Some Examples of Visualization in Data Mining

by

Bruce Golden

R. H. Smith School of Business
University of Maryland

Presented at CORS 2002, Toronto, June 2002

Collaborators

- Ed Condon, University of Maryland
- S. Lele, University of Maryland
- S. Raghavan, University of Maryland
- Edward Wasil, American University

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 > 11,000 subscribers
  - Many conferences, courses, and successful applications

Focus of Paper

- A primary focus of this paper will be on a visualization project based on adjacency data (Fiske data)
- A secondary focus will be on a visualization project based on group decision making data (AHP)
- The paper illustrates the power of visualization
- Visualization generates insights and impact

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

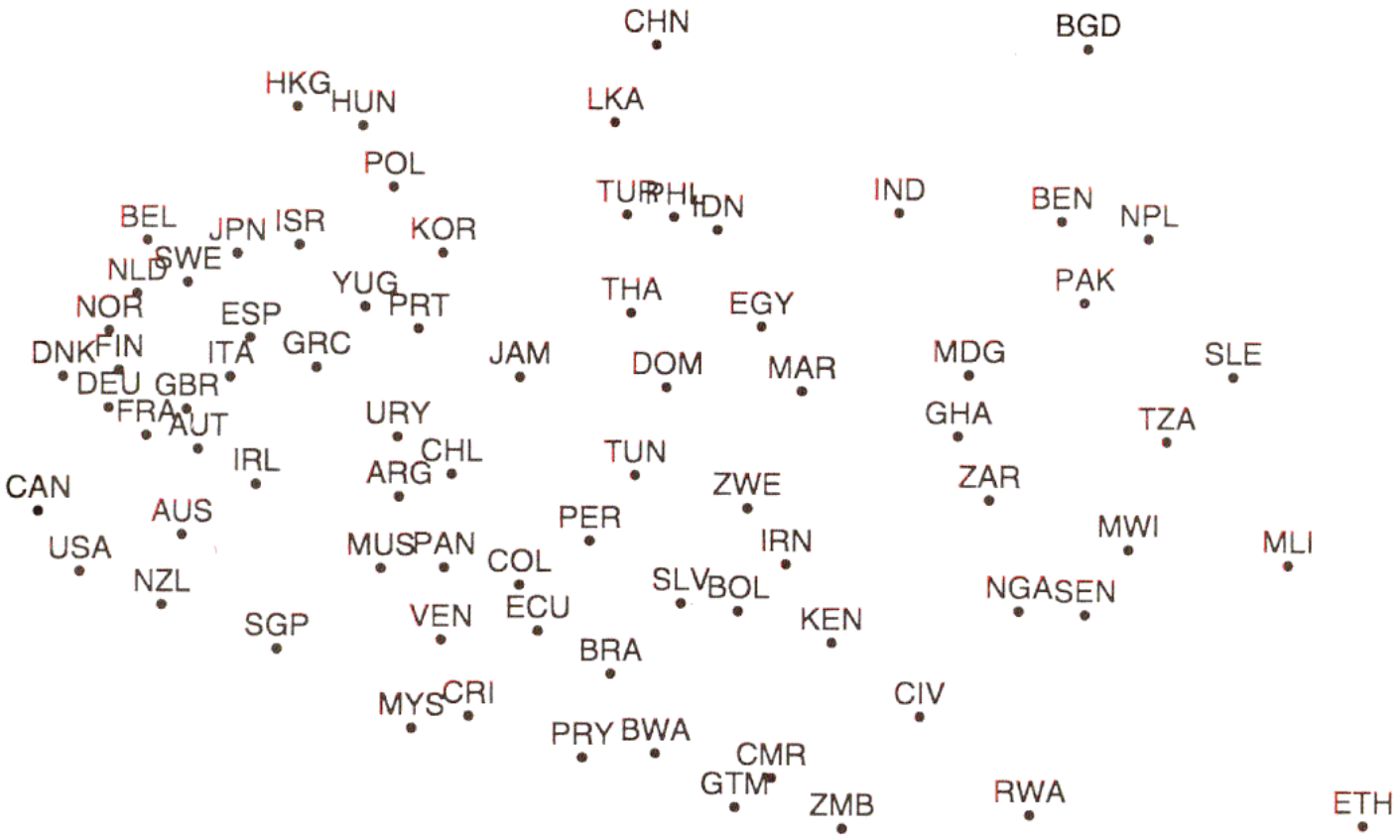
- More on adjacency
  - If the purchase of item *i* results in the recommendation of item *j*, then item *j* is adjacent to item *i*
  - Adjacency data for *n* alternatives can be summarized in an *n* x *n* adjacency matrix, $A = (a_{ij})$, where

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

  - 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?



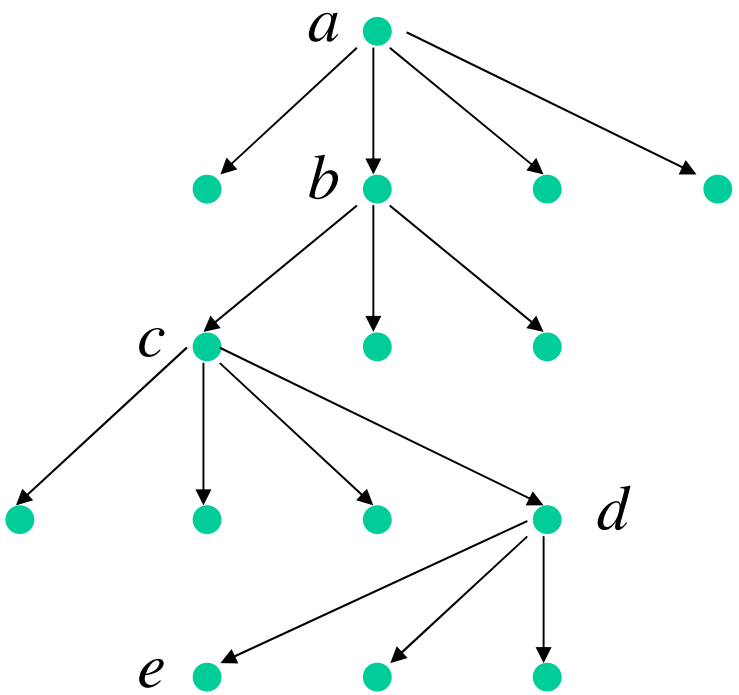
Sammon Map of World Poverty Data Set (World Bank, 1992)

Optimization & Sammon Maps

- Given distance d_{ij} between i and j (for all i, j) in n dimensions (input)
- Let f_{ij} be the distance between i and j (for all i, j) in two dimensions (output)
- We seek to minimize $\sum_{i<j} (d_{ij} - f_{ij})^2 / d_{ij}$
- Solution techniques include gradient descent, smoothing, and simulated annealing

Obtaining Distances from Adjacency Data

- How can we use linkage information to determine distances ?



