Clustering

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

Clustering

The purpose of clustering is to determine the similarity structure of the data. To determine the natural homogeneous groups in the data. Each natural group is called a cluster. The observations are densely distributed in the cluster and the observations in the spaces between clusters are sparsely distributed.

K-Means

- Let $X = \langle x_1, \ldots, x_Z \mid x_z \in R^N \rangle$ be the data set
- Each x_z is an N-tuple
- Determine a K-block partition $\pi = \{\pi_1, \ldots, \pi_K\}$ of X
- Define $\mu_k = \frac{1}{|\pi_k|} \sum_{x \in \pi_k} x$
- Such that

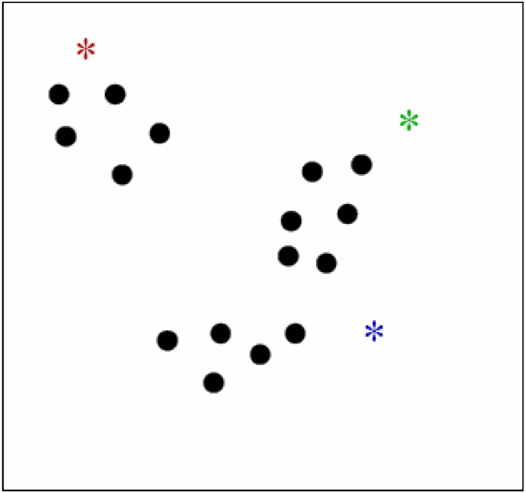
$$\sum_{K=1}^{K} \sum_{x \in \pi_k} ||x - \mu_k||^2$$

is minimized

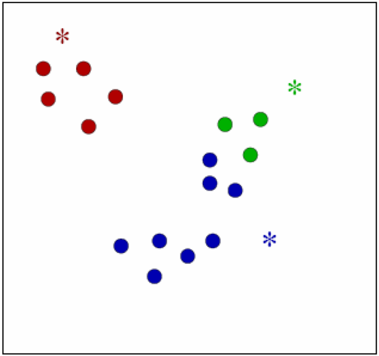
K-Means

- Choose initial K centers $\mu_1, \ldots, \mu_k$ at random
- Iterate until no change
  - For each observation, find the center to which it is closest
  - This association forms a K-block partition $\pi = \{\pi_1, \ldots, \pi_K\}$
  - Where block π_k contains all the observations closest to center μ_k
  - The new center μ_k is the mean of all the observations in π_k

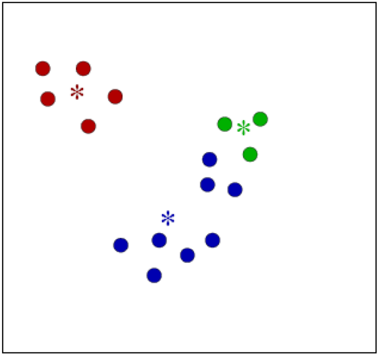
K-Means Initial



K-Means Iterate



Assign to nearest representative



Re-estimate means

K-Means Problems

K-Means result is sensitive to the initial placement of cluster centers.

K-Means has problems when

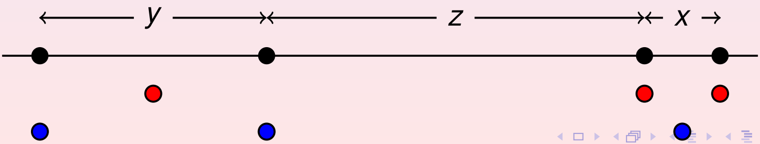
- There are outliers
- Clusters have vastly different sizes
- Cluster shapes are not spherical
- Clusters have different covariance matrices
- Clusters can become empty
- Clusters can merge
- Achieves only a local minimum

K-means is often run multiple times with different random number seeds and the best result is taken.

