CAMCOS Verizon Project

Adaptive Spectral Clustering for High-Dimensional Sparse Count Data

with Applications to Document and Web User Clustering

Team leaders: Joey Fitch, Fengmei Liu

Team Members: Shiou-Shiou Deng, Sonia Kong, Nate Kotila, Rachel Li, Ryan Quigley, Andrew Zastovnik

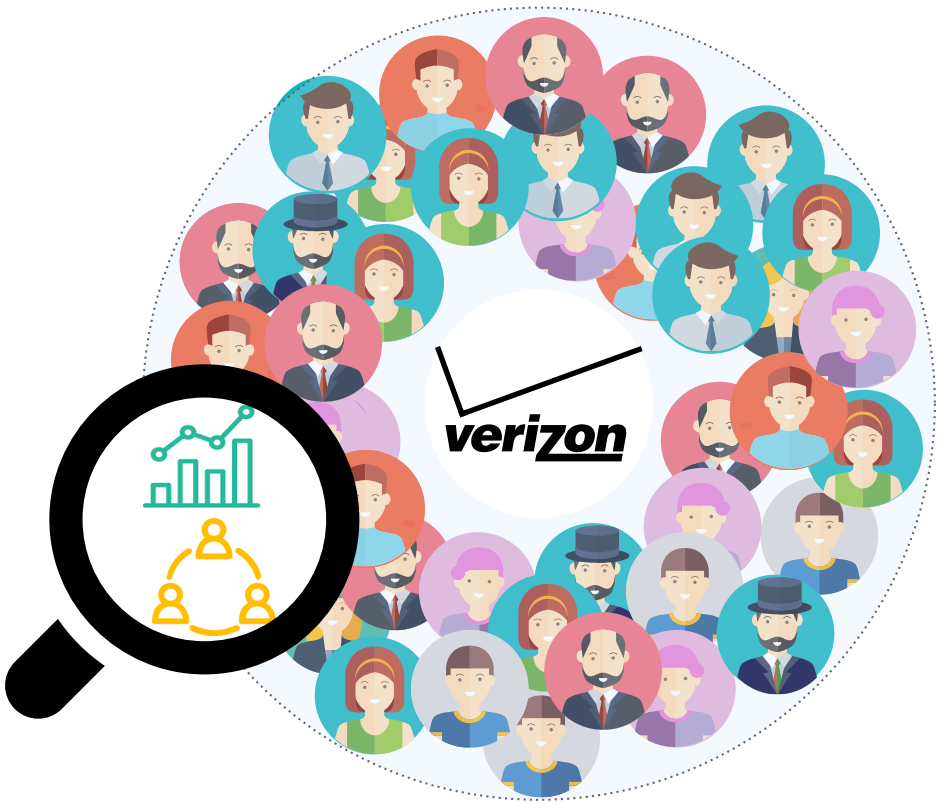
May 19, 2017

Outline

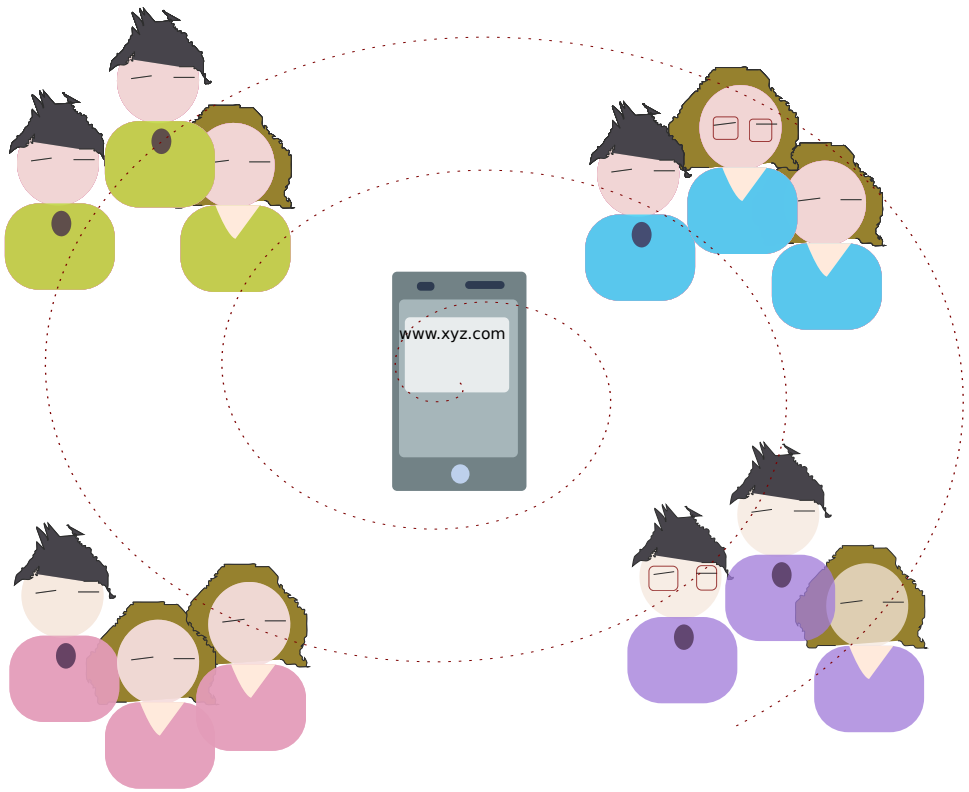
Project Overview
20 Newsgroup Dataset
Data Processing
Dimensionality Reduction
Similarity
Spectral Clustering
Insights

Project Overview - Fengmei

From users' demographic and browsing history data, get insights about customer segmentation, preference prediction, time drift...



Our focus: Discover audience with similar characteristics



Verizon Simulated Data

User	tldAggScore
Id 1	{web1: 1, webm:1}
Id 2	{webm:1}
Id 3	None
...	...
Id n	{web k:1}

999, 937 users

Filter out missing

Verizon Data

(web visits part)

	Web 1	2	...	m
Id 1	1	0	...	1
Id 2	0	0	...	1
...	...	...	...	...
Id n	0	0	...	0

Dense Matrix

User	Web	Frequency
Id 1	Web 1	1
Id 1	Web m	1
Id 2	Web m	1
...	...	...
Id n	Web k	1

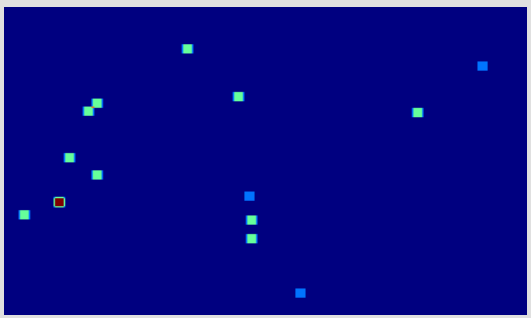
Sparse Matrix

High-Dimension

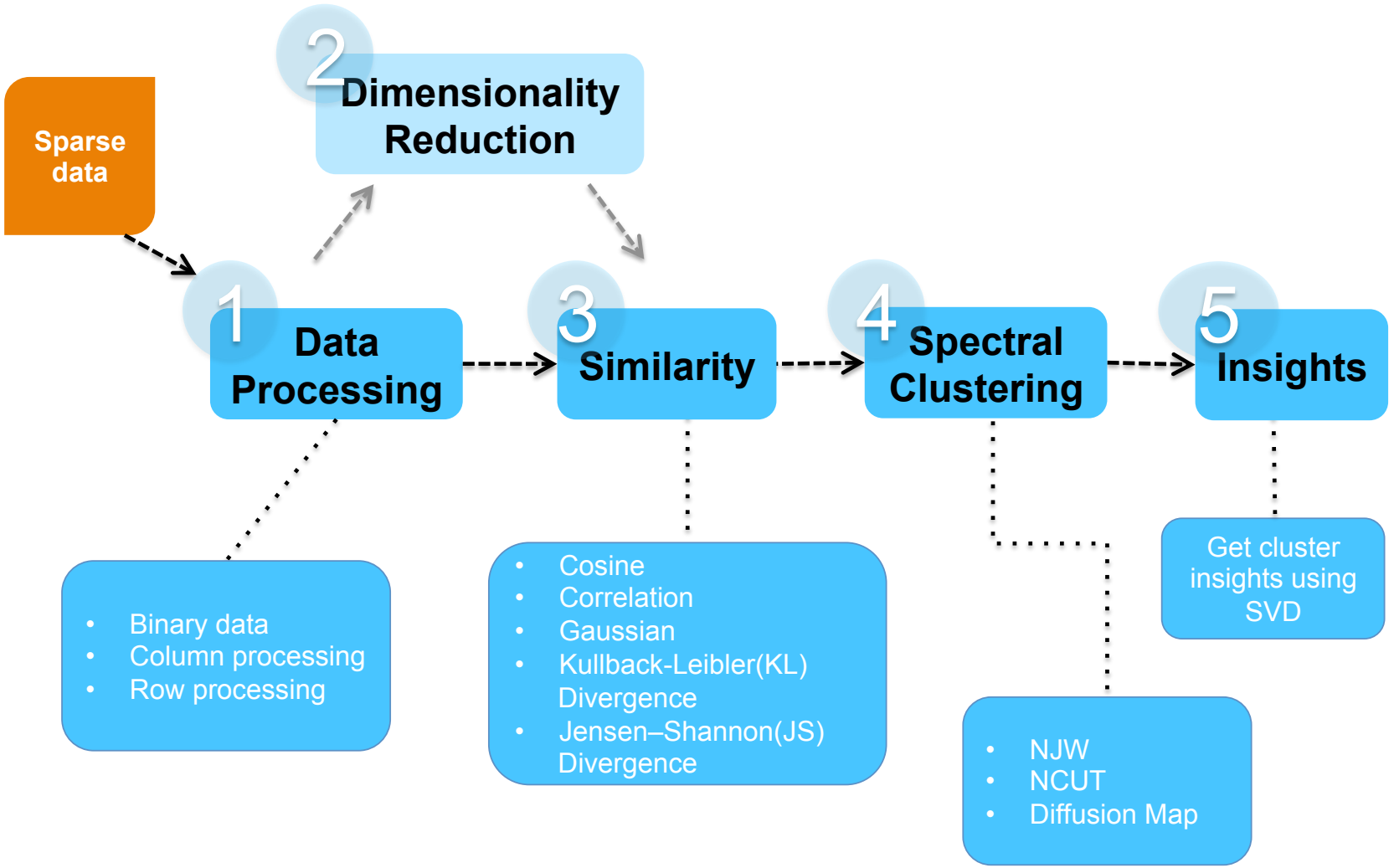
$$330K * 175K$$

High Sparsity

$$\frac{591K}{330K * 175K} = 0.001\%$$



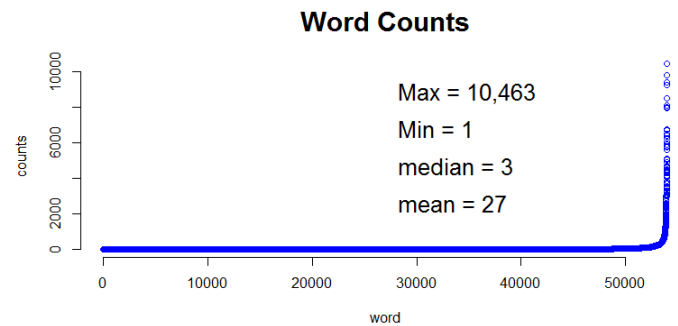
Process Framework

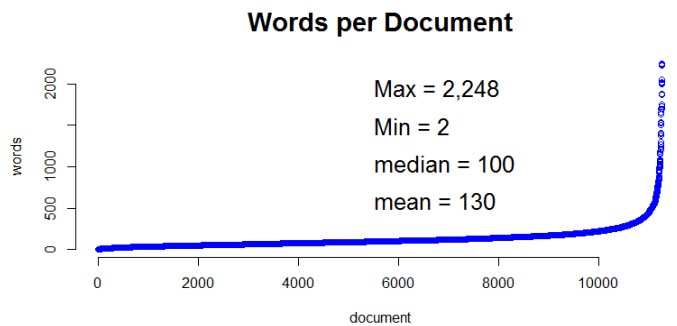


20newsgroup dataset - Rachel

20newsgroup dataset

- **Similar to Verizon dataset (proof of concept for the project):**
  - count documents – high dimension – high sparsity
- **Basic information**
  - newsgroup documents on 20 topics
  - ground truth available
  - row(documents): 11,269
  - column (words): 53,975
  - Density: 0.24%





20newsgroup dataset

Original Sparse Form

docID	wordID	count
doc 1	word 1	1
doc 1	word 2	4
doc 1	word 3	3
doc 1	word 4	1
doc 2	word 1	9
doc 2	word 2	1
doc 2	word 4	2
doc 3	word 1	2
doc 3	word 5	3
...	...	...
doc m	word n	...

