# Data Clustering and Mining

Dexin Zhou – Bard College (Presenter)
Ralph Abbey – North Carolina State U.
Jeremy Diepenbrock – Washington U. at St. Louis

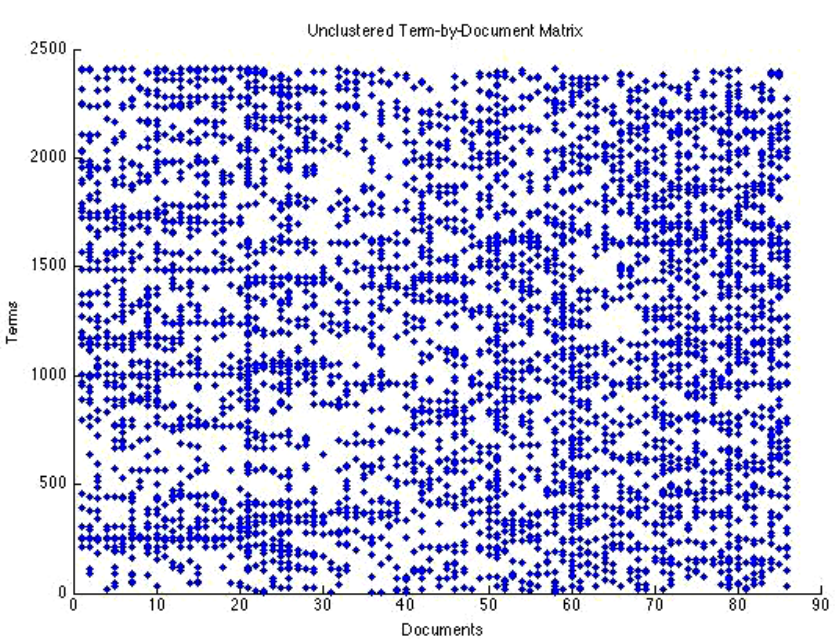
Project Advisor: Dr. Carl Meyer
Additional Advising: Dr. Amy Langville
Graduate Assistant: Shaina Race

# What is Data Clustering?

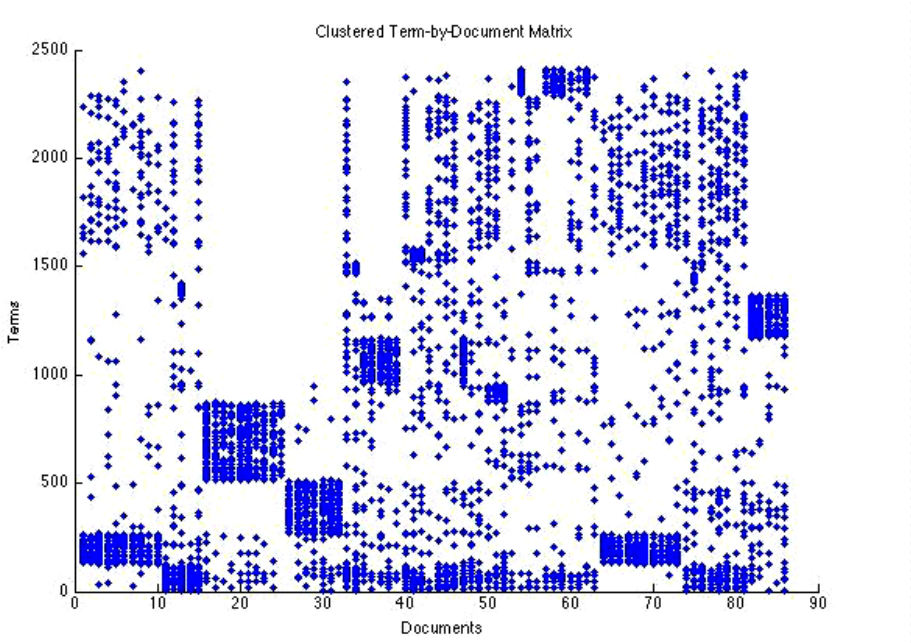
- Clustering is the partitioning of a data set into subsets (clusters).
- We are interested in creating good clusters that allow us to reorganize disordered data into a block structure so that useful information can be extracted.