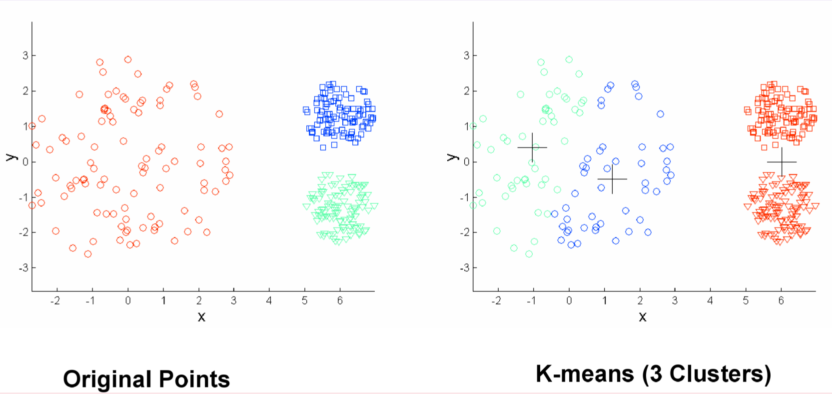
Local Minimum

Solving the global K-Means is NP-Hard.
Shown below are two fixed points of the K-means algorithm

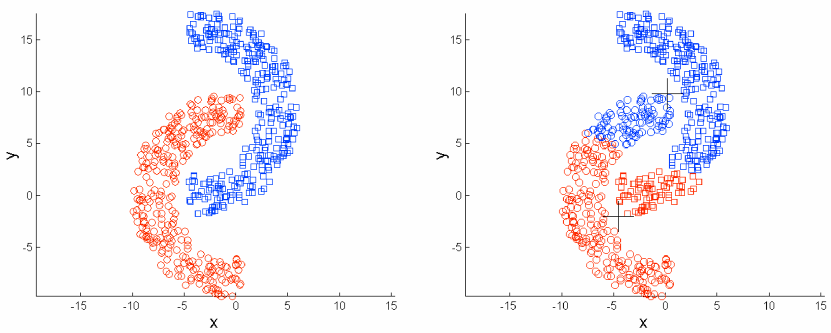
- 4 Data points (black)
  - $x < y < z$
- Non-optimal K-Means Clustering
- Optimal K-Means Clustering



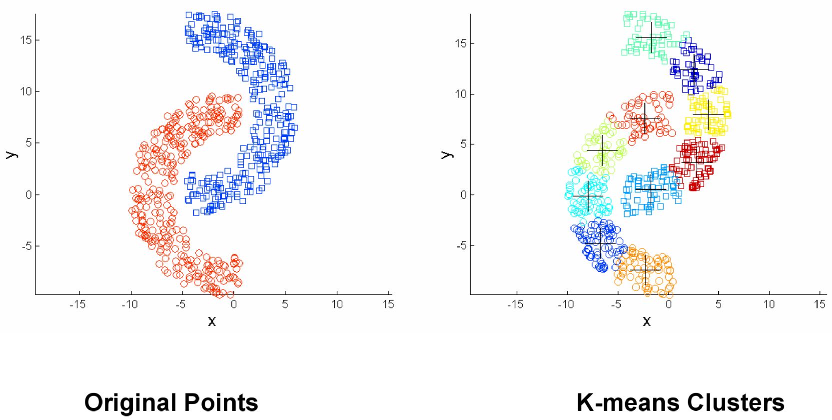
K-Means Differing Covariance Matrices



K-Means Non-Spherical Shapes



K-Means More Clusters



K-Means Only Local Minimum

- Repeat K-Means many times with different randomly chosen initial centers
- Keep the best result clusters

Near Global K-Means

- Incremental-deterministic algorithm
- Employs the K-Means Algorithm as a local search procedure
- Obtains near optimal solutions