Dense Matrix Form

	word 1	word 2	word 3	word 4	word 5	...	word n
doc 1	1	4	3	1	0	...	...
doc 2	9	1	0	2	0	...	...
doc 3	2	0	0	0	3	...	...
...	...	...	...	...	...	...	...
doc m	...	...	...	...	...	...	...

Class of the 20newsgroup data

comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x

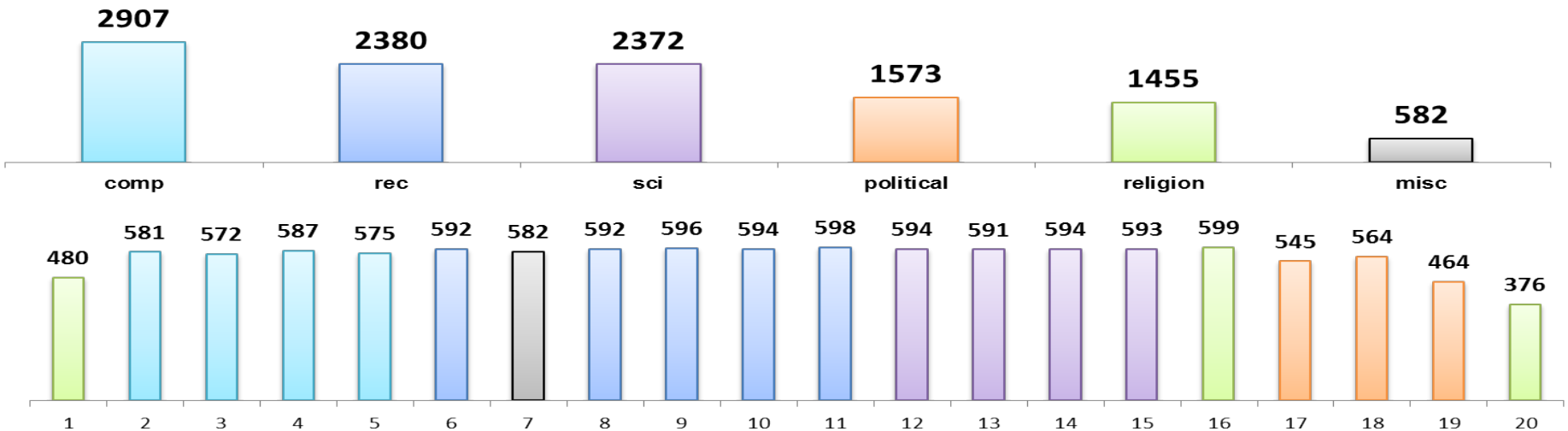
talk.**politics**.misc
talk.**politics**.guns
talk.**politics**.mideast

sci.crypt
sci.electronics
sci.med
sci.space

rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey

talk.**religion**.misc
alt.**atheism**
soc.**religion**.christian

misc.forsale



Data Processing - Ryan

Data processing steps:

1. Binarization: 0-1
2. Column Processing: trimming and weighting
3. Row Processing: trimming and normalization

Binarization

- Convert all non-zero entries to 1
- Indicator of word occurrence in a document
- Trade-off between loss of information and de-emphasis of high frequency terms

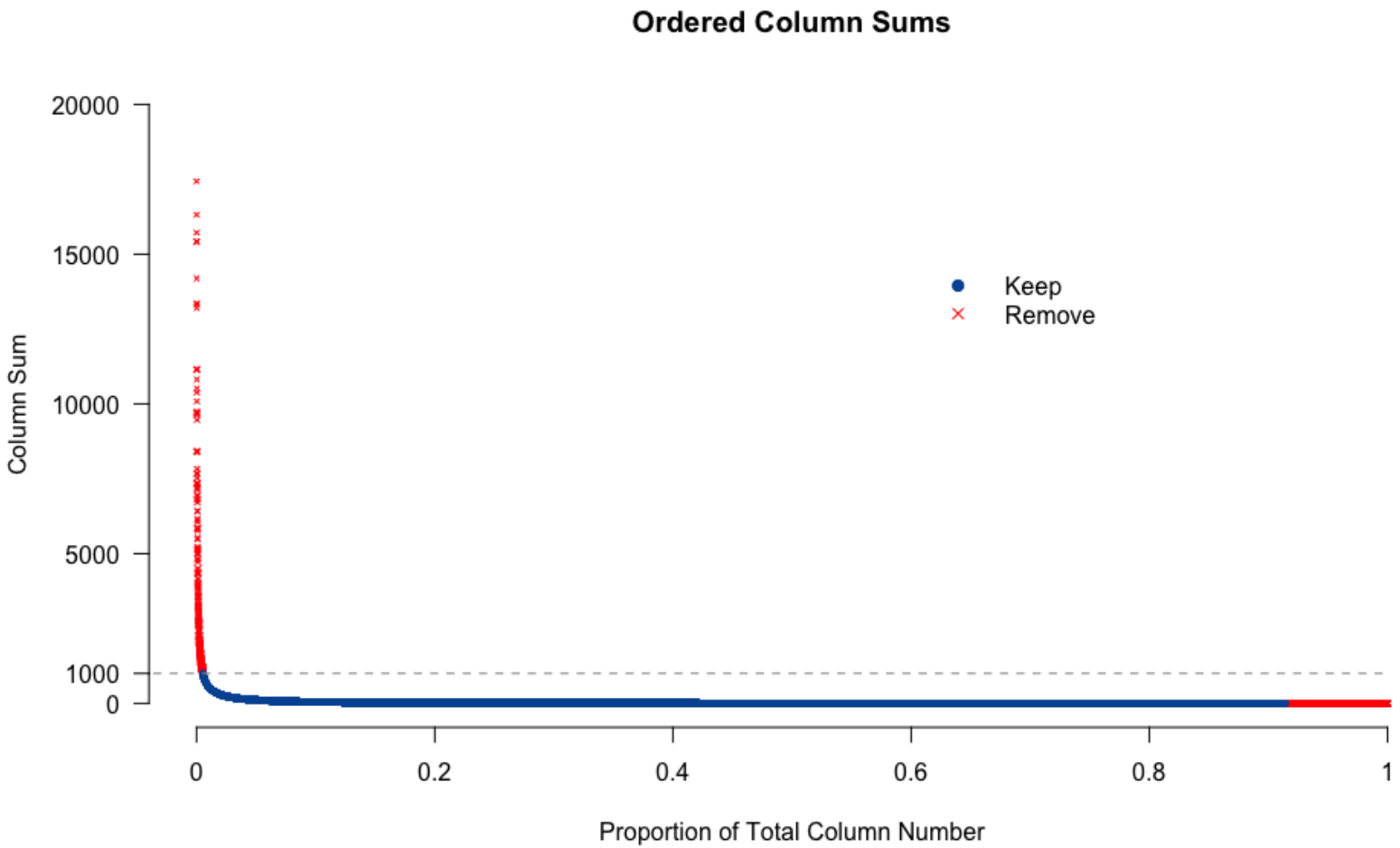
Column Processing

Trimming thresholds:

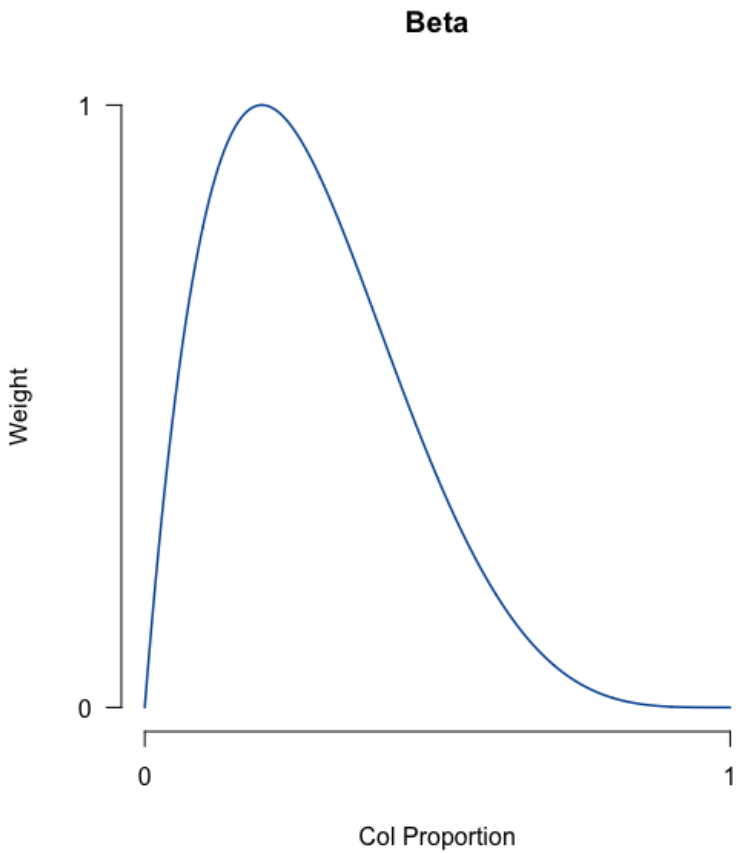
- Min document occurrence: 1
  - Removes 9%
- Max document occurrence: >1000
  - Removes 0.5%