# A Visible Example

## Before Clustering



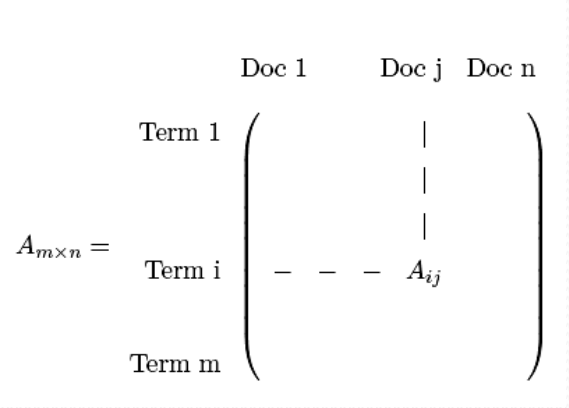

## After Clustering



# What are we clustering?

- An 86 mini-document set that we created with 13 topics
- A 185 document set used in Daniel Boley's paper with 10 topics
- SAS grocery store dataset

# Preparing the data




- Term Aij is in the following form $A_{ij} = l_{ij} g_i d_j$
- g term is a function of term i, it downplays the terms that appear frequently globally
- l term is a function of the raw frequency of a certain term in document j(eg: log)
- d term is a normalization factor

# How?

- Principal Direction Divisive Partitioning
- Principal Direction Gap Partitioning
- Non-Negative Matrix Factorization
- Clustering Aggregation

# Singular Value Decomposition

$$A = \begin{pmatrix} u_1 & u_2 & \dots & u_m \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots \\ \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \sigma_r \end{pmatrix}_{m \times n} \begin{pmatrix} v_1^T \\ v_2^T \\ \dots \\ v_n^T \end{pmatrix}$$

Where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$

# Principle Direction Divisive Partitioning

- PDDP uses the singular vector associated with the largest singular value
- The left singular vector, $u_1$, can be geometrically understood as the principal direction of the spread of the data
- The right singular vector, $v_1$, gives the value of the projection of the data onto the principal direction
- This information is easily obtained from the SVD

# PDDP

- PDDP is an iterative algorithm
- Each iteration the data must be recentered
- The user inputs the number of clusters, k
- There are $k-1$ iterations

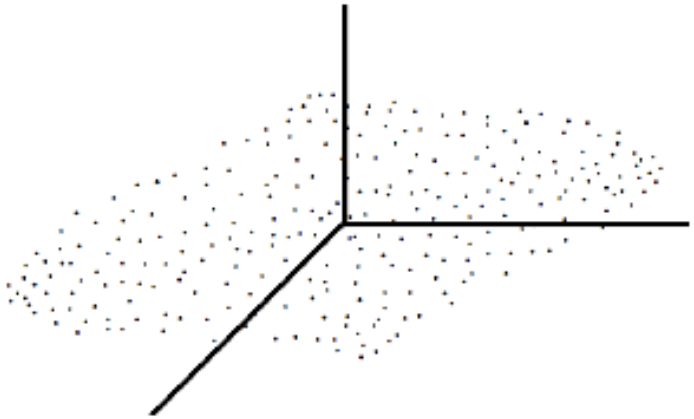
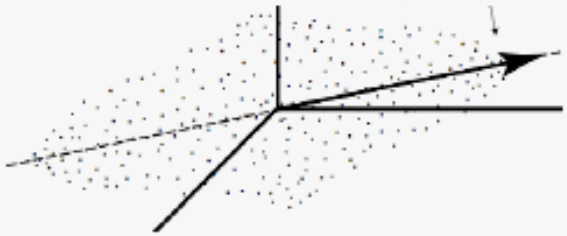