K-Means Clustering

Problem Statement

- Given dataset $\langle x_1, \ldots, x_Z \mid x_z \in R^N \rangle$
- Partition the data set into K disjoint clusters
  - Clusters $\pi_1, \ldots, \pi_K$
  - Means $\mu_1, \ldots, \mu_K$
- To Minimize

$$E(\mu_1, \ldots, \mu_K) = \sum_{k=1}^{K} \sum_{x \in \pi_k} ||x - \mu_k||^2$$

Near Global K-Means

- Solve the 1-Means Cluster problem
  - Find μ_1 that minimizes $\sum_{z=1}^{Z} ||x_z - \mu_1||^2$
  - $\mu_1 = \frac{1}{Z} \sum_{z=1}^{Z} x_z$
- Solve the 2-Means Cluster problem
- The center for first cluster of 2-means is μ_1, the solution to 1-Means,
  - For each $z \in \{1, \ldots, Z\}$ set the second cluster center to x_z
  - Define $\pi_1 = \{x \in X \mid ||x - \mu_1|| \leq ||x - x_z||$
  - Define $\pi_2 = \{x \in X \mid ||x - x_z|| < ||x - \mu_1||$
  - Find that z such that $\mu_2 = x_z$ minimizes

$$\sum_{k=1}^{2} \sum_{x \in \pi_k} ||x - \mu_k||^2$$

- Set $\mu_k = \frac{1}{|\pi_k|} \sum_{x \in \pi_k} x, \;\; k \in \{1, 2\}$

Global K-Means: m^{th} Iteration

- Let $\mu_1, \ldots, \mu_{m-1}$ be the means associated with the solution to the $m-1$ clustering problem
- For each $y \in X$
  - Set $\mu_m = y$
  - Use $(\mu_1, \ldots, \mu_{m-1}, \mu_m)$ as the cluster centers for the m^{th} run
  - For each $n \in \{1, \ldots, m\}$ determine
    - $\pi_n = \{x \in X \mid ||x - \mu_n|| < ||x - \mu_i||, i \neq n\}$
    - Evaluate $E = \sum_{k=1}^{m} \sum_{x \in \pi_k} ||x - \mu_k||^2$
- μ_m is the resulting center with smallest error over the N runs
- Set $\mu_k = \frac{1}{|\pi_k|} \sum_{x \in \pi_k} x, \; k \in \{1, \ldots, m\}$

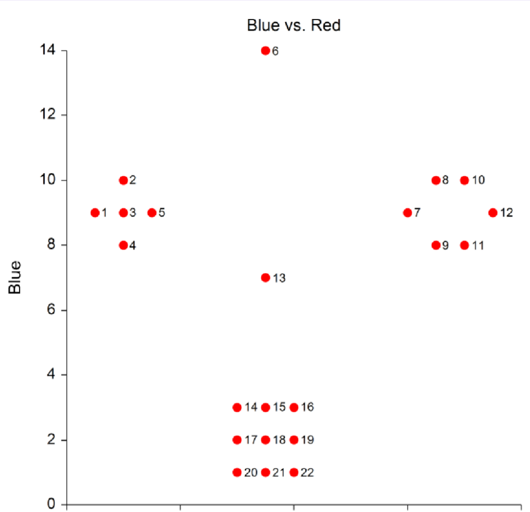
Near Global K-Means

- Does not suffer from the Initialization problem
- Computes clustering in a deterministic way
- Provides all intermediate solutions with $1, \ldots, M$ clusters when solving the M-clustering problem
- Experiments show Global K-Means is better than K-Means with multiple random starts

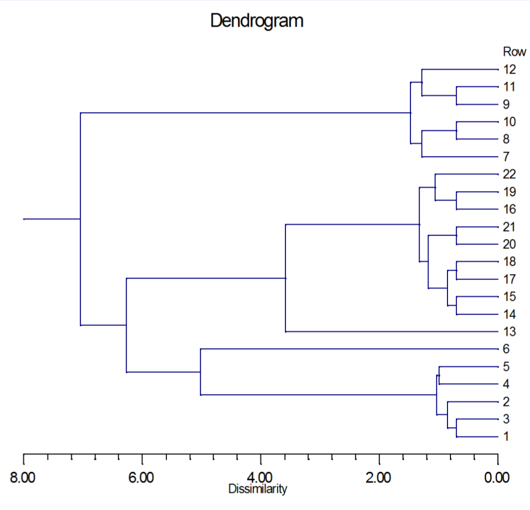
Agglomerative Hierarchical Clustering

- Initialization: Each observation is in its own cluster
- At each step, the two clusters that are most similar are joined into a new cluster
- The clustering is shown as a dendrogram