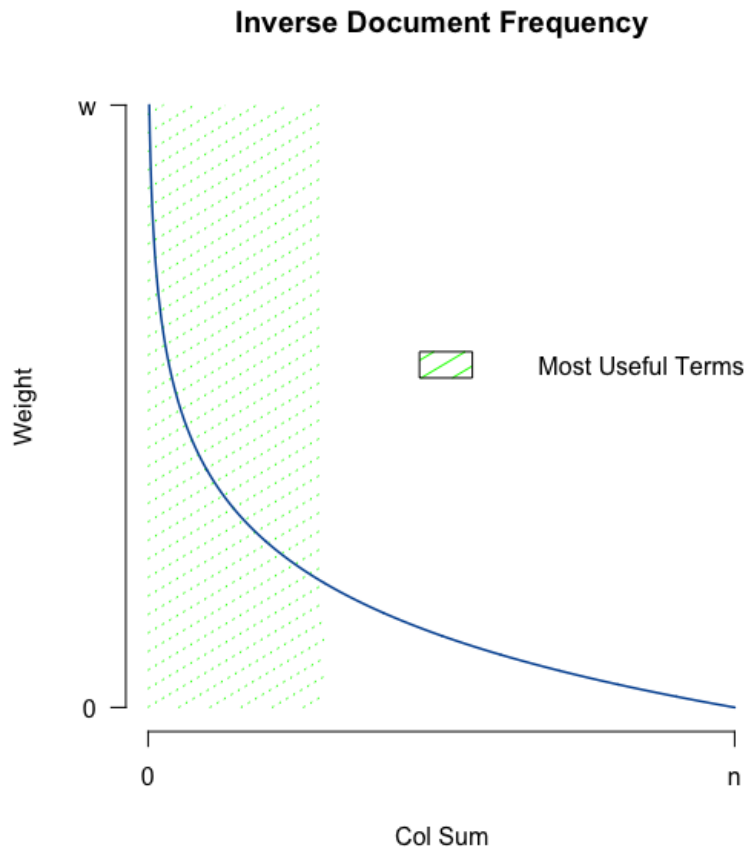
	$Term_1$	$Term_2$	$Term_3$	$Term_4$	. . .	$Term_p$
Doc_1	1	0	0	0	. . .	1
Doc_1	1	1	1	0	. . .	0
Doc_1	0	1	0	0	. . .	1
.						
.					Document Term Matrix	
.						
Doc_N	0	1	1	1	. . .	0
Column Sums	42	1251	7	1		23

20newsgroup dataset



Column Weighting

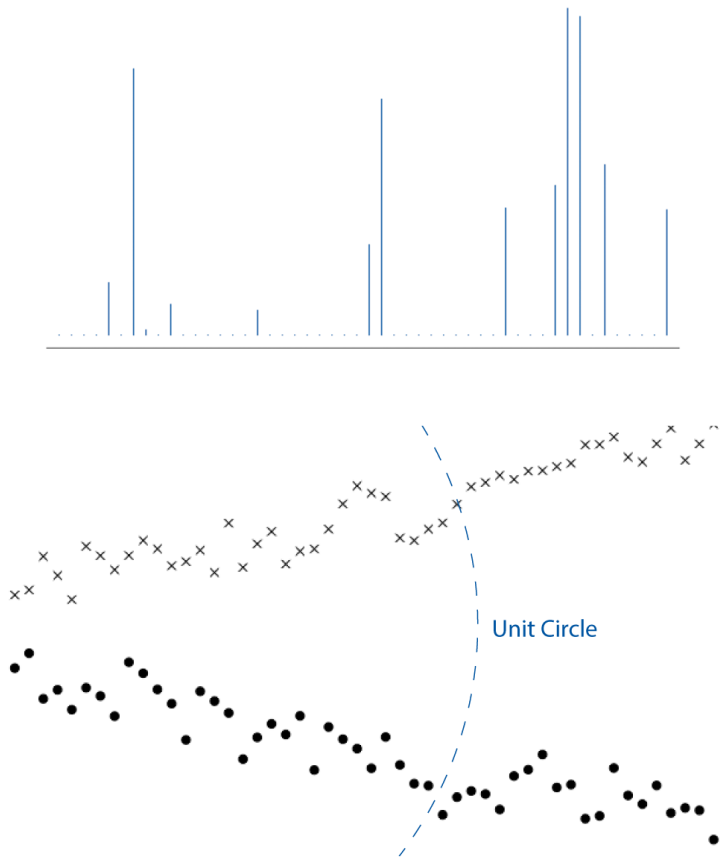




Row Processing

Normalization

- L1 - $\frac{W}{||W||_1}$
- L2 - $\frac{W}{||W||_2}$



Dimensionality Reduction - Qingbin

Motivation

- Feature Extraction
- Curse of dimensionality
  - sparsity, problematic for any method that requires statistical significance
  - Noise, covers the true structure or pattern in the data
  - Time Consuming, Memory demanding

Idea of Latent Semantic Indexing

- Assume lower dimensions reveal latent features of original data space
- Map documents (and terms) to this lower dimensional space
- Reduced data reflects true structure of original data
- Compute document similarity based on Reduced data

SVD (Mathematical LSI)

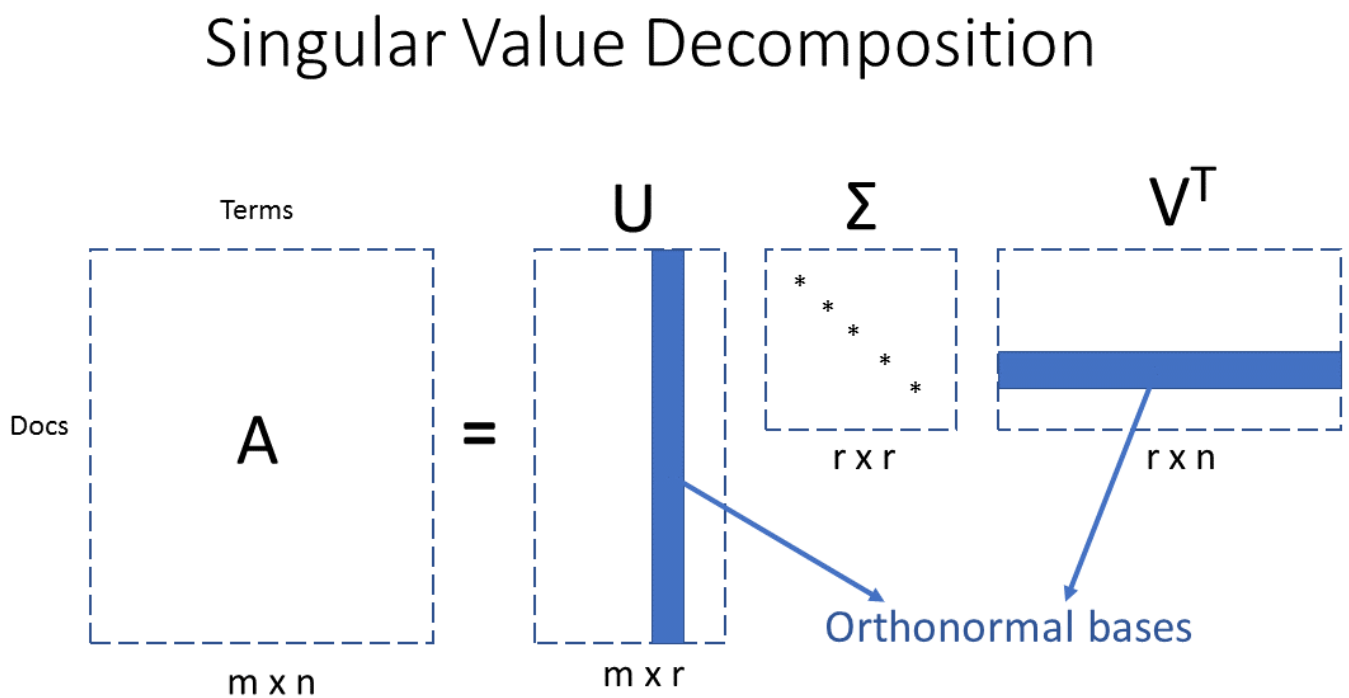
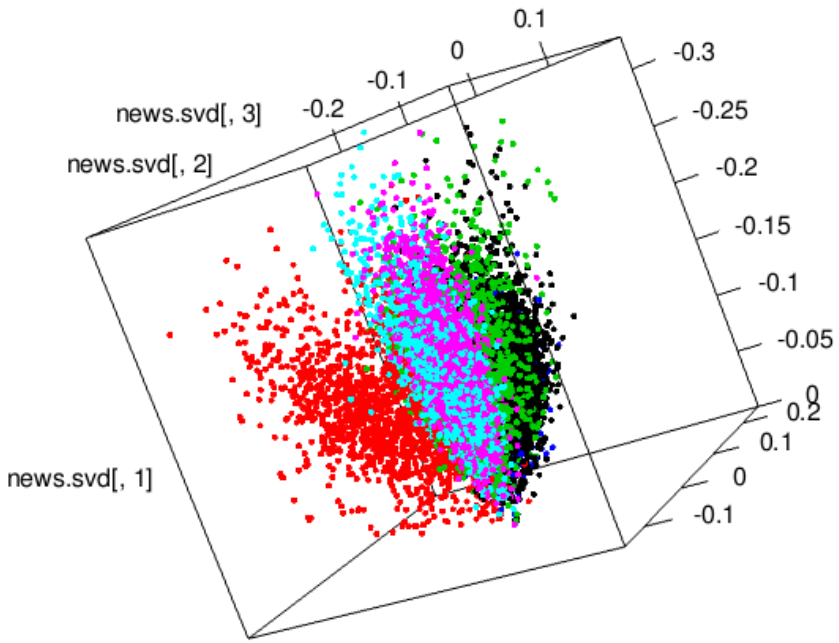


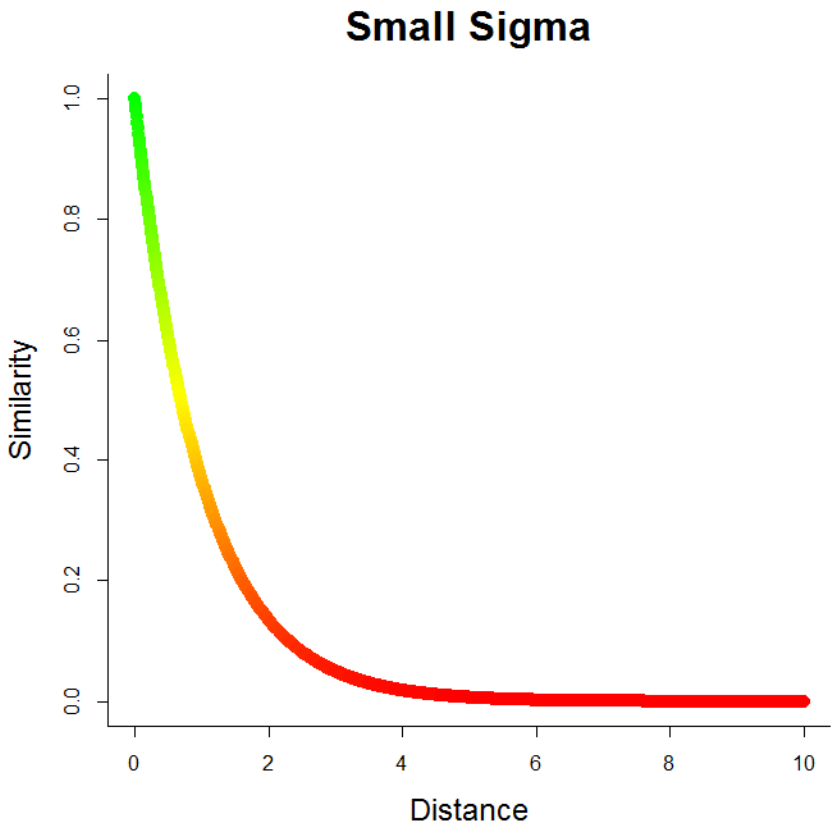
Illustration of 20newsgroup data with SVD