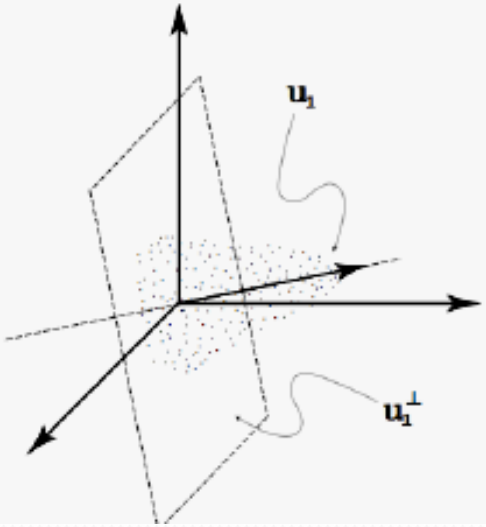
Figure 1: Example of Centered Data

# PDDP



Above: Data Cloud and Principal Direction

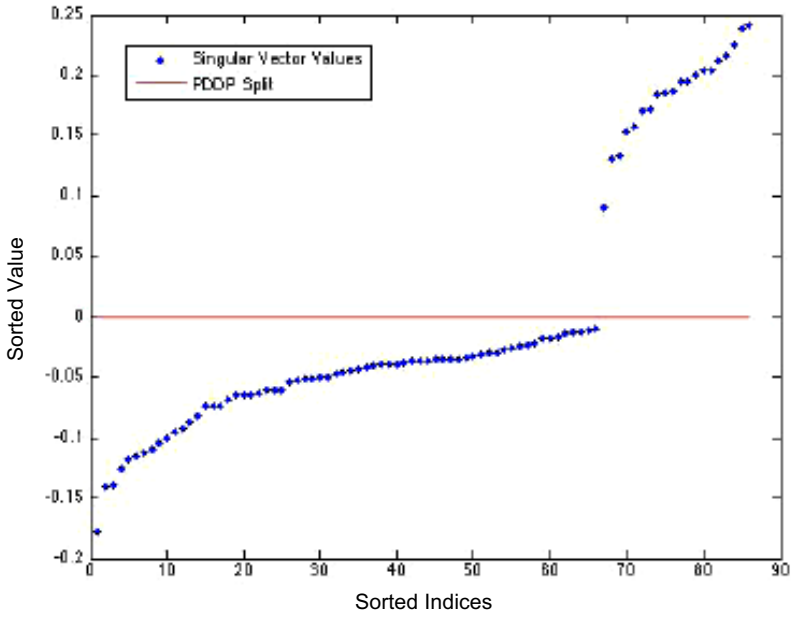
Below: Data Cloud and Principal Direction with Hyper-plane



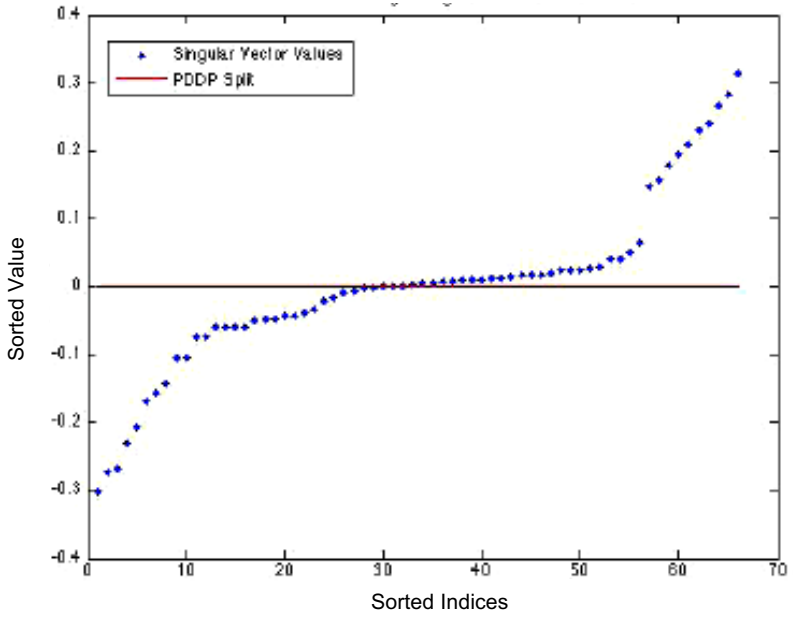
# Principle Direction Gap Partitioning

- A split at zero may not be the best split as done in PDDP

Plot of the First Right Singular Vector




Plot of the Second Right Singular Vector



# A Comparison of PDGP w/ PDDP

## Disadvantages

- The edge values can skew where splitting occurs
- Slower than PDDP

## Advantages

- Better density and/or entropy rankings in some cases
- While edge values can skew where splitting occurs, this can also be changed by the user

# Centering Vs. Non-Centering

## Centering Destroys Sparsity

- Increased time for computation, though using Lanczos there are ways around large computation time
- Increased storage space needed

## Non-centering With SVD

- Now first singular vector cannot be used
- Second singular vector shows trends less reliably
- Maintains sparsity