Example Data



Dendrogram


Dendrogram
Row
12
11
9
10
8
7
22
19
16
21
20
18
17
15
14
13
6
5
4
2
3
1
8.00
6.00
4.00
2.00
0.00
Dissimilarity

Hierarchical Algorithms

- d_{ij}: Distance between clusters i and j
- n_i: Number of observations in cluster i
- D set of all remaining d_{ij}
- Repeat until D contains a single
- Find the smallest element d_{ij} in D
- Merge clusters i and j into a single new cluster k
- Calculate a new set of distances d_{km} by:
  - $d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma |d_{im} - d_{jm}|$
- The new distances replace d_{im} and d_{jm} in D
- $D \leftarrow D - \{d_{im}, d_{jm} | m = 1, \ldots, M\} \cup \{d_{km} | m = 1, \ldots, M\}$
- $n_k = n_i + n_j$

Distance Between Clusters

- Single Linkage: $d_{ij} = \min_{x \in C_i} \min_{y \in C_j} \rho(x, y)$
- Complete Linkage: $d_{ij} = \max_{x \in C_i} \max_{y \in C_j} \rho(x, y)$
- Average Linkage: $d_{ij} = \frac{1}{n_i n_j} \sum_{x \in C_i} \sum_{y \in C_j} \rho(x, y)$
- Centroid Linkage: $d_{ij} = \rho(\mu_i, \mu_j)$
- Median Linkage: $d_{ij} = \frac{n_i n_j}{(n_i + n_j)^2} \rho(\mu_i, \mu_j)$
- Group Linkage: $d_{ij} = \frac{n_i}{n_i + n_j} \sum_{x \in C_i} \frac{n_j}{n_i + n_j} \sum_{y \in C_j} \rho(x, y)$

Variations

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma |d_{im} - d_{jm}|$$

- Single Linkage: $\alpha_i = \alpha_j = .5, \beta = 0, \gamma = -.5$
- Complete Linkage: $\alpha_i = \alpha_j = .5, \beta = 0, \gamma = .5$
- Average Linkage: $\alpha_i = \alpha_j = .5, \beta = 0, \gamma = 0$
- Centroid Linkage: $\alpha_i = n_i/n_k, \alpha_j = n_j/n_k, \beta = -\alpha_i\alpha_j, \gamma = 0$
- Median Linkage: $\alpha_i = \alpha_j = .5, \beta = -.25, \gamma = 0$
- Group Linkage: $\alpha_i = n_i/n_k, \alpha_j = n_j/n_k, \beta = 0, \gamma = 0$

K-Center Clustering

- Given observation data $x_1, \ldots, x_N$
- Partition in K clusters $C_1, \ldots, C_K$
- Cluster spread of C_k
  - The least value of D_k for which all points are
  - Within distance D_k of each other
  - Or within distance $D_k/2$ of the cluster center
- The cluster size D of the partition is $D = max_{k=1,\ldots,K} D_k$
- Find the partition that minimizes D