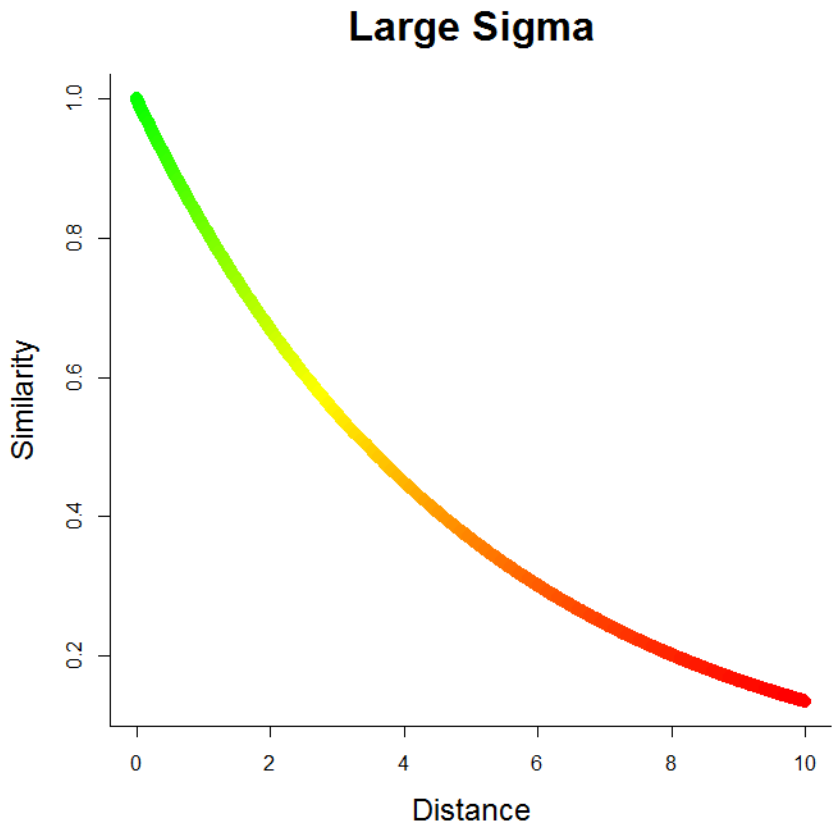
- Significant structure
- Basis for similarity calculation



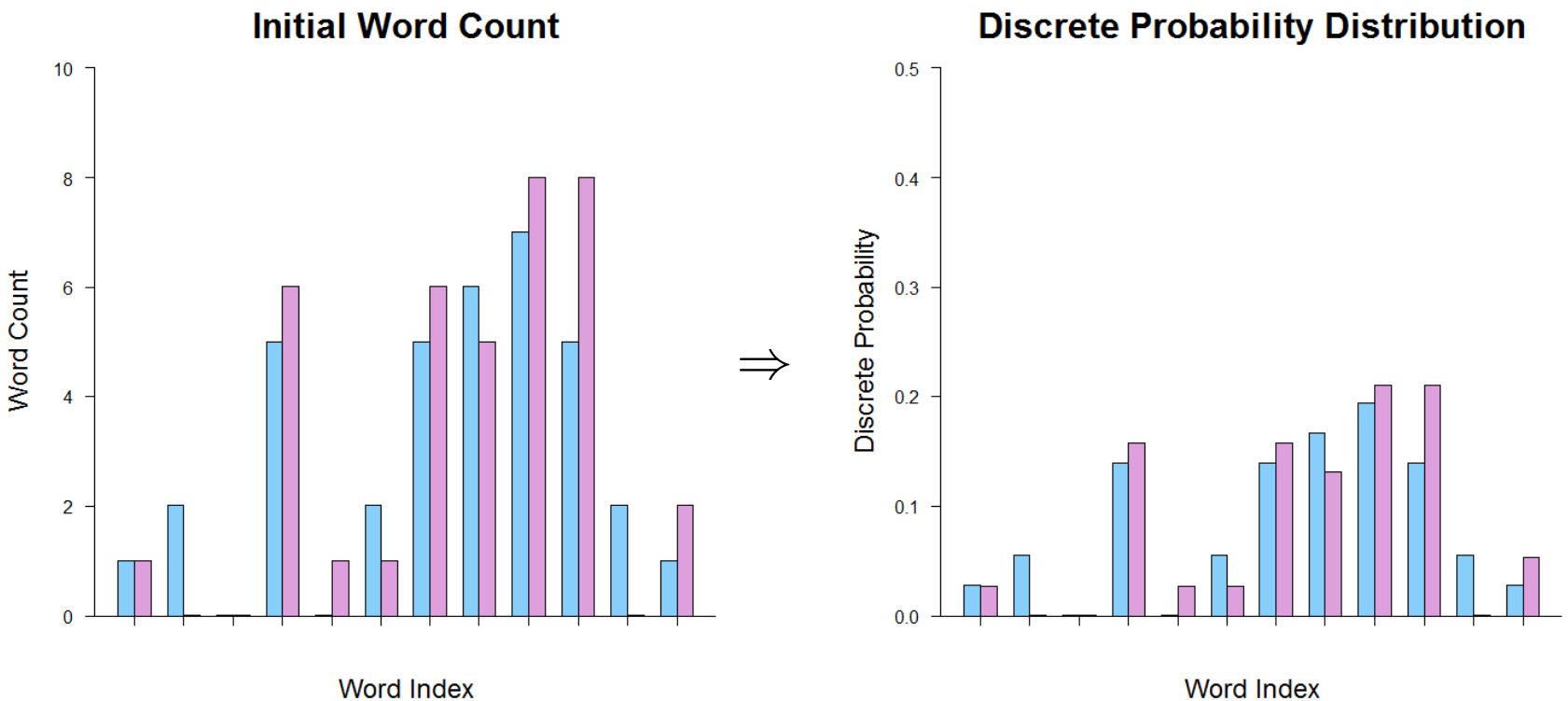
Similarity - Joey

1) Gaussian Kernel: $\text{Sim}(x, y) = e^{\frac{-\text{dist}(x,y)^2}{2\sigma^2}}$

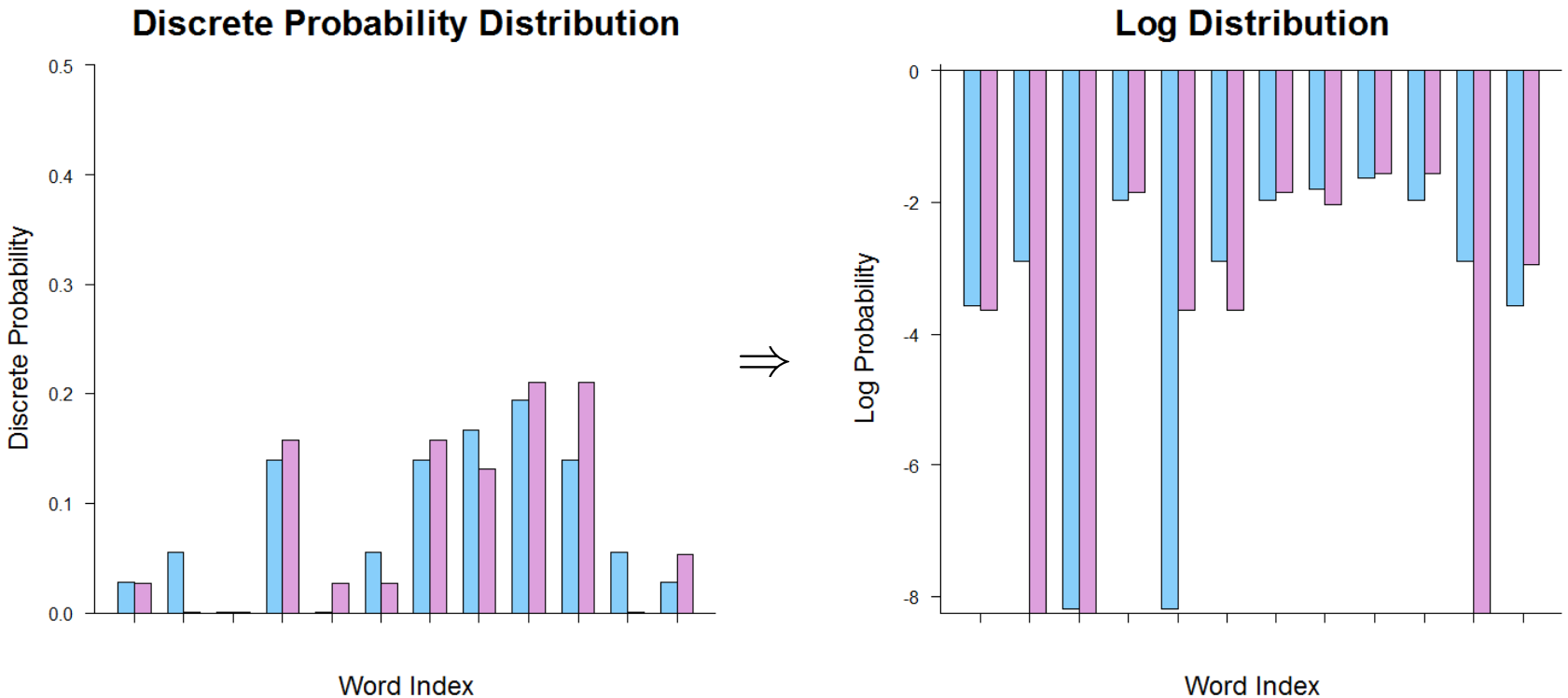




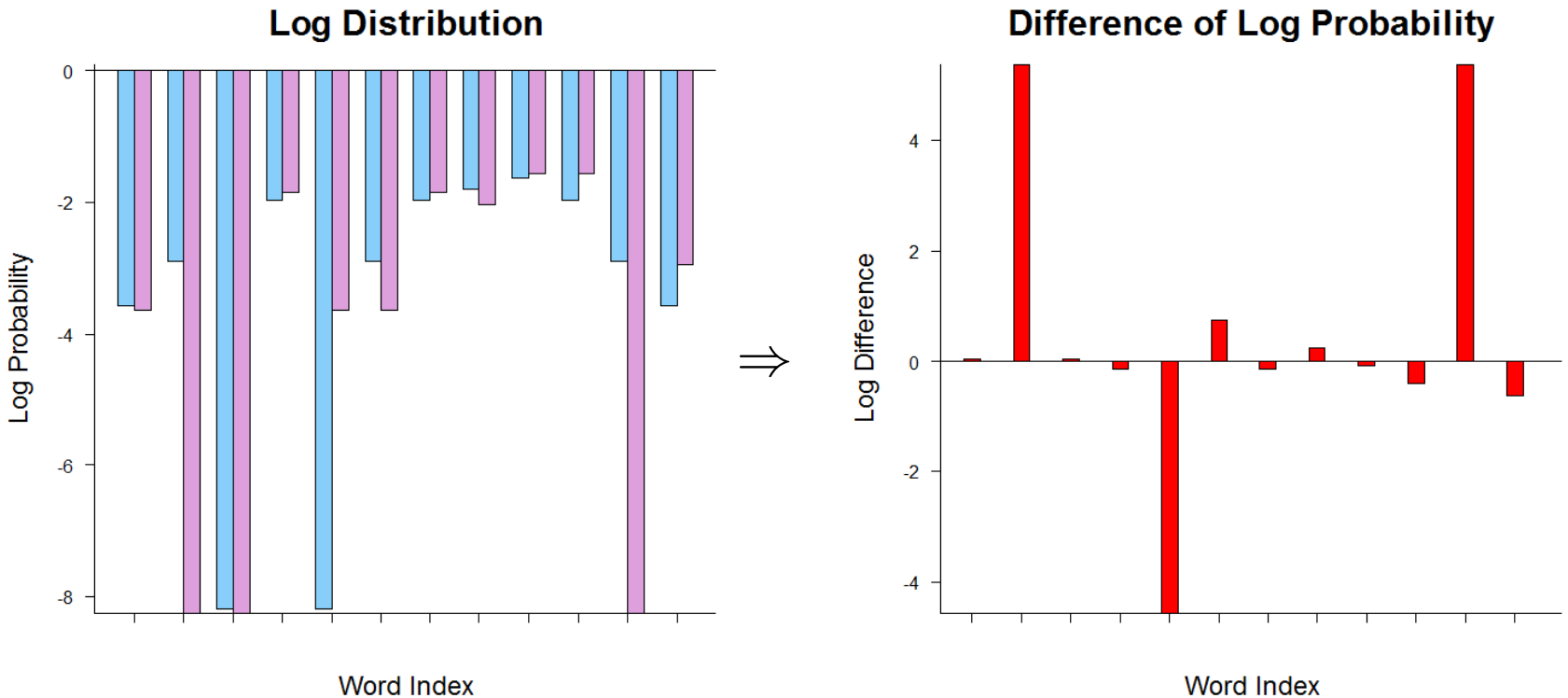
1.1) Kullback-Leibler Divergence:



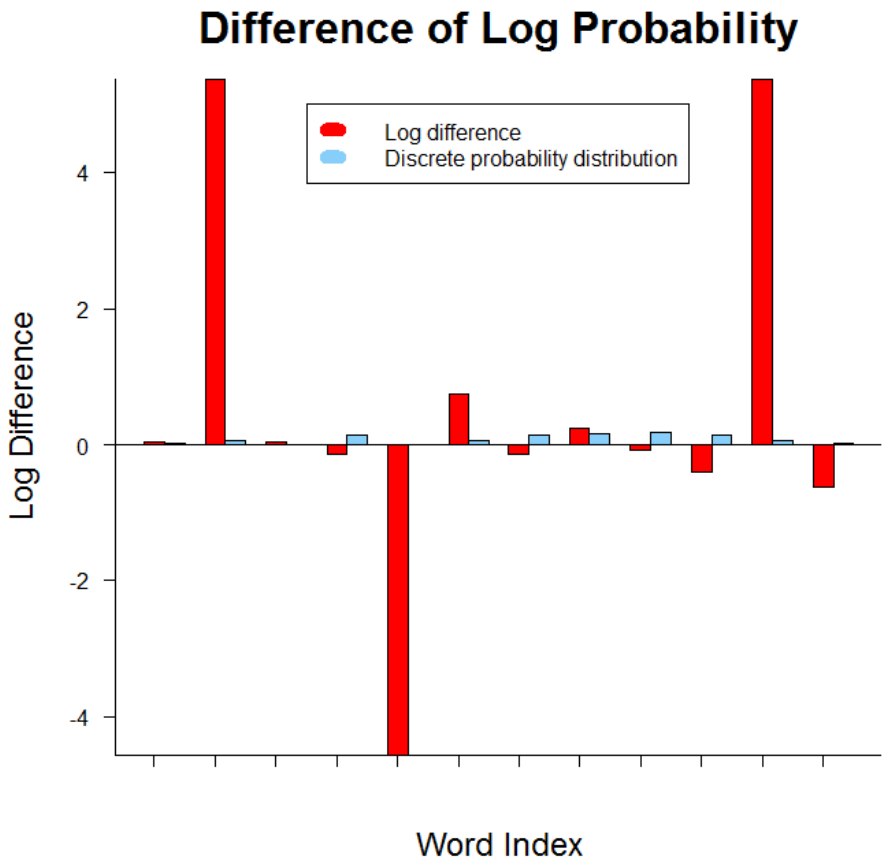
1.1) Kullback-Leibler Divergence:



1.1) Kullback-Leibler Divergence:



1.1) Kullback-Leibler Divergence:



For discrete distributions x and y:

Divergence(x, y):

$$= \mathrm{E}\left[\log(x) - \log(y)\right]$$

$$= \sum_k x_k \left[\log\left(\frac{x_k}{y_k}\right)\right]$$

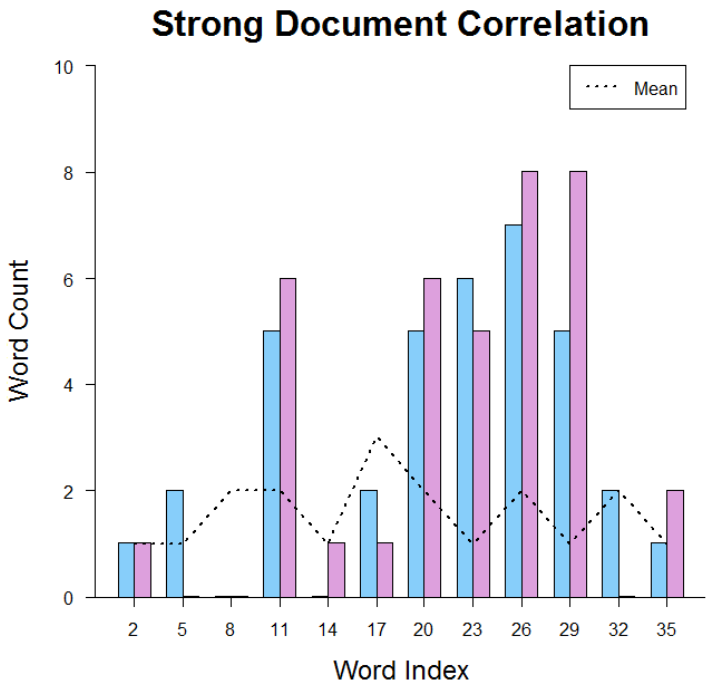
1.2) Jensen-Shannon Divergence:

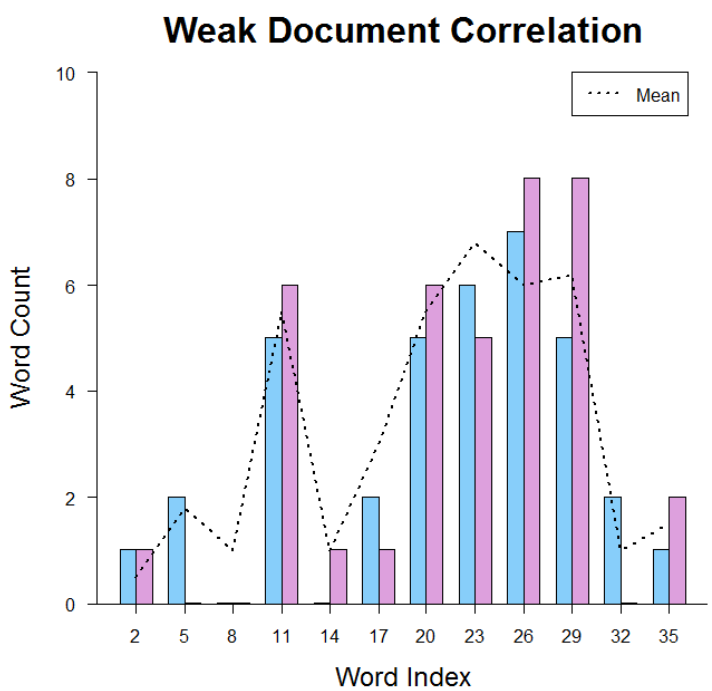
"Average Distribution" Symmetry: $M = \frac{x+y}{2}$

$$\text{JS.Div}(x, y) = \frac{\text{Divergence}(x,M) + \text{Divergence}(y,M)}{2}$$

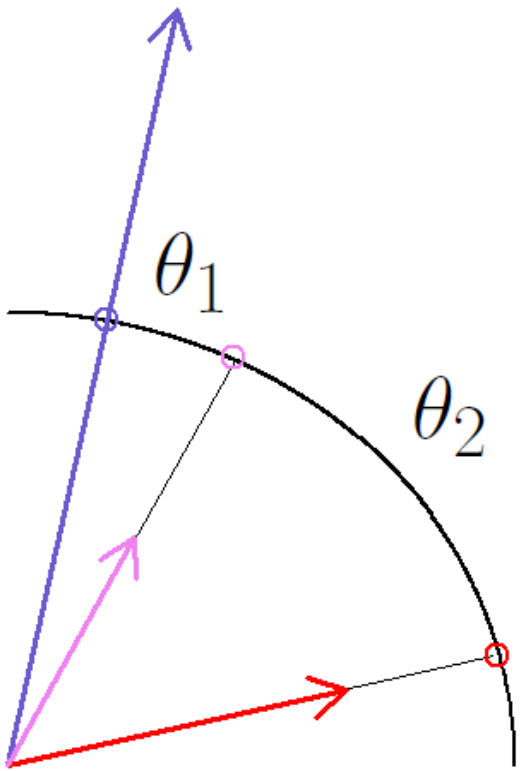
$$\Rightarrow \text{Sim}(x, y) = e^{\frac{-\text{JS.Div}(x,y)^2}{2\sigma^2}}$$

2) Correlation: $\text{Sim}(x,y) = \dfrac{(\vec{x}-\vec{\mu})\cdot(\vec{y}-\vec{\mu})}{\sqrt{||\vec{x}-\vec{\mu}||^2\ ||\vec{y}-\vec{\mu}||^2}}$





3) Cosine: $\text{Sim}(x, y) = \dfrac{\vec{x} \cdot \vec{y}}{\sqrt{||\vec{x}||^2 \, ||\vec{y}||^2}} = \cos(\theta_{xy})$



Outliers Removal - Nate

Outlier Removal

What are outliers?