# Non-Negative Matrix Factorization

$$A \approx \begin{pmatrix} w_{11} & \dots & w_{1k} \\ w_{21} & \dots & w_{2k} \\ \dots & \dots & \dots \\ w_{m1} & \dots & w_{mk} \end{pmatrix}_{m \times k} \begin{pmatrix} h_{11} & \dots & h_{1n} \\ h_{21} & \dots & h_{2n} \\ \dots & \dots & \dots \\ h_{k1} & \dots & h_{kn} \end{pmatrix}_{k \times n}$$

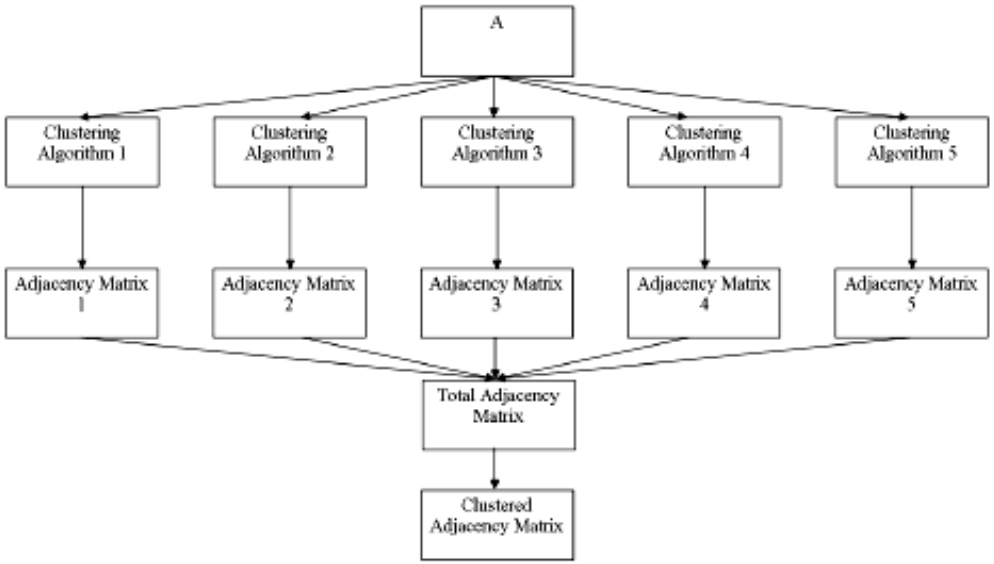
# NMF Clustering

- $A \approx WH$ where A is m by n, W is m by k, and H is k by n
- k refers to the number of topics.
- Thus W is term by topic, and H is topic by document
- NMF clusters both rows and columns at the same time
- Similarily, the user must choose the k

# Cluster Aggregation

1. Run a set of clustering algorithms on the data
2. Form an adjacency matrix for each algorithm by assigning 1 to $A_{ij}$ if document $i$ is in the same cluster as document $j$, and add them together
3. Run clustering on the overall adjacency matrix

---



# Cluster Aggregation

## Motivation

- Different algorithms can yield very different clustering results
- With unknown data sets, knowing which algorithm to use may be difficult
- We would like to use good results from multiple clustering algorithms

## Claims

- Cluster aggregation can yield better clustering than using a single algorithm
- While slow, cluster aggregation can theoretically be made faster by use of parallel computing
- Cluster aggregation can eliminate poor clusters in various algorithms while maintaining good clustering