K-Means Versus K-Center

K-Means minimizes

$$\sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu_k||^2$$

where μ_k is the centroid for cluster C_k

K-Center minimizes

$$\max_{k=1,\ldots,K} \max_{x \in C_k} ||x - c_k||^2$$

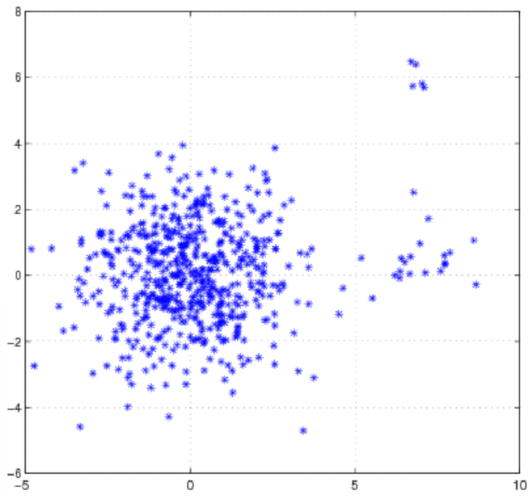
where c_k is the center of cluster C_k

Alternate Formulation

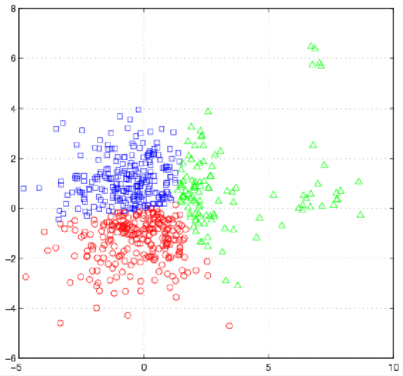
Find the partition $C_1, \ldots, C_K$ that minimizes

$$\max_{k=1,\ldots,K} \max_{x,y \in C_k} \rho(x, y)$$

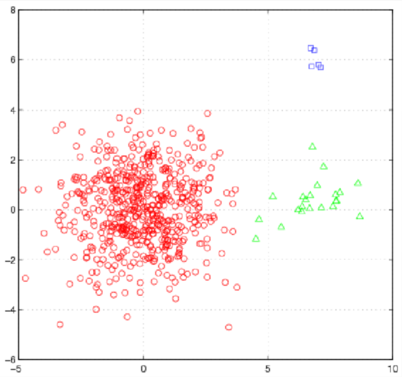
Example Data



K-Means and K-Center



Clustering by k-means. K-means focuses on average distance.



Clustering by k-center. K-center focuses on worst scenario.

Greedy Algorithm

- Choose a subset H consisting of K points that are farthest apart from each other
- Point $c_k \in H$ represents a cluster center for cluster C_k
- $C_k = \{x \mid \rho(x, c_k) \leq \rho(x, c_j), j = 1, \ldots, K\}$

Greedy Algorithm

Let D^* minimize

$$D^* = \max_{k=1,\ldots,K} \max_{x,y \in C_k} \rho(x, y)$$

Let D be the cluster spread produced by the greedy algorithm. Then $D^* \leq D \leq 2D^*$.

Faculty Evaluation: Journals and Research

Column		J-score	Weight
D	1	Number of Journal papers	1
E	2	Number of Conference papers	.75
F	3	Number of Books	2
G	4	Number of Books edited	.5
H	5	Number of Book chapters	1
I	6	Number of Patents	1
J	7	Total dollars of external research grants	.000005
K	8	Total dollars of external education grants	.000005
L	9	Total dollars of external equipment grants	.0000005
	10	Number of recognition awards	0

Faculty Evaluation: PhD Student Interaction

Column		P-score	Weight
M	1	Number of completed doctoral students	1
N	2	Number of current doctoral student mentoring	.5
O	3	Number of doctoral exam committees	.1
P	4	Number of doctoral courses taught	.5

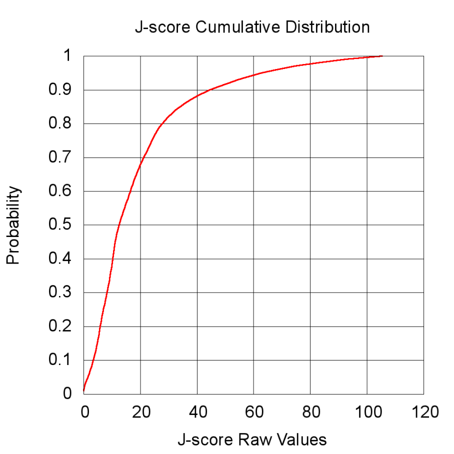
Faculty Evaluation: Professional Service