items adjacent to *a*

items adjacent to *b*

items adjacent to *c*

items adjacent to *d*

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}

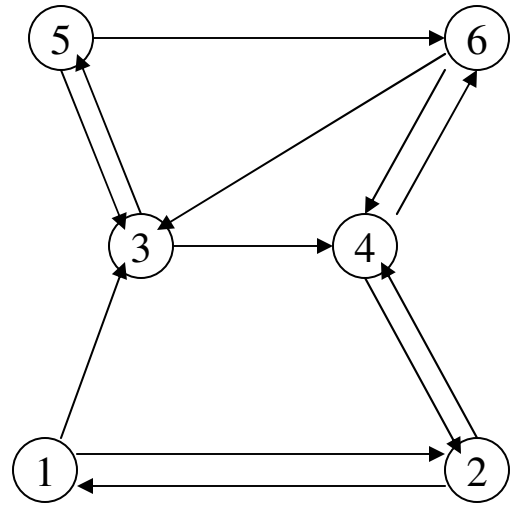
Obtaining Distances from Adjacency Data -- continued

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

$D =$

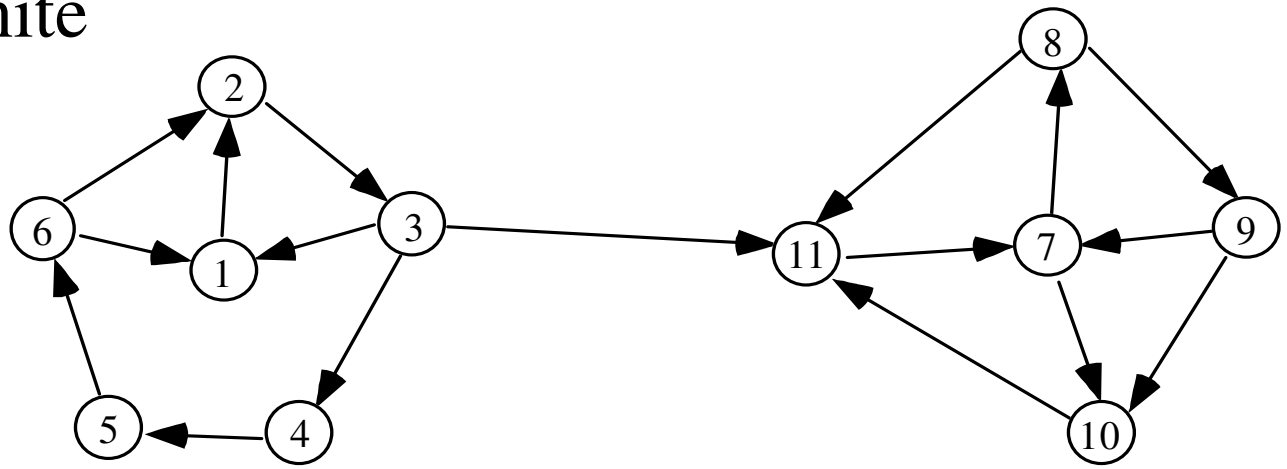
	1	2	3	4	5	6
1	0	1	1	2	2	3
2	1	0	2	1	3	2
3	3	2	0	1	1	2
4	2	1	2	0	3	1
5	4	3	1	2	0	1
6	3	2	1	1	2	0

$X =$

	1	2	3	4	5	6
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

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



A B

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

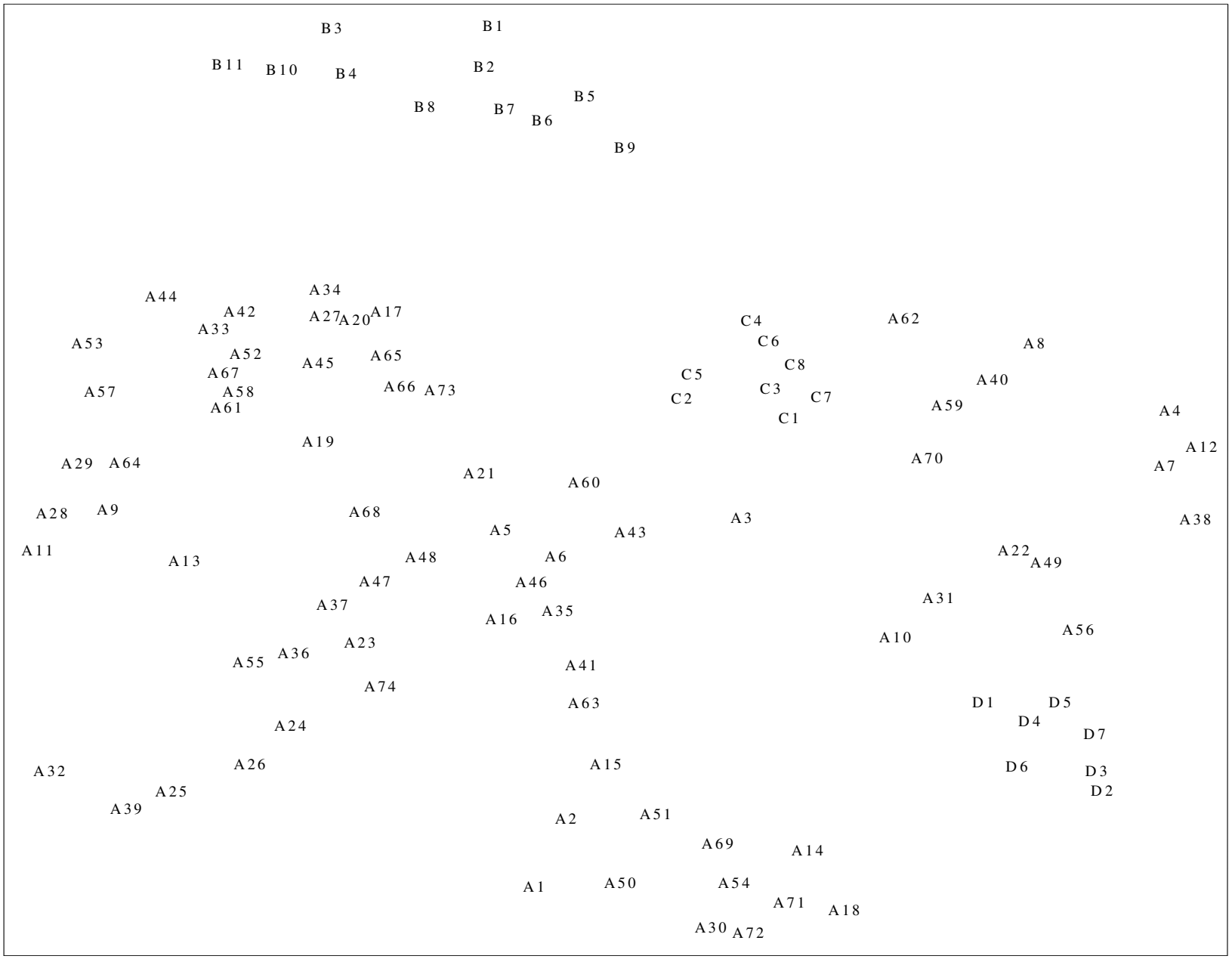
- Data source: The Fiske Guide to Colleges, 2000 edition
  - Contains information on 300 colleges
  - Approx. 750 pages
  - Loaded with statistics and ratings
  - For each school, its biggest overlaps are listed
- Overlaps: "the colleges and universities to which its applicants are also applying in greatest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