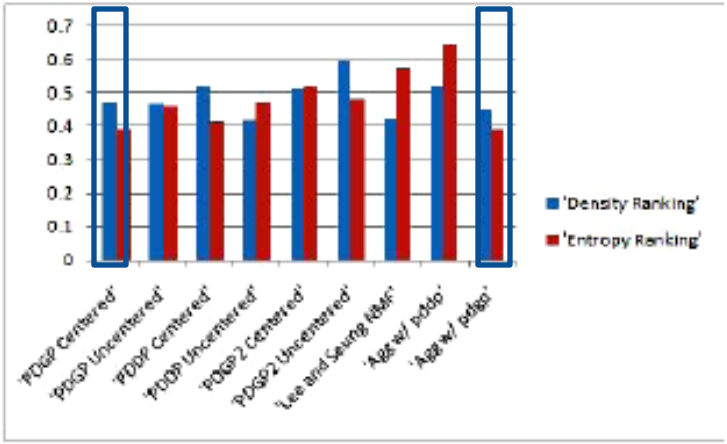
# Metrics

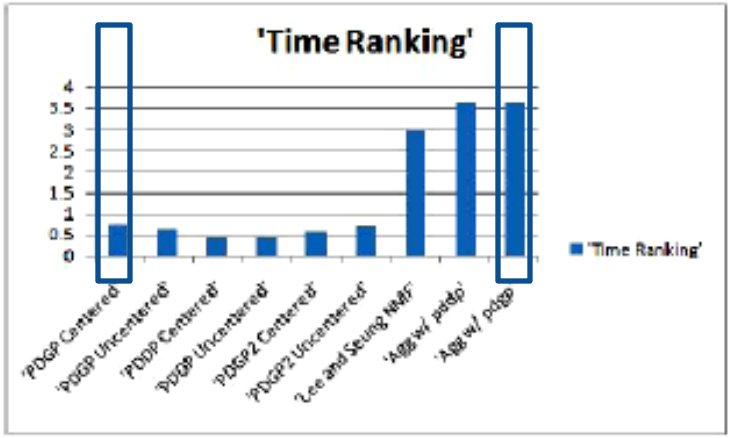
- Entropy Method
  - A standard measurement based on our prior knowledge to the data file.
- Density Method
  - Does not require prior knowledge to the data file.
  - Less accurate.

# Mini-document dataset

- Made from excerpts from webpages with 13 clusters
- Lengths are less than 100 words long
- There is no global weighting, all words are used, with augmented normalized term frequency
- The total term list is 2413 words along the rows with 86 documents along the columns

# Mini-document dataset Result



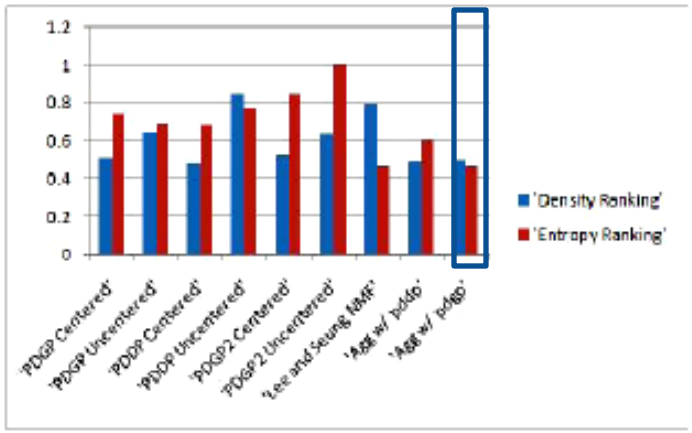



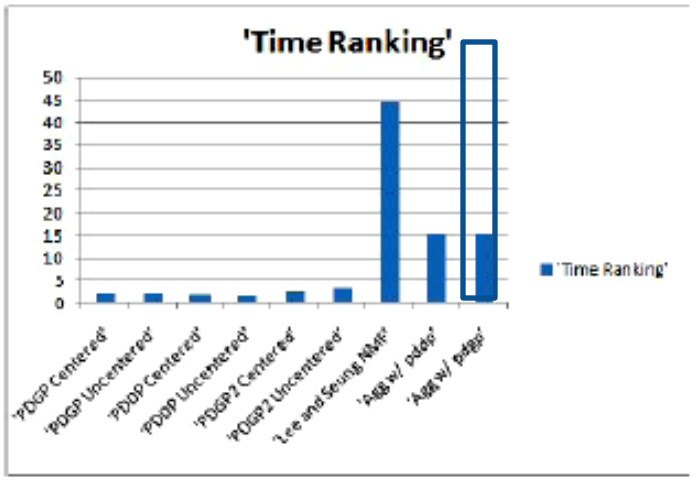

| Methods | Density Ranking | Entropy Ranking | Time Ranking |
|---|---|---|---|
| PDGP Centered | 0.4707 | 0.389 | 0.4846 |
| PDGP Uncentered | 0.4675 | 0.4578 | 0.6865 |
| PDDP Centered | 0.5177 | 0.4146 | 0.4342 |
| PDDP Uncentered | 0.4183 | 0.5054 | 0.4006 |
| Lee and Seung | 0.4229 | 0.5054 | 3.1115 |
| Agg w/ pddp | 0.5194 | 0.5819 | 3.5607 |
| Agg w/ pdgp | 0.4491 | 0.3905 | 3.5649 |

