
- The resulting maps convey more information than is immediately available in The Fiske Guide
- Visualization encourages what-if analysis of the data
- Can be applied in other settings (e.g., web-based recommender systems)