Overlaps and the Adjacency Matrix

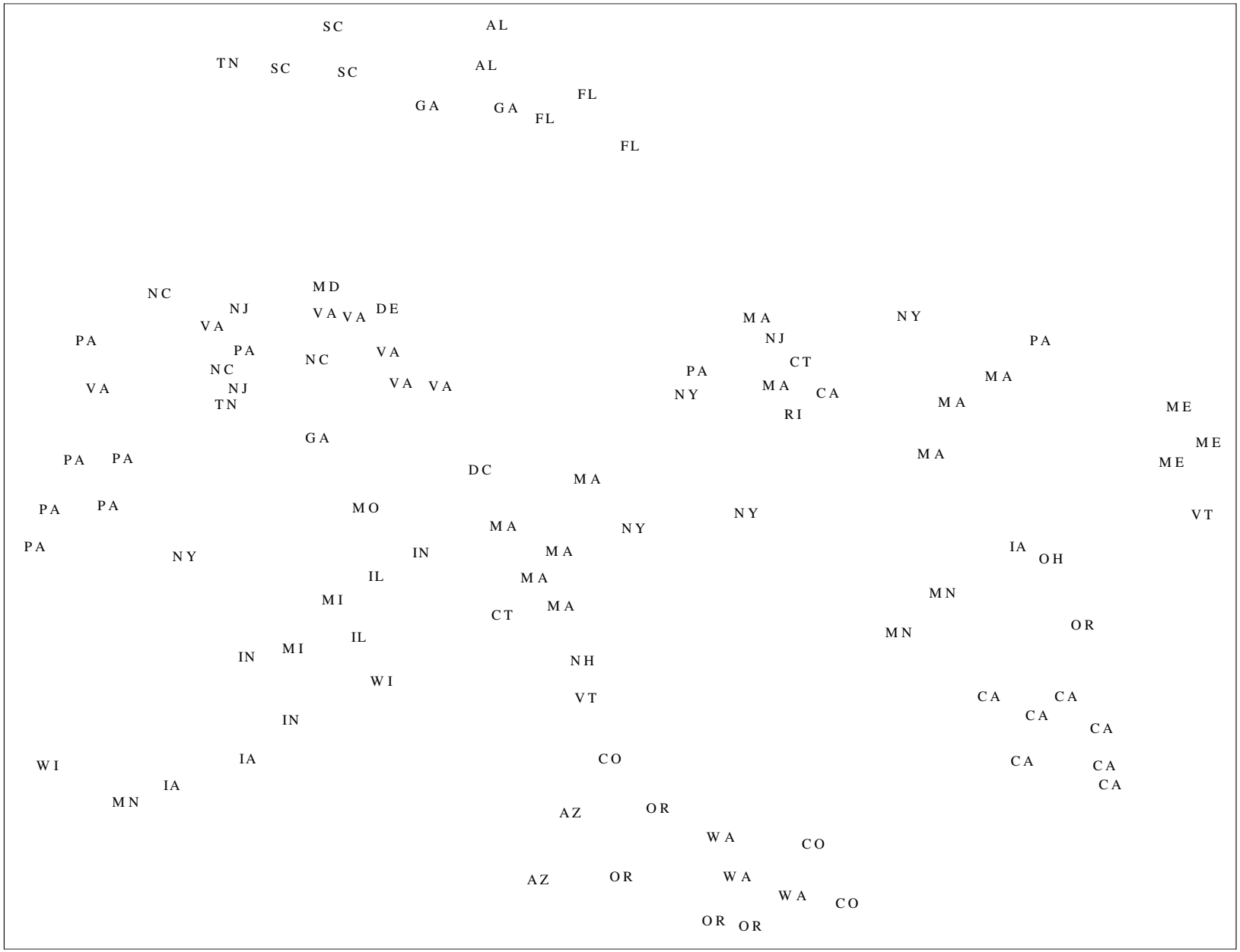
- Penn's overlaps are Harvard, Princeton, Yale, Cornell, and Brown
- Harvard's overlaps are Princeton, Yale, Stanford, M.I.T., and Brown
- Note the lack of symmetry
  - Harvard is adjacent to Penn, but not vice versa

Proof of Concept

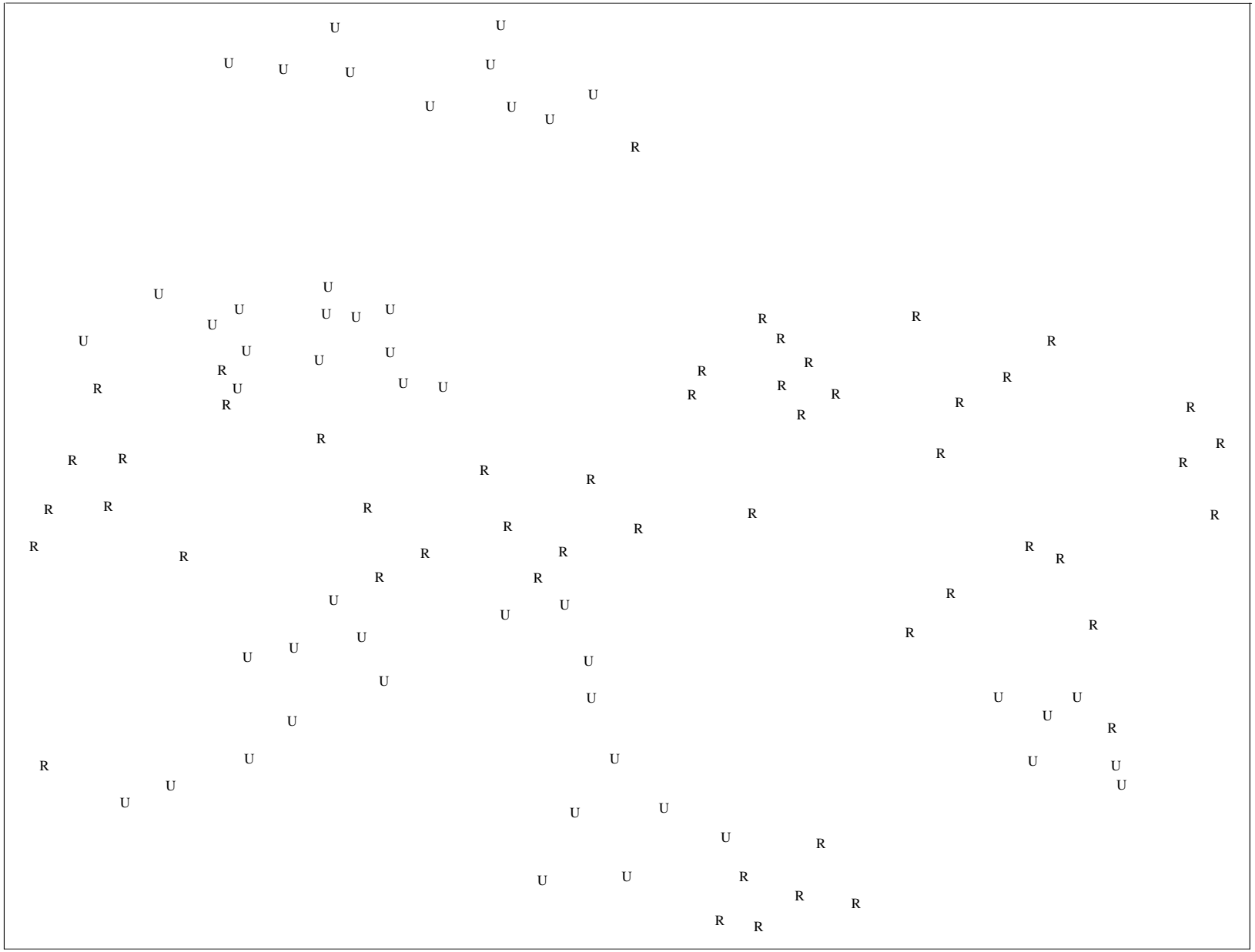
- Start with 300 colleges and the associated adjacency matrix
- From the directed graph, several strongly connected components emerge
- We focus on the four largest to test the concept (100 schools)
  - Component A has 74 schools
  - Component B has 11 southern colleges
  - Component C has 8 mainly Ivy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