- Documents that have low connectivity to other documents
- Documents that have low information
  - Very common words
  - Very few words

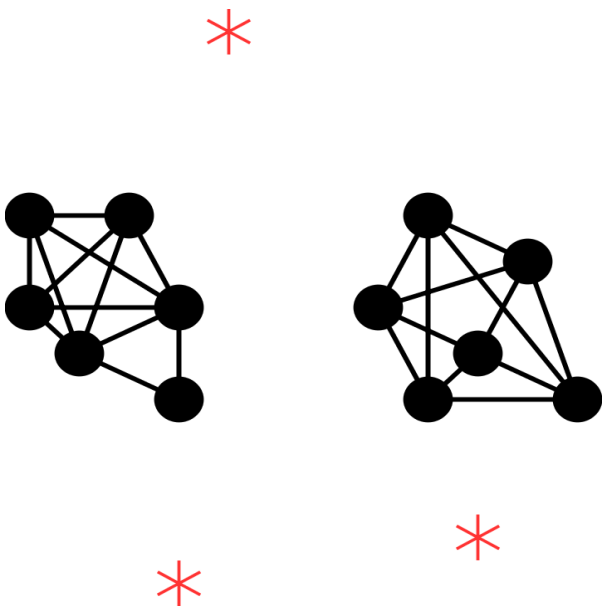
Determining Connectivity or Information

- Row sums of our data matrix:

    1. Raw data tells us how many words are in the document

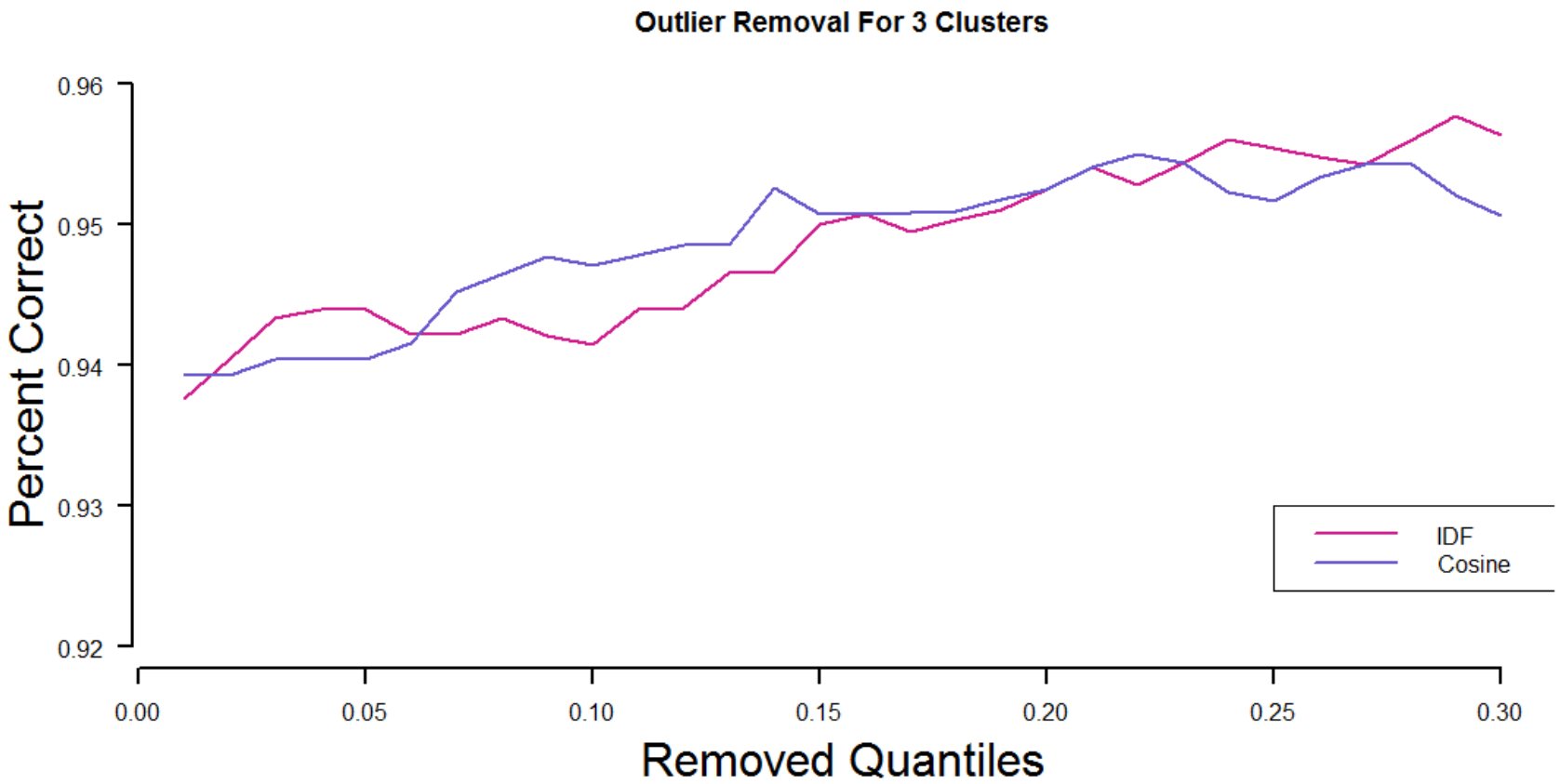
    2. IDF weighted data tells us how many *useful* words are in the document

Outlier Removal

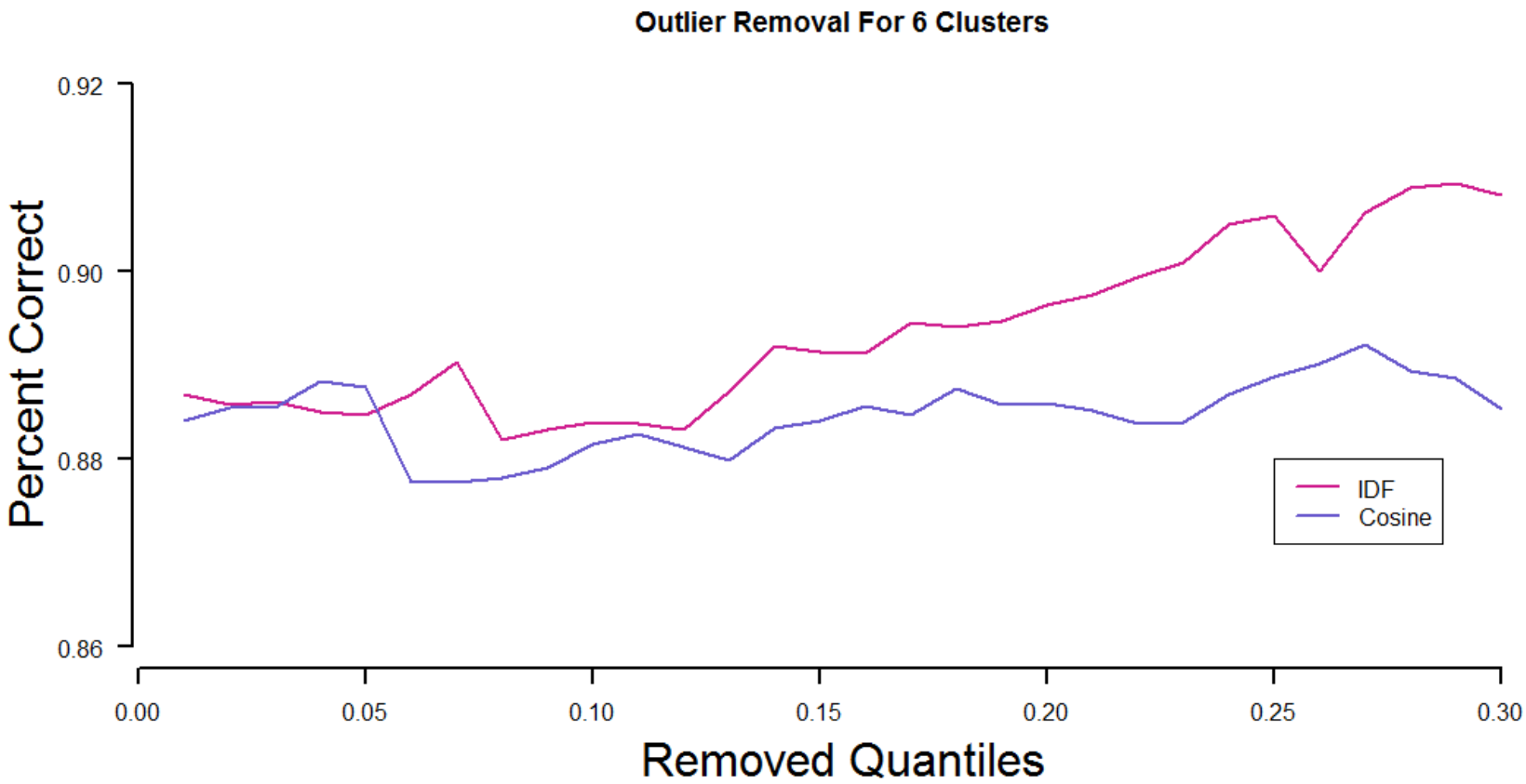


- Row sums of the cosine similarity matrix count as a measure of connectivity to other documents