Column		S-score	Weight
Q	1	Journal Editorial boards	.1
R	2	Major conference organization	.5
S	3	Program committees	.25
T	4	Number of conferences or journals reviewer for	.1

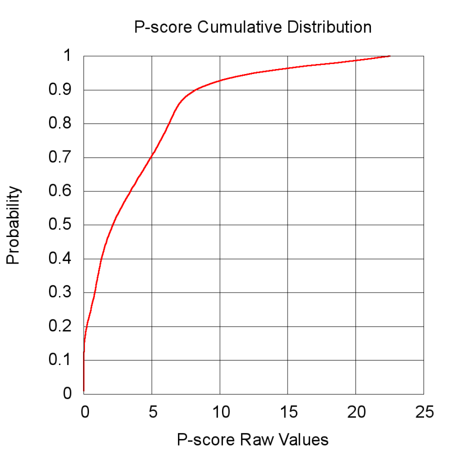
Faculty Evaluation: Career Standing

Column		G-score	Weight
U	1	Google log (number of citations+50)	2
V	2	Google H-index	1
	3	Google I10-index	0
W	4	Google total number of documents cited	.05

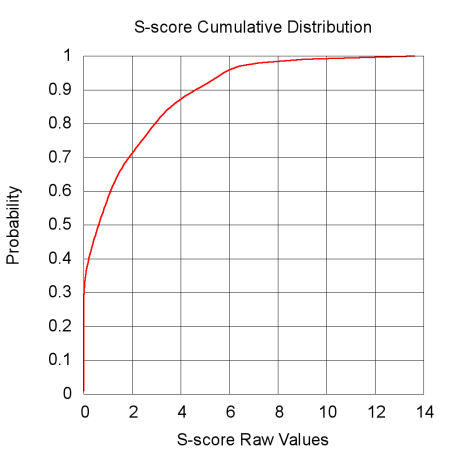
J-Score


J-score Cumulative Distribution
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
J-score Raw Values

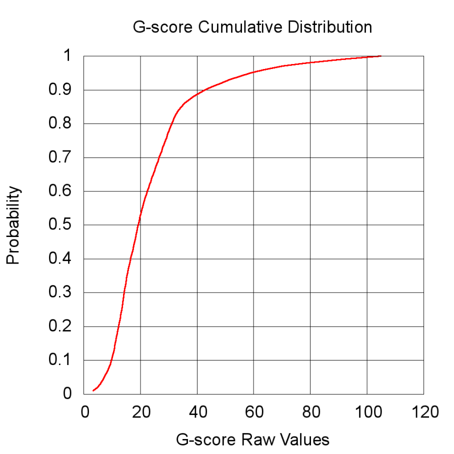
P-Score


P-score Cumulative Distribution
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25
P-score Raw Values

S-Score


S-score Cumulative Distribution
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
S-score Raw Values

G-Score


G-score Cumulative Distribution
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
120
G-score Raw Values

Correlations

	J-score	P-score	S-score	G-score
J-score	1.0000	0.3137	0.3960	0.5306
P-score	0.3137	1.0000	0.2218	0.5406
S-score	0.3960	0.2218	1.0000	0.2088
G-score	0.5306	0.5406	0.2088	1.0000

Data Normalization: z-scores

For each field independently,

- Let μ be the mean of the field's value over all records
- Let σ be the standard deviation of the field's value over all records
- Let x be a raw value of the field

$$x_{normalized} = \frac{x - \mu}{\sigma}$$

Range Normalization

For each field independently,

- Let x_{min} be the minimum value in the field over all records
- Let x_{max} be the maximum value in the field over all records
- Let x be a raw value of the field

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Normalizes the values to between 0 and 1

Rank Normalization

For each field independently,

- Let $x_1, \ldots, x_N$ be the values of the field in record 1 through record N
- Sort these values from smallest to largest $x_{(1)}, \ldots, x_{(N)}$
- $x_{n\ normalized} = k$ where $x_n = x_{(k)}$