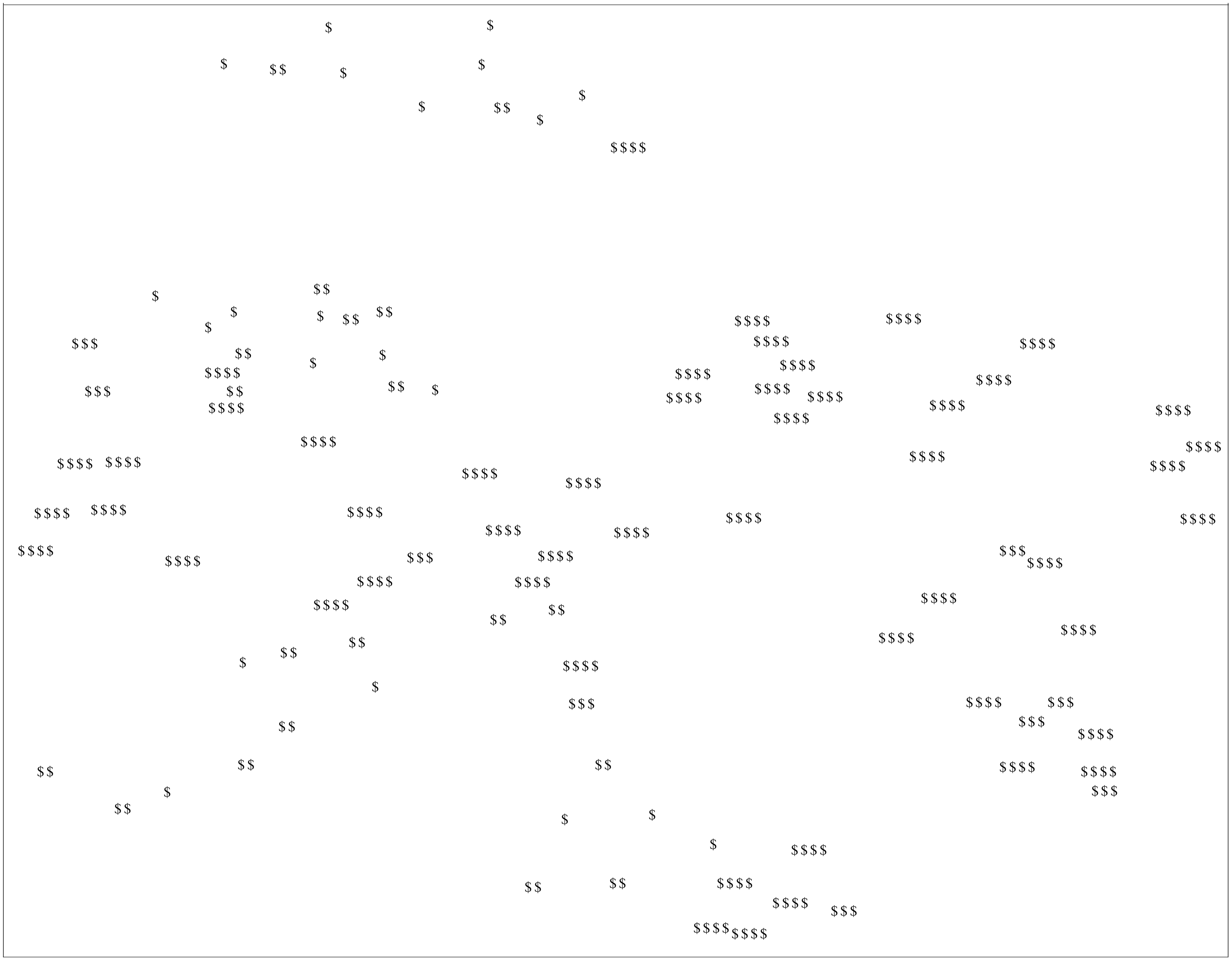
Sammon Map with Each School Labeled by its Component Identifier



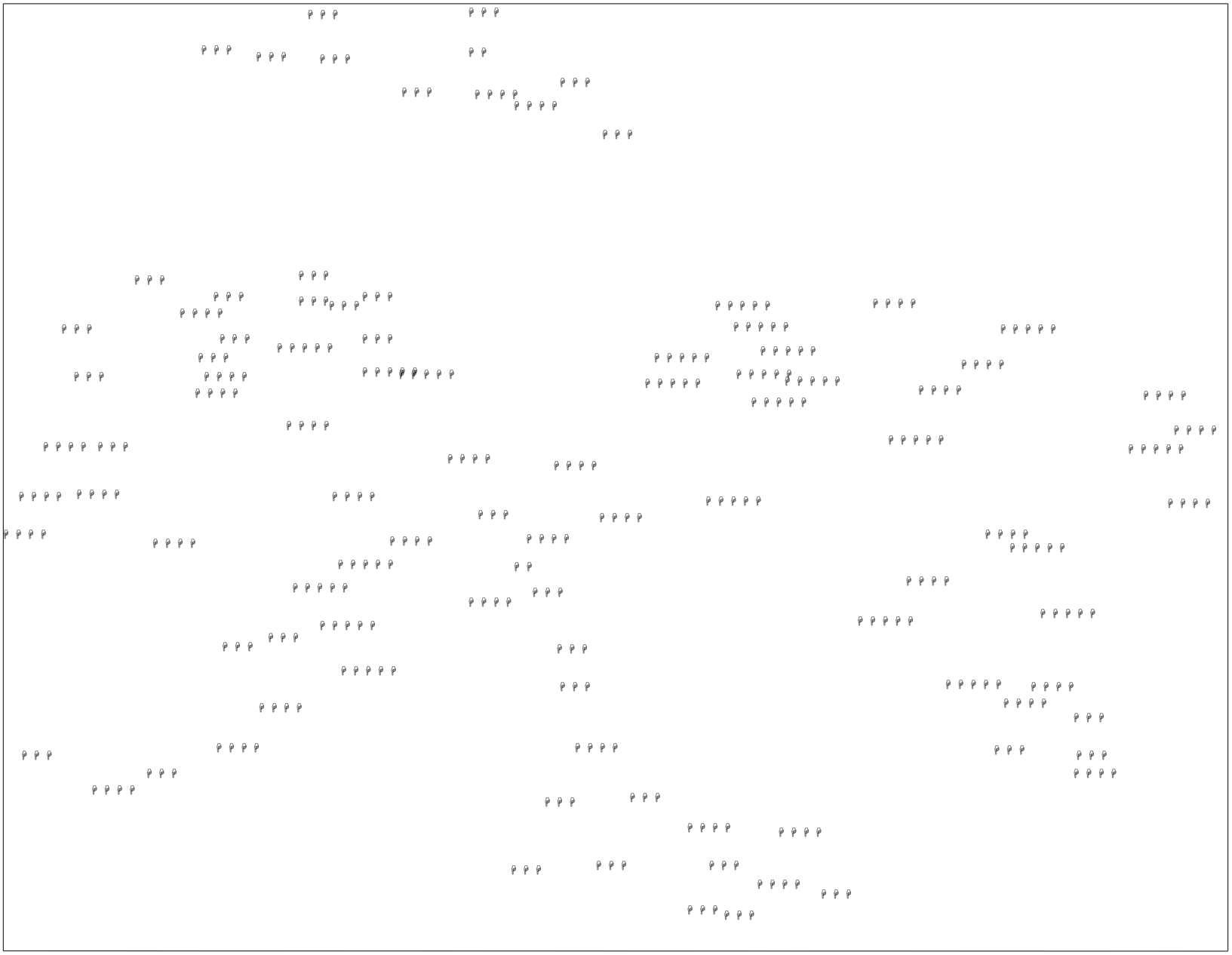
Sammon Map with Each School Labeled by its Geographical Location