Outlier Removal

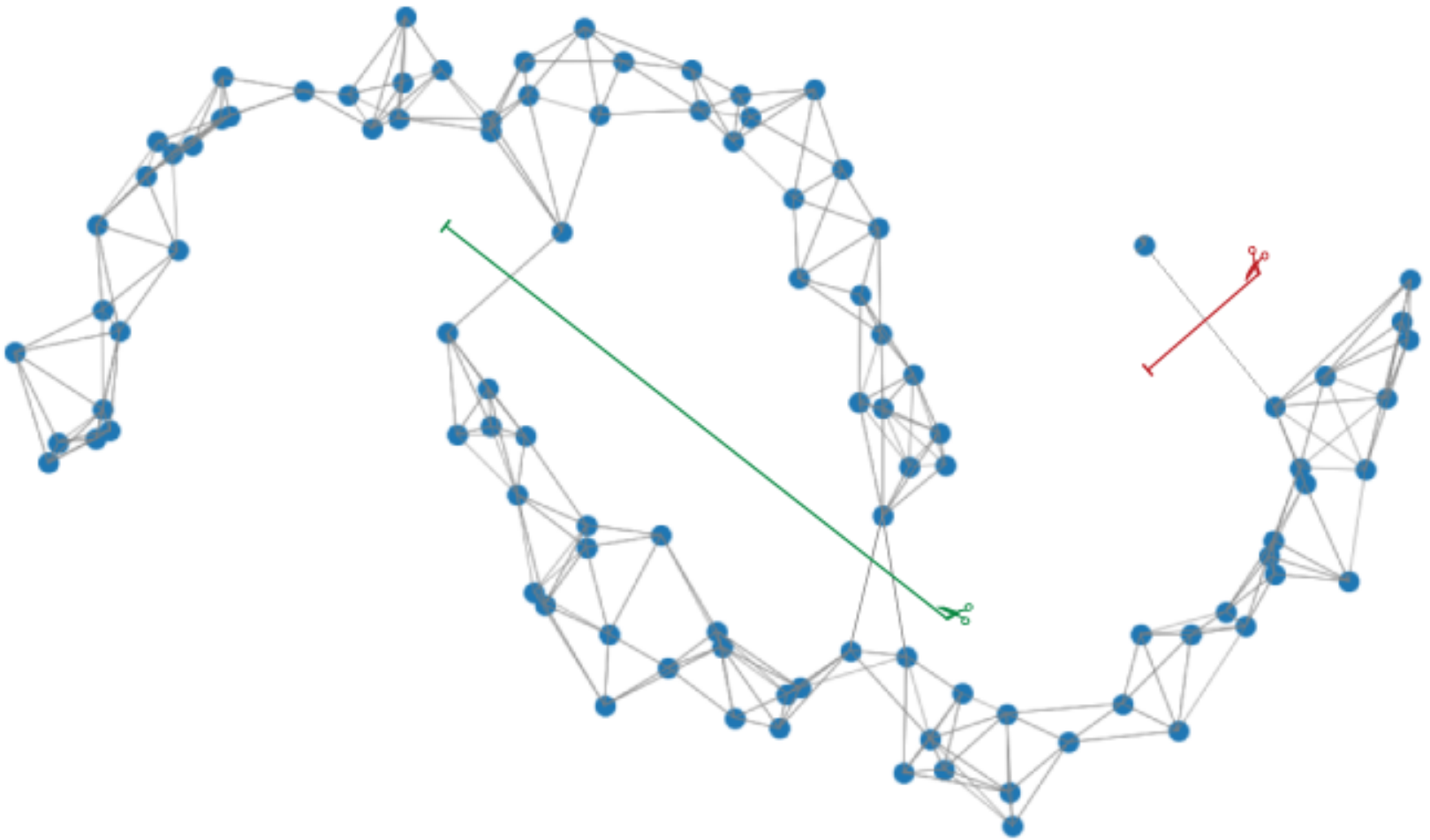


Outlier Removal



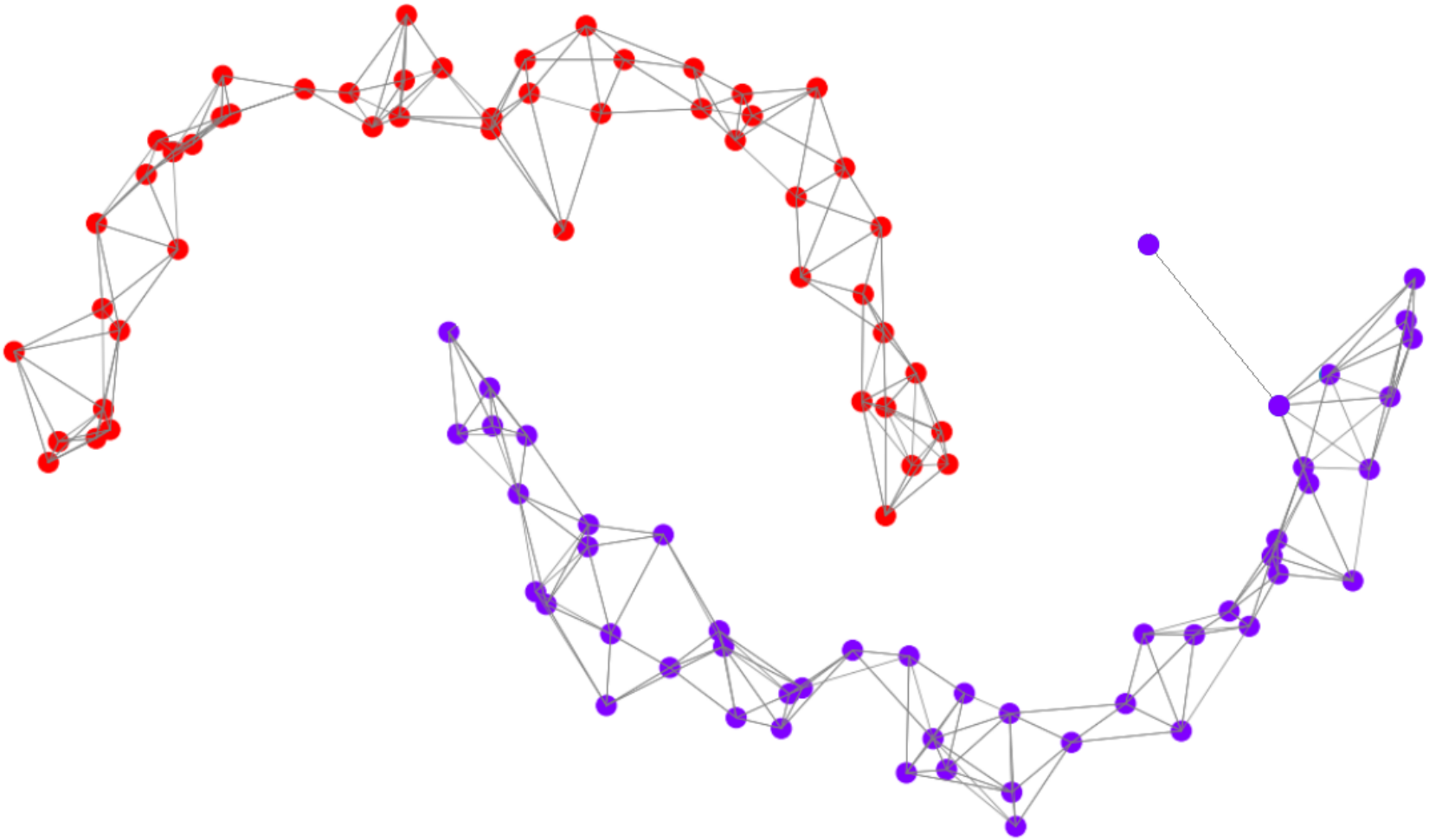
Spectral Clustering - Andrew

Spectral Clustering- Normalized Cut



$= \min Cut(A, B)$ $= \min \frac{Cut(A,B)}{Vol(A)} + \frac{Cut(A,B)}{Vol(B)}$

An ideal way to remove edges to create two clusters



Let D be a diagonal matrix where $D_{ii} = \sum_{j=1}^{n} w_{ij}$

Also let $x_i = \begin{cases} 1, \text{ if the observation i is in cluster A} \\ -1, \text{ otherwise} \end{cases}$

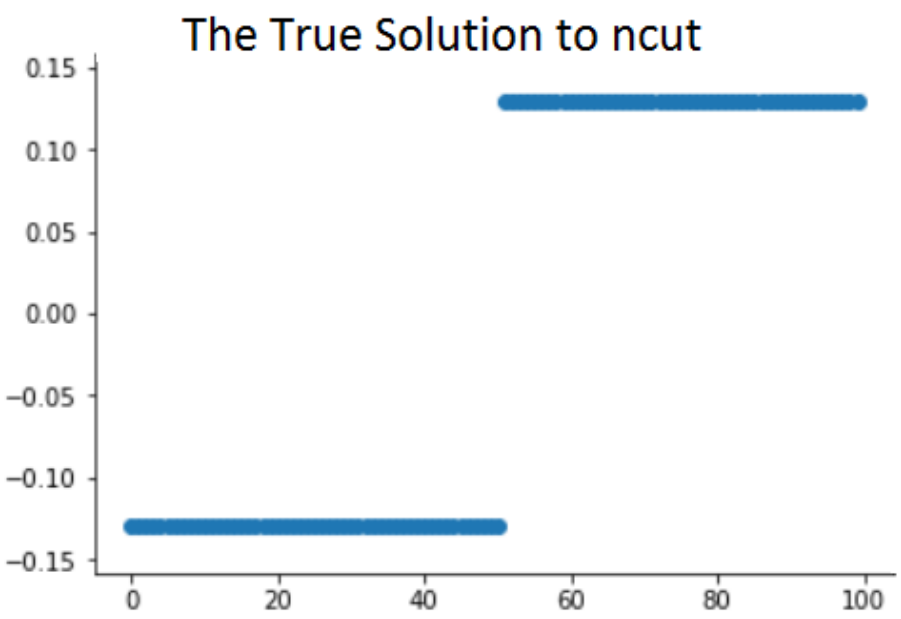
It can be shown that

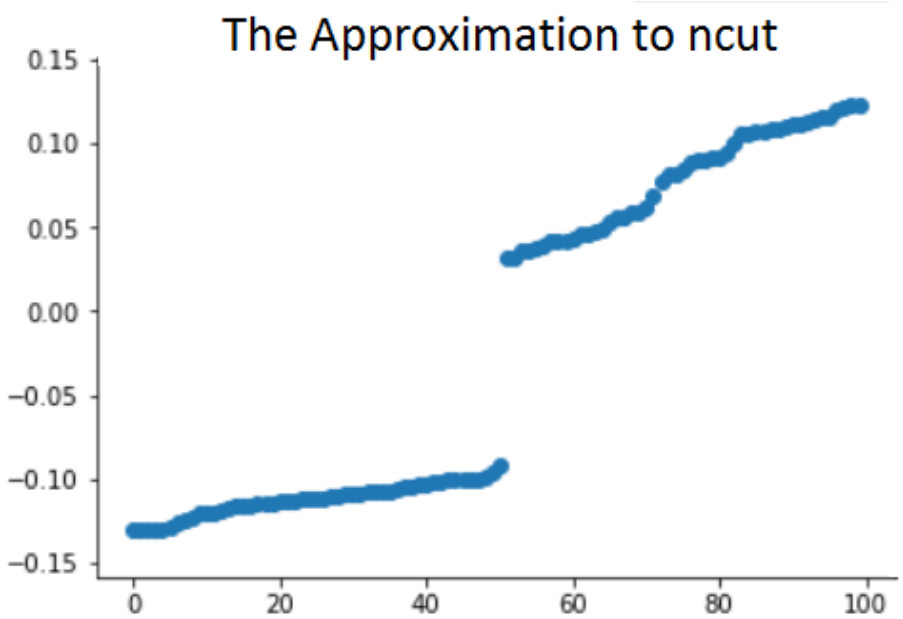
$$\min \left(\frac{Cut(A,B)}{Vol(A)} + \frac{Cut(A,B)}{Vol(B)} \right) = \min_{\mathbf{y}} \frac{\mathbf{y}^T \mathbf{L} \mathbf{y}}{\mathbf{y}^T \mathbf{D} \mathbf{y}}$$

Where $y = (1 + \mathbf{x}) - \frac{Vol(A)}{Vol(B)}(1 - \mathbf{x})$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

The solution to the normalized cut problem can be approximated by the sign of the second largest eigenvector of $\mathbf{D}^{-1}\mathbf{L}$

Spectral Clustering- Normalized Cut





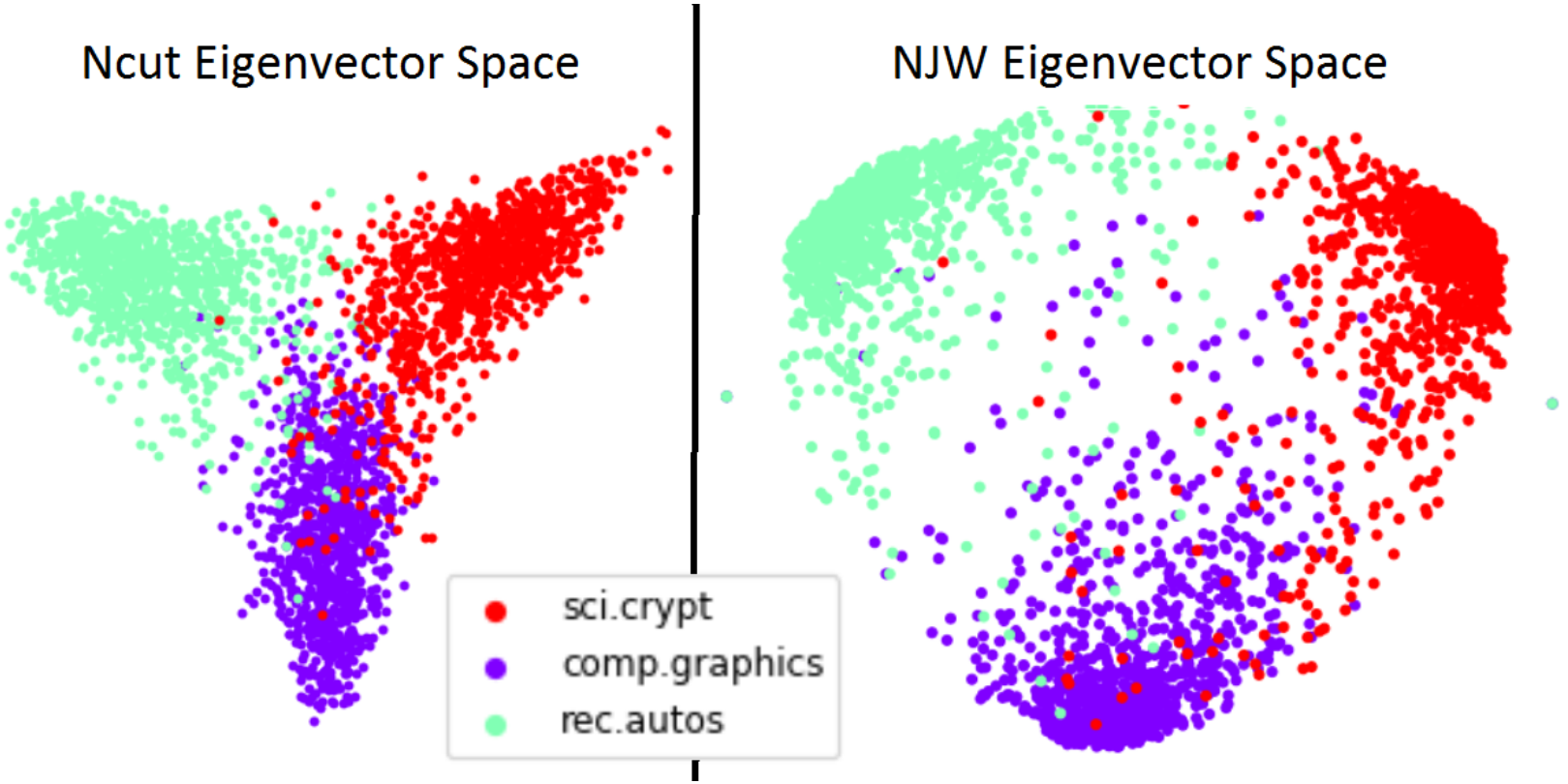
Algorithm (Ncut)