# Boley's J1 Dataset

- Document set used in Daniel Boley's paper "Principal Direction Divisive Partitioning" (1998).
- Documents are from various websites with 10 different clusters
- The columns of the term by document matrix are normalized so that the column sums are one
- The total term list is 10536 words along the rows with 185 documents along the columns

# Boley's Dataset Result



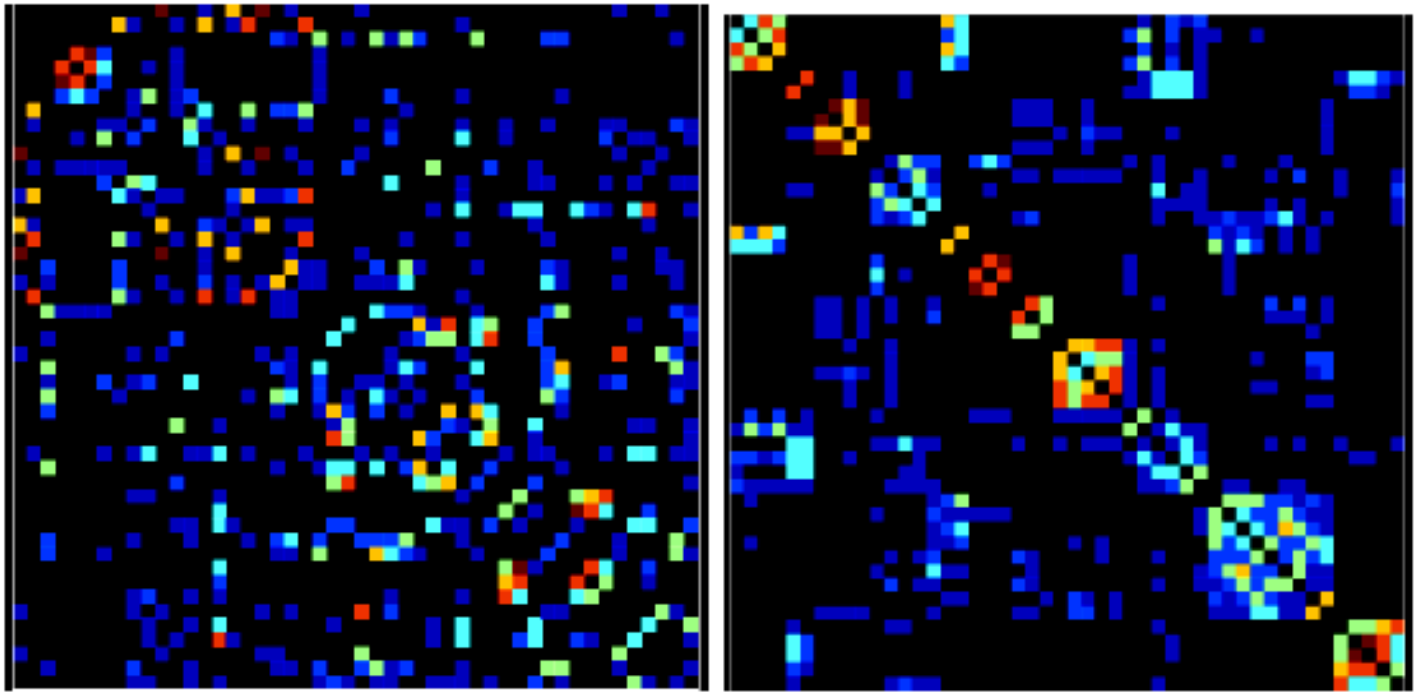




| Methods | Density Ranking | Entropy Ranking | Time Ranking |
|---|---|---|---|
| PDGP Centered | 0.505 | 0.7405 | 2.3876 |
| PDGP Uncentered | 0.6496 | 0.6946 | 2.5744 |
| PDDP Centered | 0.4854 | 0.6886 | 2.0479 |
| PDDP Uncentered | 0.8488 | 0.7742 | 1.762 |
| Lee and Seung | 0.7924 | 0.4661 | 44.1468 |
| Agg w/ pddp | 0.4947 | 0.6012 | 15.5568 |
| Agg w/ pdgp | 0.5014 | 0.4655 | 15.5855 |