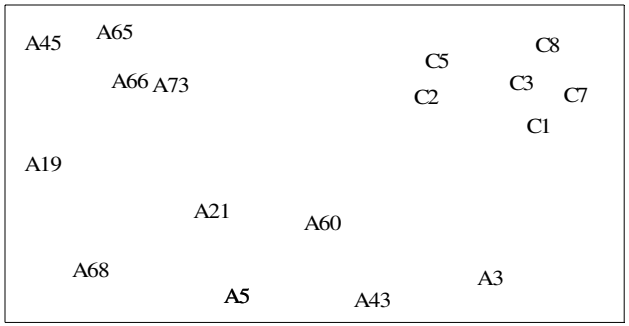
Sammon Map with Each School Labeled by its Designation
(Public (U) or Private (R))



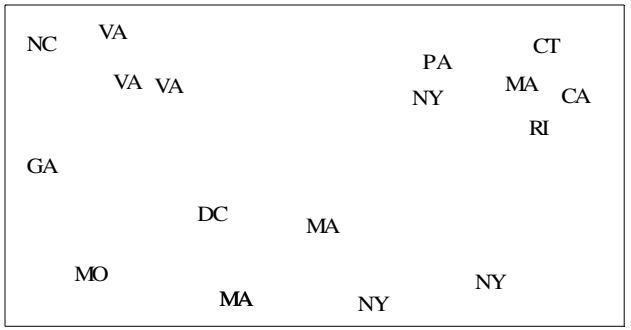
Sammon Map with Each School Labeled by its Cost



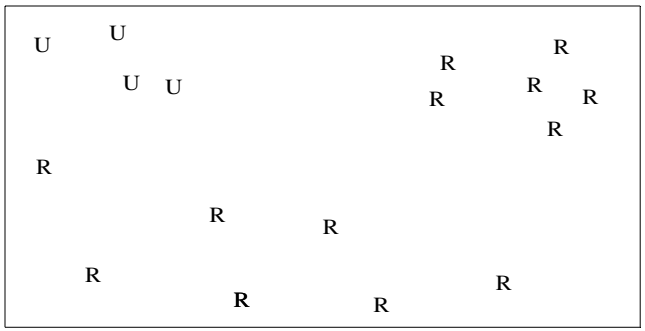
Sammon Map with Each School Labeled by its Academic Quality



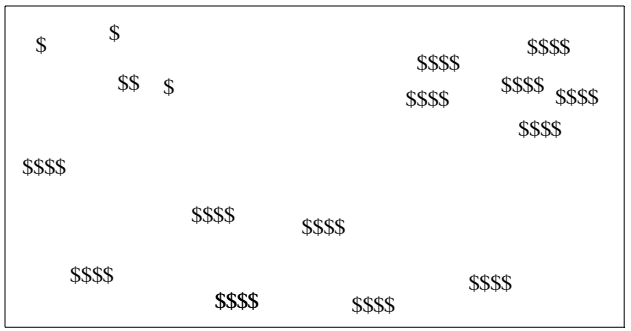
(a) Identifier



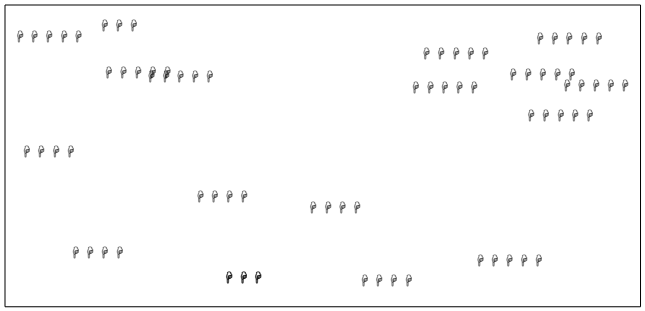
(b) State



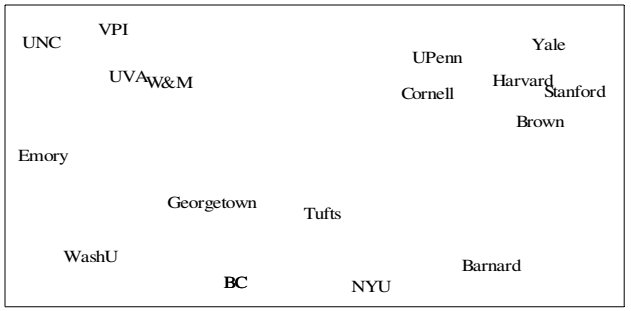
(c) Public or private



(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 ?