Rank Normalization Example

Original Data

x_1	79.2
x_2	1.58
x_3	191.6
x_4	4.63

Sorted Data

$x_{(1)}$	1.58
$x_{(2)}$	4.63
$x_{(3)}$	79.2
$x_{(4)}$	191.6

Rank Normalized Data

x_1 *normalized*	3
x_2 *normalized*	1
x_3 *normalized*	4
x_4 *normalized*	2

Correlation For Rank Normalized Data

	J-score	P-score	S-score	G-score
J-score	1.0000	0.4630	0.4929	0.6467
P-score	0.4630	1.0000	0.2807	0.4436
S-score	0.4929	0.2807	1.0000	0.3553
G-score	0.6467	0.4436	0.3553	1.0000

Initial Centers

Profile	J-score	P-score	S-score	G-score
Less good in Research	23.50	74.50	74.50	74.50
Less good in PhD student interaction	74.50	23.50	74.50	74.50
Less good in Professional service	74.50	74.50	17.50	74.50
Less good in Career standing	74.50	74.50	74.50	23.50
Good in all four areas	74.50	74.50	74.50	74.50
Not good in any of the four areas	23.50	23.50	17.50	23.50

Final K-means Centers

Profile	J-score	P-score	S-score	G-score
Less good in Research	44.2500	65.9375	74.2500	71.8750
Less good in PhD student interaction	68.9500	26.4000	69.9000	75.5000
Less good in professional service	59.3333	56.4722	19.7500	53.3889
Less good in career standing	64.0909	52.5909	78.5455	29.0455
Good in all four areas	80.7647	84.7941	72.5588	81.9706
Not good in any of the four areas	18.0147	28.6471	31.0588	23.4706

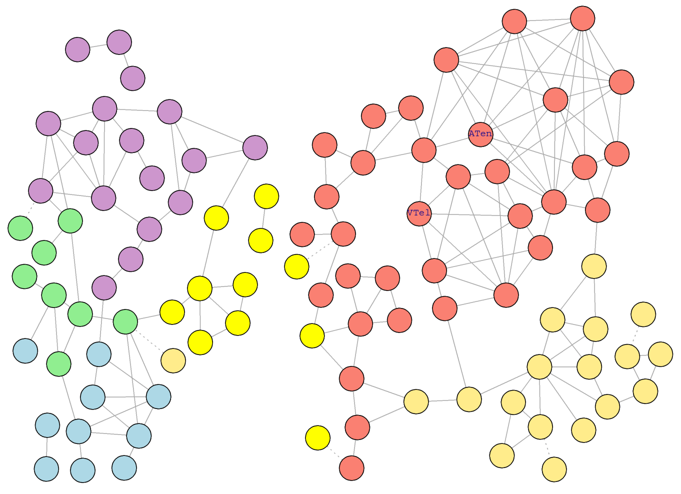
K-means Inter-Cluster Distances

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Cluster 1	0.000	46.961	60.242	49.240	42.352	79.293
Cluster 2	46.961	0.000	63.251	54.243	59.987	82.554
Cluster 3	60.242	63.251	0.000	63.931	69.765	59.199
Cluster 4	49.240	54.243	63.931	0.000	64.436	70.586
Cluster 5	42.352	59.987	69.765	64.436	0.000	110.610
Cluster 6	79.293	82.554	59.199	70.586	110.610	0.000

Graph Clustering

- Each faculty member has a normalized rank score in the four evaluation dimensions.
- This can be thought of as a point in a four dimensional space.
- Between every pair of points we define the Manhattan distance as the sum of the absolute values of the differences.
- We make a graph where each node is associated with a doctoral faculty member and a pair of nodes are joined with an edge if their Manhattan distance is less than 42.
- Any isolated node is joined to its nearest node with a dotted line.
- This yields about 165 edges plus 6 dotted edges.

Graph Clustering



Graph Clustering

There are six k-means clusters with the colors of the nodes indicating cluster type.