# SAS Grocery Dataset

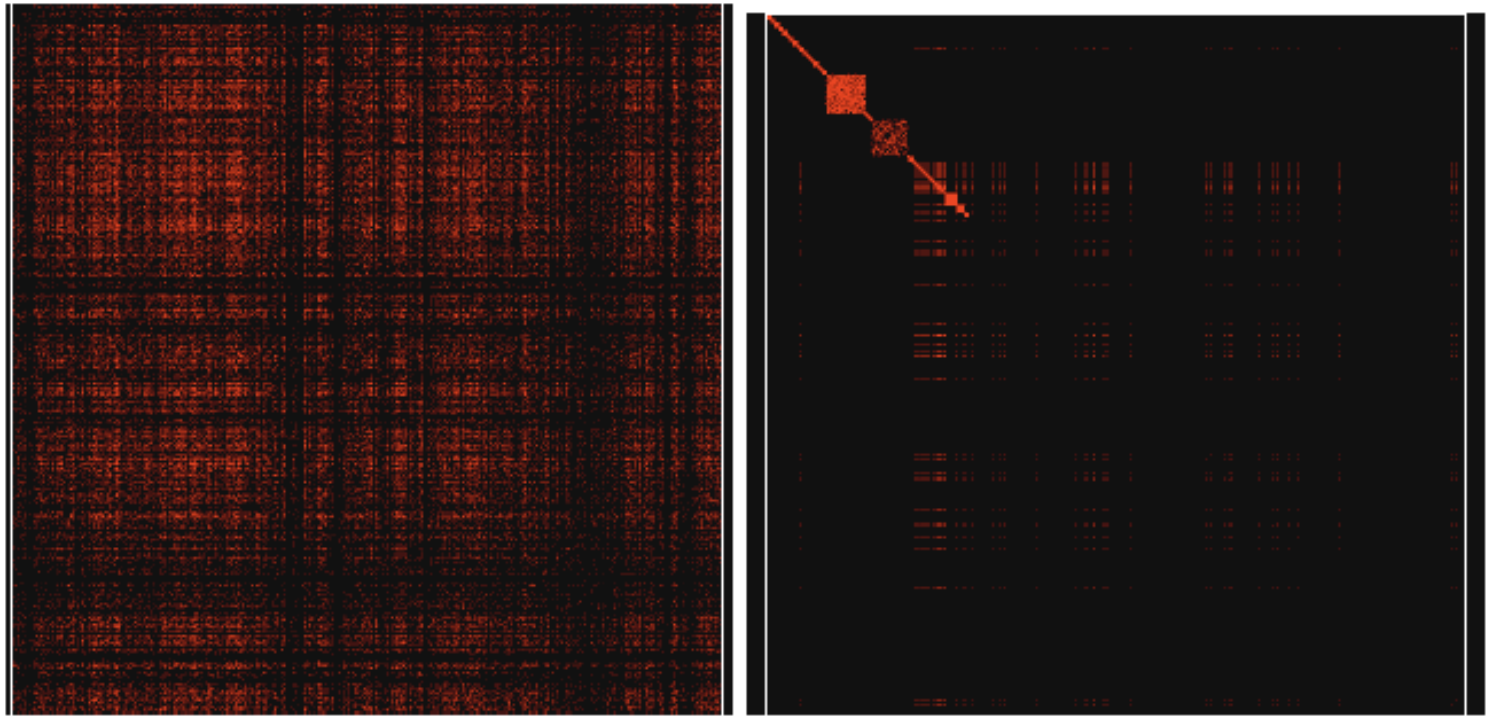
- Provided by SAS for testing purposes
- The total term list is 10982 terms along the rows with 48 stores along the columns
- The actual structure of the data set is not known, unlike in previous tests in which the data was known beforehand

# SAS Grocery Dataset Results



Store by Store aggregation matrices: unclustered and clustered with PDGP

# SAS Grocery Dataset Result



Aggregation matrices: unclustered and clustered with PDGP

# Conclusion

- In tests on several document sets, PDGP has shown promise as a clustering algorithm
- Running clustering with centering does not take significantly more time, however the question of whether to use centered data or uncentered data is still unanswered
- On small, known data sets, cluster aggregation outperforms several other SVD based clustering algorithms

# For Additional Information

- Please Visit
  - http://meyer.math.ncsu.edu/Meyer/REU/REU.html