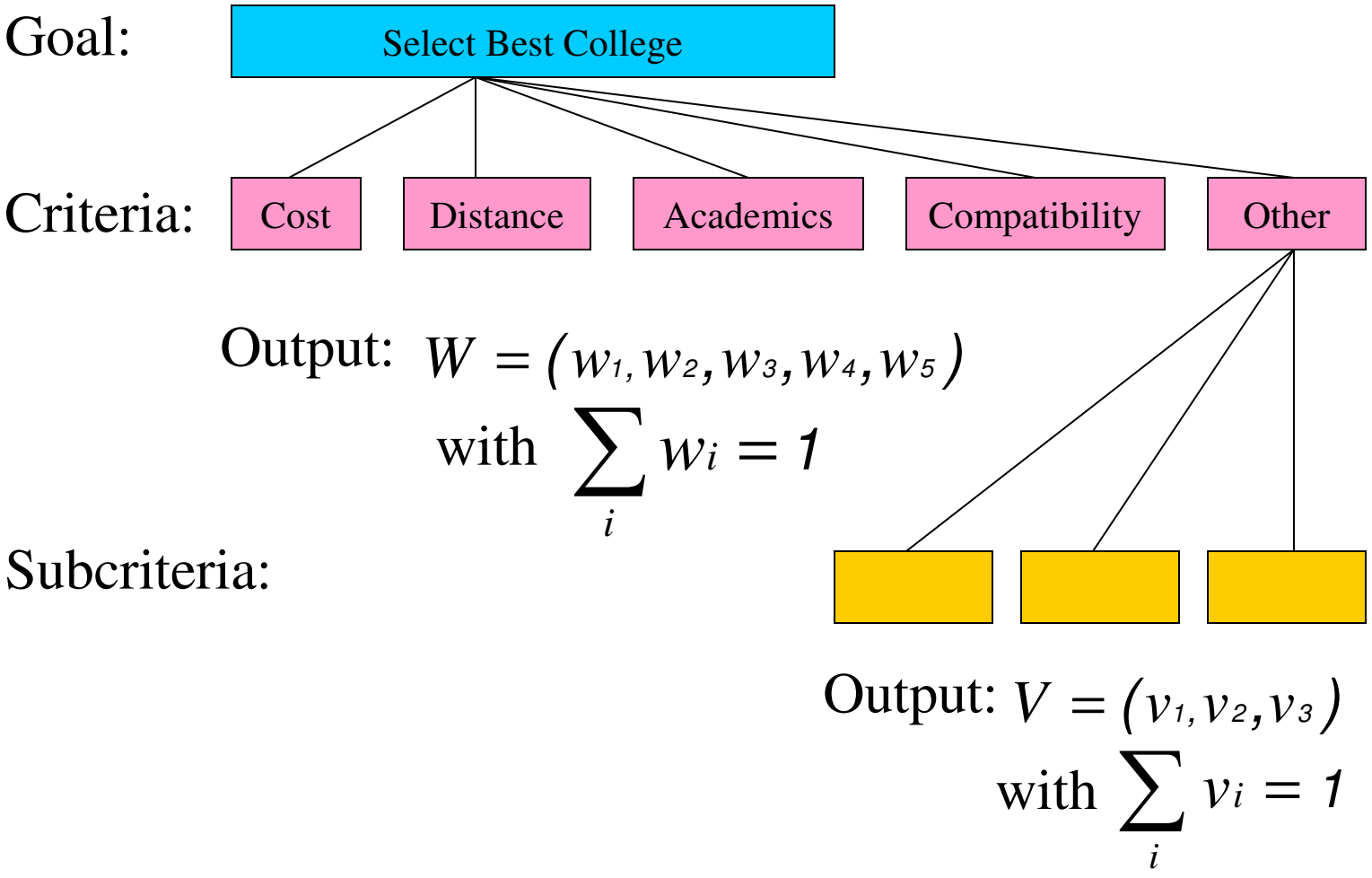
Benefits of Visualization -- continued

- 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: Part One

- 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)

AHP at a Glance



AHP at a Glance -- continued

- The vector of weights *W* indicates the relative importance of criteria with respect to the goal
- The vector of weights *V* indicates the relative importance of subcriteria with respect to the criteria
- In group AHP, some decision makers may distort preferences in order to enhance a hidden agenda
- Visualization makes this more difficult

Group Decision Making with AHP

- Find examples of group decision making with AHP in the research literature
- Address the question: Is it possible to visualize the priorities of the various decision makers ?
- Demonstrate this connection between data mining and AHP on several data sets
- What insights are provided by visualization ?

The Challenge: Getting the Group to Agree

- Alternative approaches (first two due to Saaty)
  - consensus agreement on each matrix entry
  - combine individual judgments using the geometric mean, e.g.,

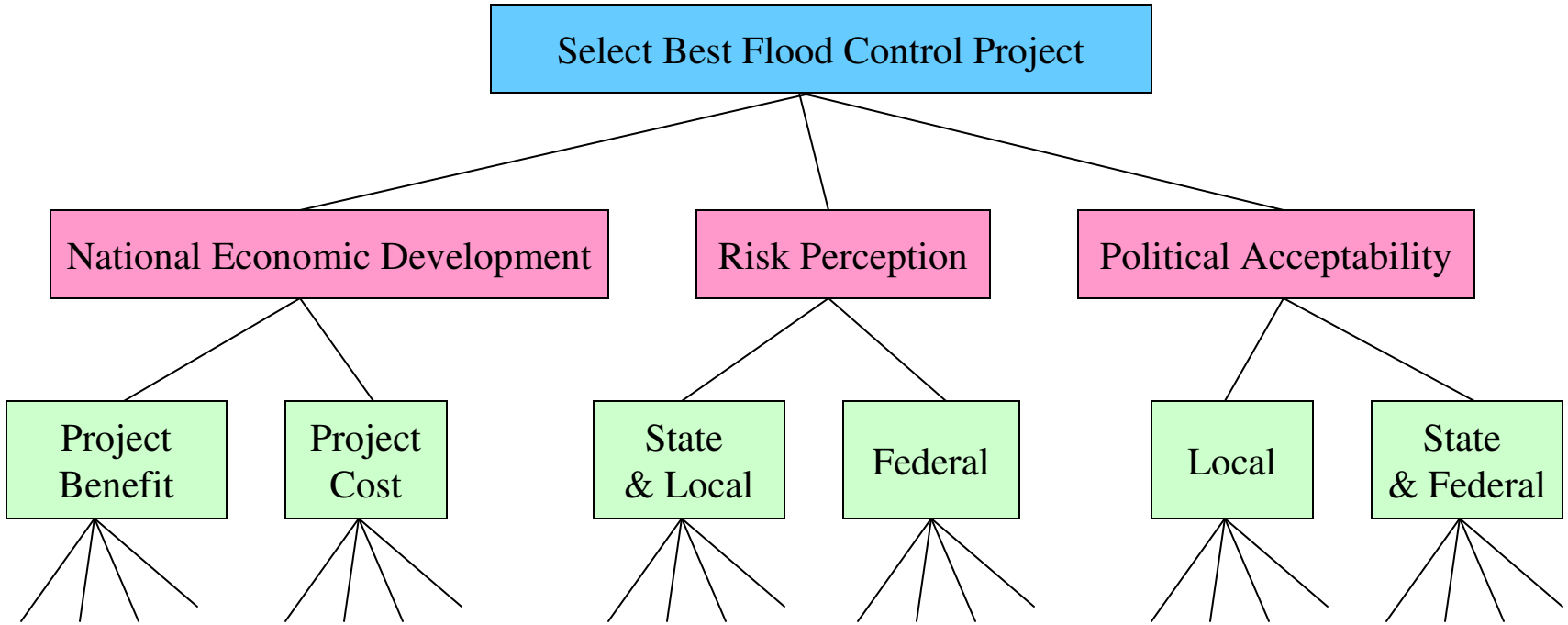
$$a_{12} = \left[a_{12}^{1} \times a_{12}^{2} \times \ldots \times a_{12}^{N} \right]^{1/N}$$

  - apply the geometric mean to eigenvectors (see Willett and Sharda, 1991)

Two Data Sets

- Water resource planning: selection of flood control projects
  - authors: Willett, Sharda
  - source: Socio – Economic Planning Sciences (1991)
  - 7 decision makers
  - 4 comparison matrices plus overall ratings
- Selection of alternatives in a U.S. Governmental Agency
  - project director: Daniel Saaty
  - source: personal communication
  - 8 decision makers
  - 6 comparison matrices