- Green – (cluster 1) productive in the PhD student interaction, professional service, and career standing areas;
- Light blue – (cluster 2) productive in the research, professional service, and career standing areas;
- Gold – (cluster 3) productive in the research, PhD student interaction, and career standing areas;
- Yellow – (cluster 4) productive in the research,PhD student interaction, and professional service areas;
- Purple – (cluster 5) productive in all four evaluation dimensions;
- Salmon – (cluster 6) unproductive in all four evaluation areas

Student Progress Data

Students in the Computer Science Doctoral Program who have completed their PhD degree have five dates that mark their progress.

- Date Entered Program
- Date Passes First Exam
- Date Completed Survey Exam
- Date Completed Dissertation Proposal Exam
- Date Defended Dissertation

Coding Data: Relative Time from Date of Entry

- Number of Months to Pass First Exam
- Number of Months to Complete Survey
- Number of Months to Complete Proposal
- Number of Months to Defend Dissertation

Coding Data: Intervals Between Successive Milestones

- Number of Months to Pass First Exam
- Number of Months to Complete Survey After Passing First Exam
- Number of Months to Complete Proposal After Completing Survey Exam
- Number of Months to Defend Dissertation After Completing Proposal Exam

Means and Medians

Given scalar data $x_1, \ldots, x_Z$ the number c that minimizes

$$\sum_{z=1}^{Z} (x_z - c)^2$$

is the sample mean

$$\mu = \frac{1}{Z} \sum_{z=1}^{Z} x_z$$

Means and Medians

Given scalar data $x_1, \ldots, x_Z$ and its sorted form $x_{(1)}, \ldots, x_{(Z)}$ the number c that minimizes

$$\sum_{z=1}^{Z} |x_z - c|$$

is the sample median

$$c_{median} = x_{(Z/2)}$$

Manhattan Distance

The Manhattan Distance ρ between two vectors $u = (u_1, \dots, u_N)$ and $v = (v_1, \dots, v_N)$ is defined by

$$\rho(u, v) = \sum_{n=1}^{N} |u_n - v_n|$$

K-Medians: Two Clusters

- Select in turn all pairs of observations as cluster centers
- Determine the Manhattan Distance between each observation and cluster center
- Associate each observation with its closest cluster center
- For each cluster center, there is the sum of all the Manhattan distances from its observations to its cluster center
- Define the objective function as the sum over two clusters of their total Manhattan distance
- Choose that pair of observations which when made as cluster centers produces the smallest total Manhattan distance

Cluster Centers

Number of Months From Date of Entry

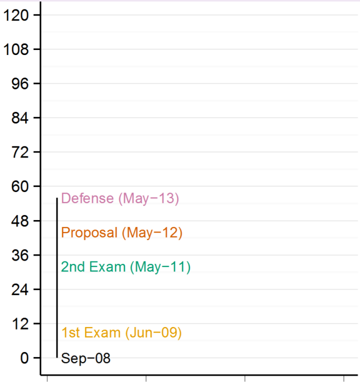
#	First Exam	Survey	Proposal	Defense
50	9	32	44	56
24	9	63	80	96

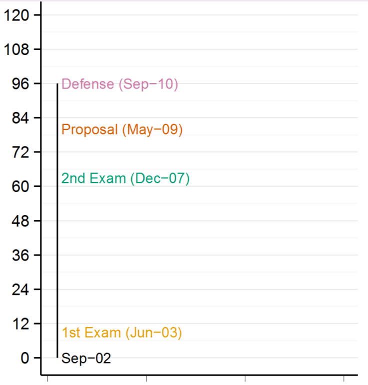
Data In Interval Form

#	First Exam	Survey	Proposal	Defense
50	9	23	12	12
24	9	54	17	16

Cluster Centers

Shows in graphic form the cluster centers of the two clusters. The left graphic is cluster 1 center. The right graphic is cluster 2 center





Graph

Shows the graph connecting each pair of students whose distance is less than 16. The center for cluster 1 is 17. The center for cluster 2 is 46. Nodes which are disconnected from all other nodes by the threshold 16 are connected with their closest neighbor by a dotted edge.