Data Set # 1



Alternatives

50 yr. levee

100 yr. levee

acquisition plan

status quo

Eigenvectors and Geometric Means

Flood Control Criteria Ratings

	NED	RiskPerc	PolAcc
Econ1	0.670	0.209	0.121
Econ2	0.746	0.193	0.060
Econ3	0.397	0.500	0.102
EngPM	0.321	0.495	0.182
EngSH	0.772	0.102	0.125
SciEnv	0.622	0.247	0.130
SciSoc	0.109	0.581	0.309
GM	0.517	0.330	0.153

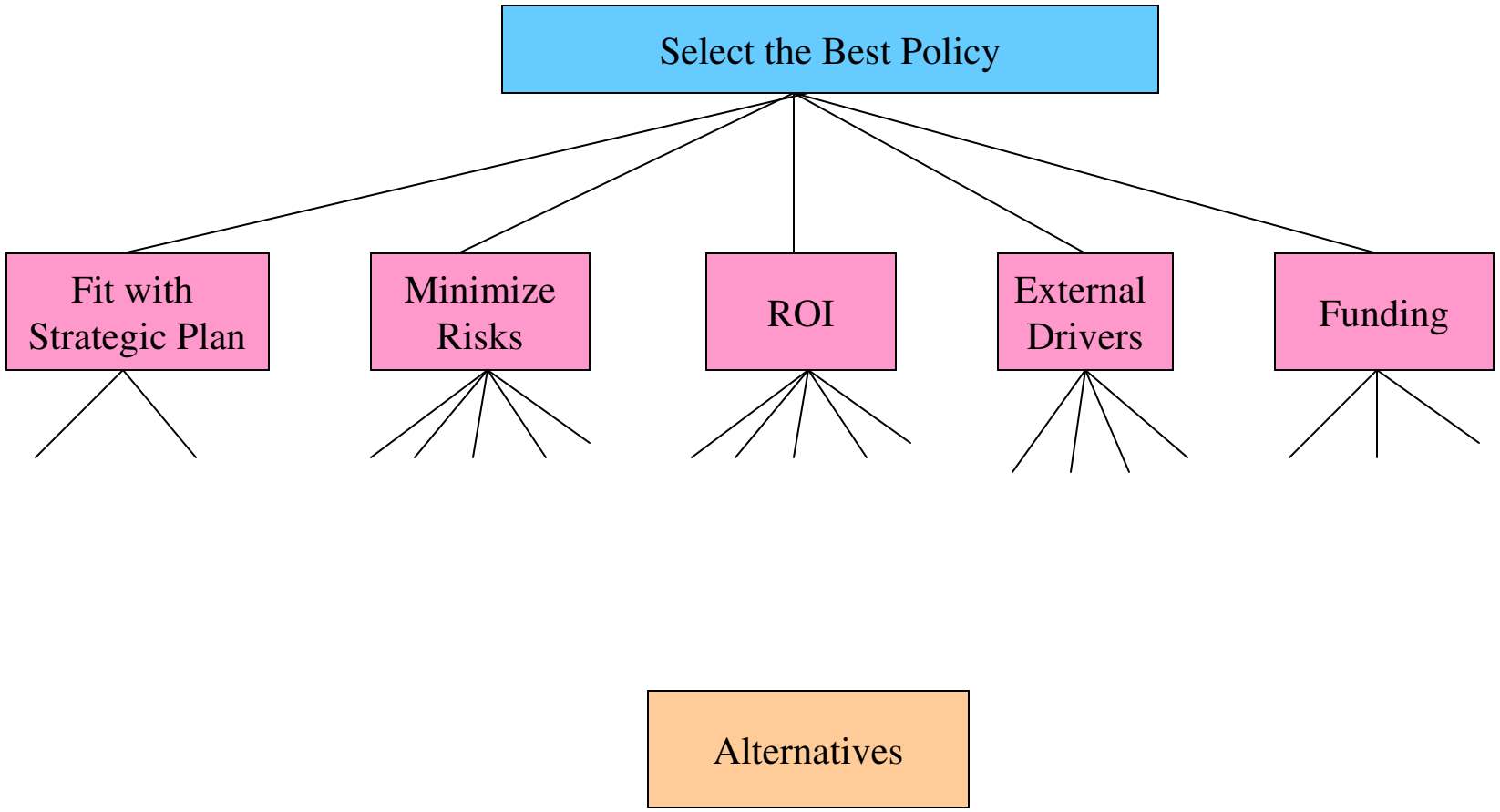
Flood Control Subcriteria Ratings

	ProjBen	ProjCost	S&LRisk	FedRisk	LPolAcc	S&FPolAcc
Econ1	0.500	0.500	0.800	0.200	0.857	0.142
Econ2	0.637	0.362	0.888	0.111	0.750	0.250
Econ3	0.609	0.390	0.274	0.725	0.343	0.656
EngPM	0.817	0.187	0.142	0.857	0.875	0.125
EngSH	0.500	0.500	0.125	0.875	0.857	0.142
SciEnv	0.333	0.666	0.750	0.250	0.200	0.800
SciSoc	0.657	0.342	0.660	0.339	0.500	0.500
GM	0.586	0.414	0.519	0.481	0.659	0.341

Flood Control Alternatives

	50yr levee	100yr levee	Acquisition	StatusQuo
Econ1	0.409	0.388	0.159	0.043
Econ2	0.249	0.633	0.067	0.049
Econ3	0.303	0.541	0.118	0.037
EngPM	0.331	0.466	0.120	0.081
EngSH	0.386	0.401	0.187	0.025
SciEnv	0.243	0.357	0.271	0.127
SciSoc	0.265	0.578	0.097	0.058
GM	0.318	0.489	0.138	0.055

Data Set # 2

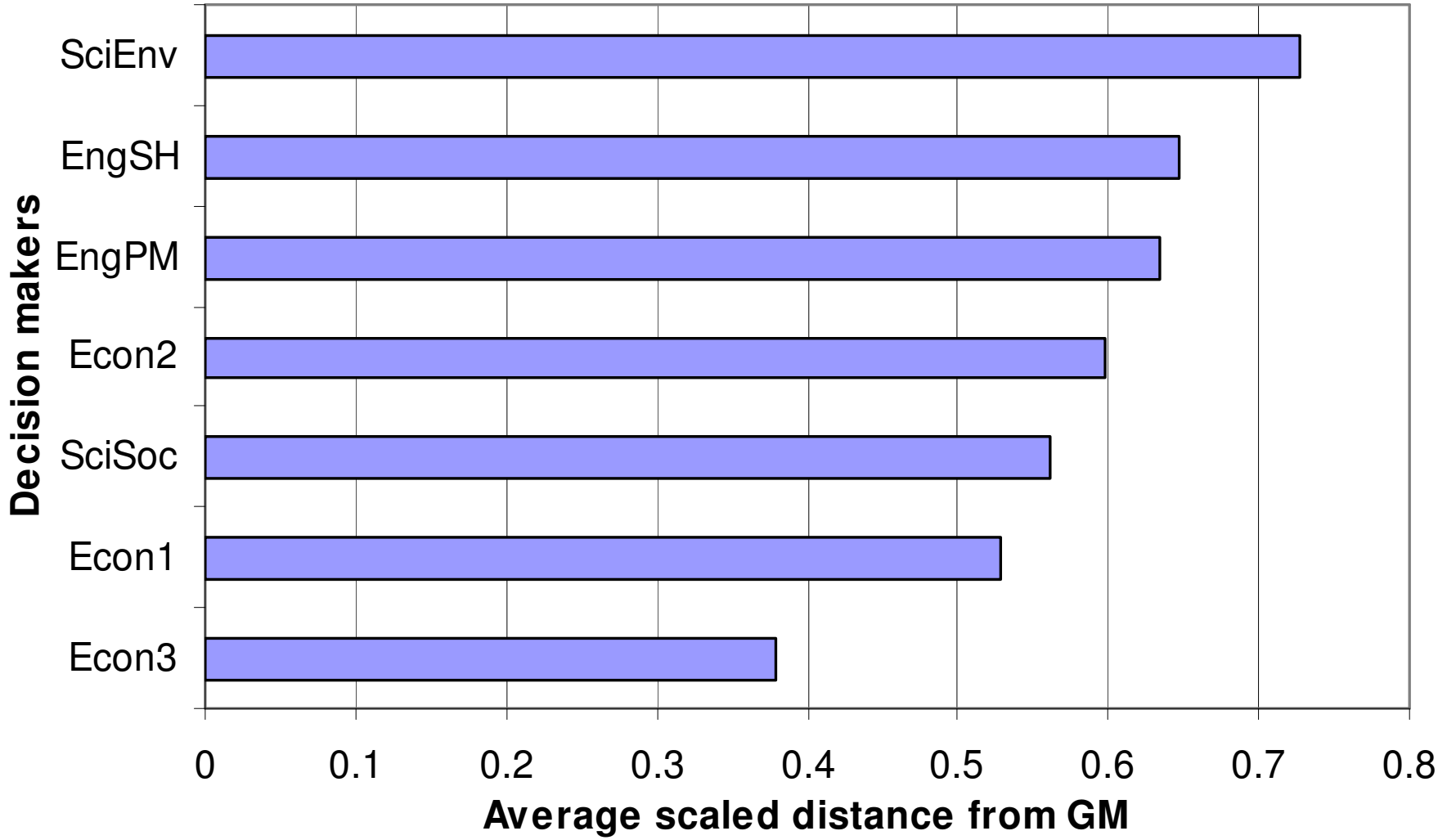

Select the Best Policy
Fit with Strategic Plan
Minimize Risks
ROI
External Drivers
Funding
Alternatives

Government Agency Data from Expert Choice

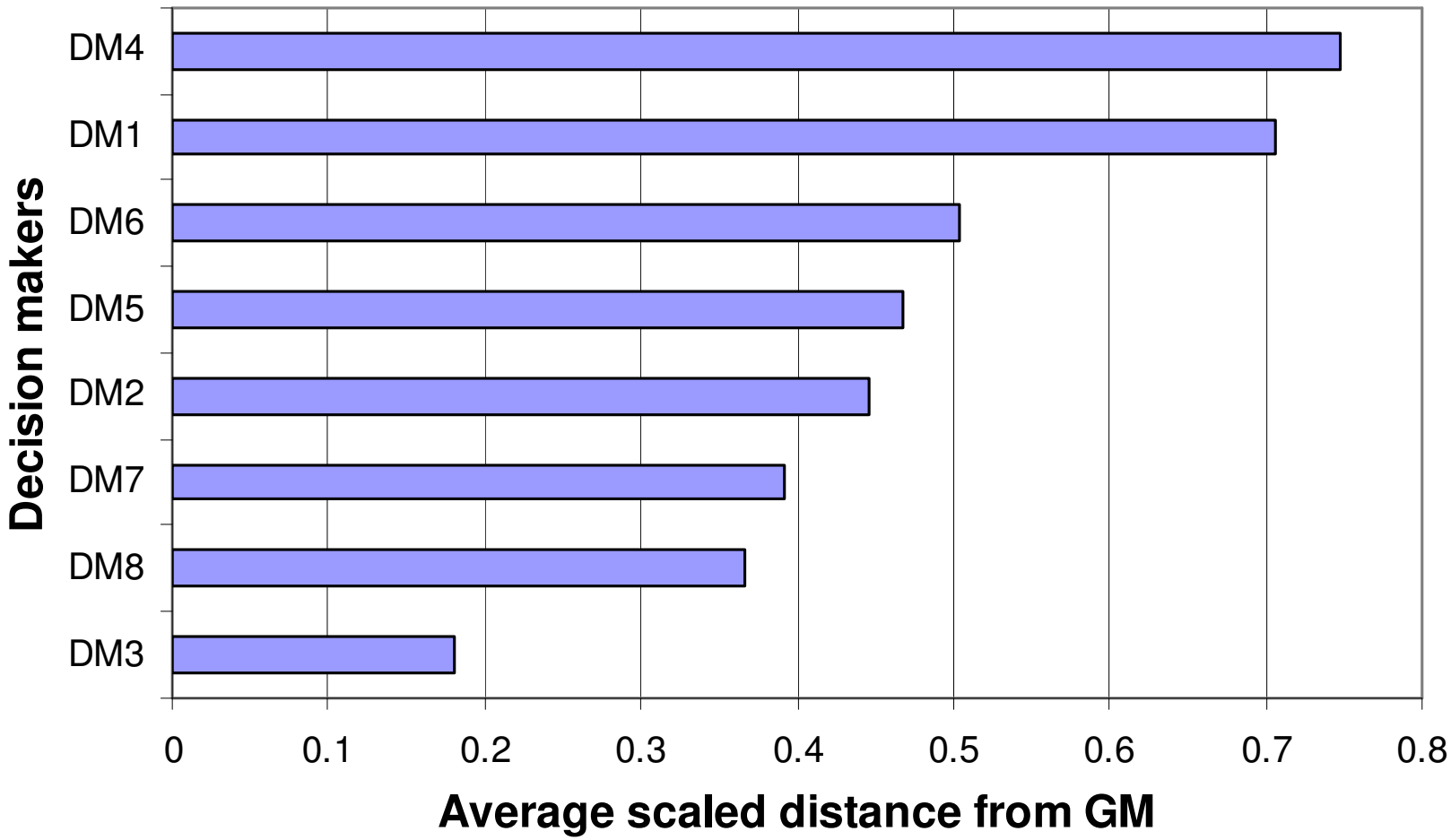
Visualizing the Judgments of the Decision Makers

- In the tables presented, an eigenvector (row) is associated with an individual decision maker
- Each eigenvector (row) can be thought of as a point in n-dimensional space
- We can build a Sammon map to visualize the judgments of the decision makers and the geometric mean in 2 dimensions
- It is easy to spot clusters, outliers, and decision makers whose judgments merit questioning

Data Set # 1 (4 maps)



Data Set # 2 (6 maps)



Conclusions: Part Two

- AHP is a powerful and well-established decision-making tool
- High-quality commercial AHP software exists and is widely used (e.g., Expert Choice and Criterium)
- Sammon maps add value by allowing users to see clusters, outliers, and suspicious judgments that are not easily discernible from a table of eigenvectors