1. Construct similarity matrix $\mathbf{W}$
2. $\mathbf{L} = \mathbf{D} - \mathbf{W}$
3. Find the first k eigenvectors of $\mathbf{D}^{-1}\mathbf{L}$
4. Make a matrix $\mathbf{V}$ by stacking the 2nd to kth eigenvectors
5. Cluster using kmeans using $\mathbf{V}$ where each row represents a point

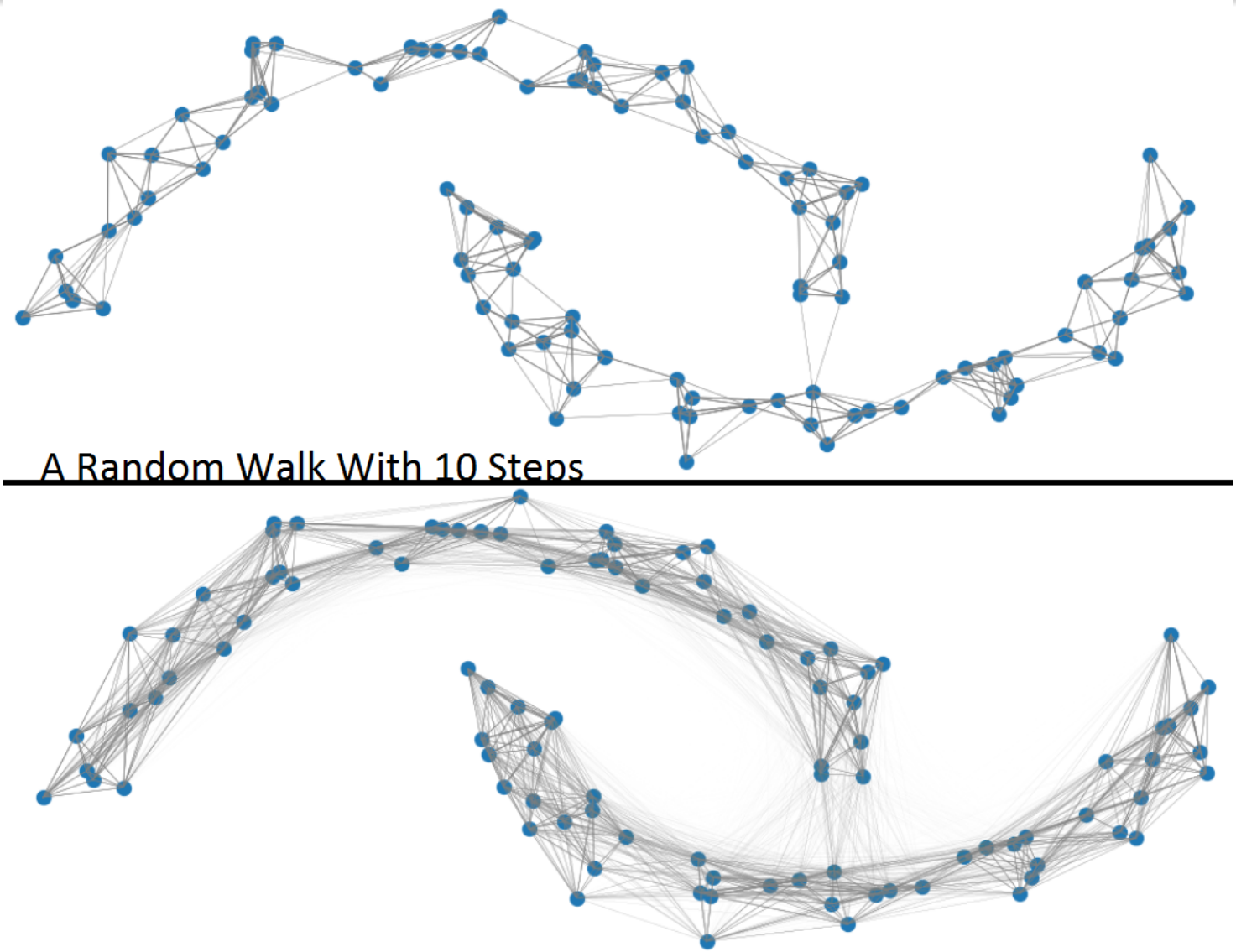
Algorithm (NJW)

1. Construct similarity matrix $\mathbf{W}$
2. $\mathbf{L} = \mathbf{D} - \mathbf{W}$
3. Find the first k eigenvectors of $\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$
4. Make a matrix $\mathbf{V}$ by stacking the first k eigenvectors
5. Normalize the rows of $\mathbf{V}$
6. Cluster using kmeans using $\mathbf{V}$ where each row represents a point

Spectral Clustering- Ng, Jordan, Weiss



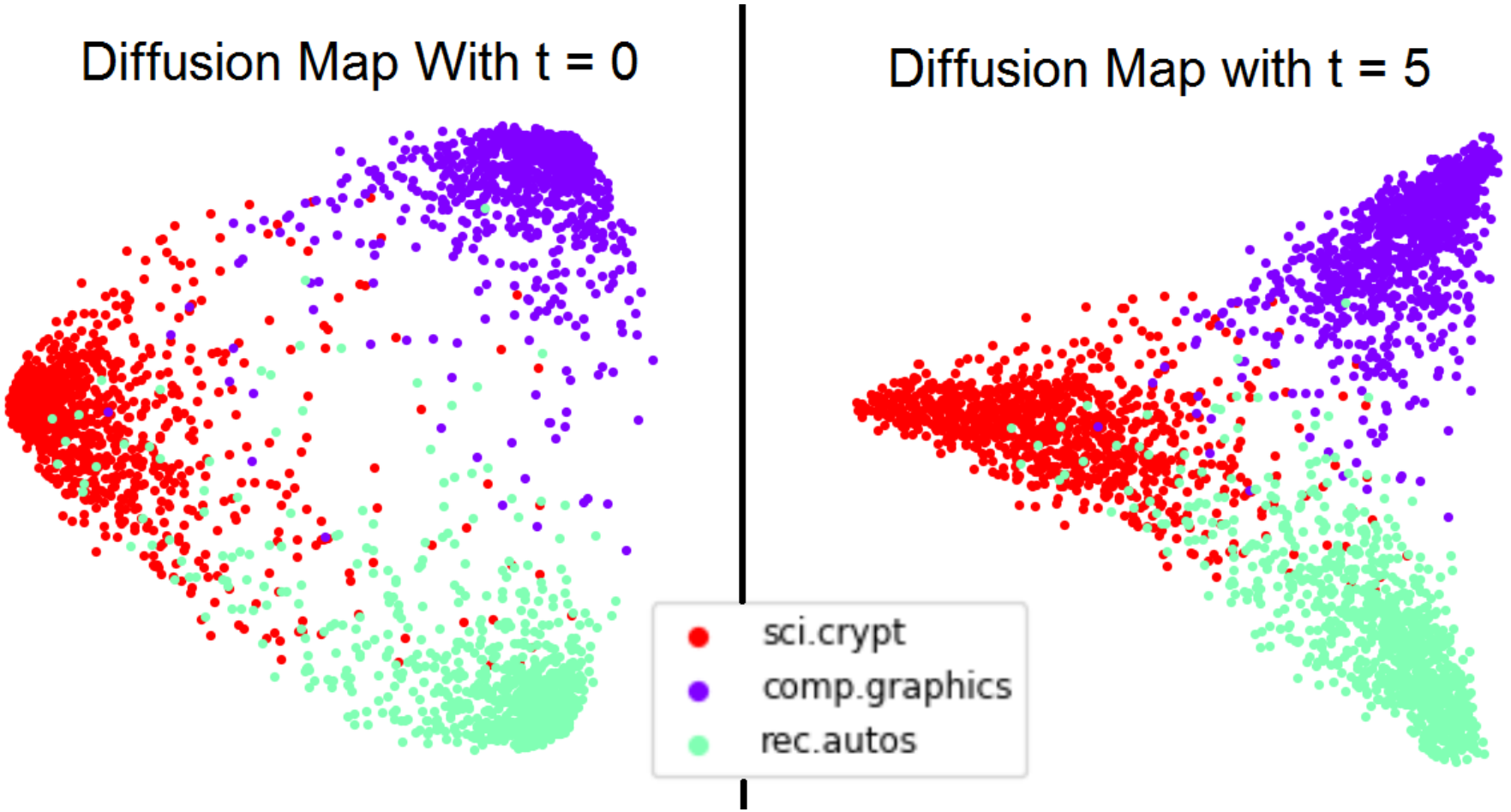
Spectral Clustering- Diffusion Map



Algorithm (Diffusion Map)

1. Construct similarity matrix $\mathbf{W}$
2. $\mathbf{L} = \mathbf{D} - \mathbf{W}$
3. Find the first k eigenvectors of $\mathbf{D}^{-1}\mathbf{L}$
4. Make a matrix $\mathbf{V}$ by stacking the 2nd to kth eigenvectors
5. Normalize the rows of $\mathbf{V}$
6. $V = (\lambda_1^t v_1, \lambda_2^t v_2, ..., \lambda_k^t v_k)$
7. Cluster using kmeans using $\mathbf{V}$ where each row represents a point

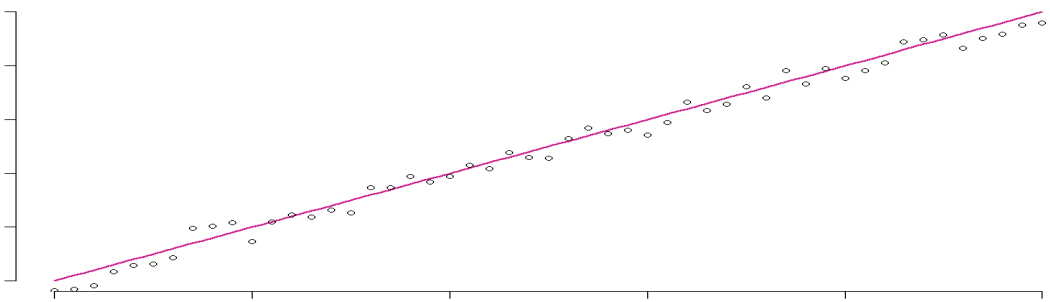
Spectral Clustering- Diffusion Map



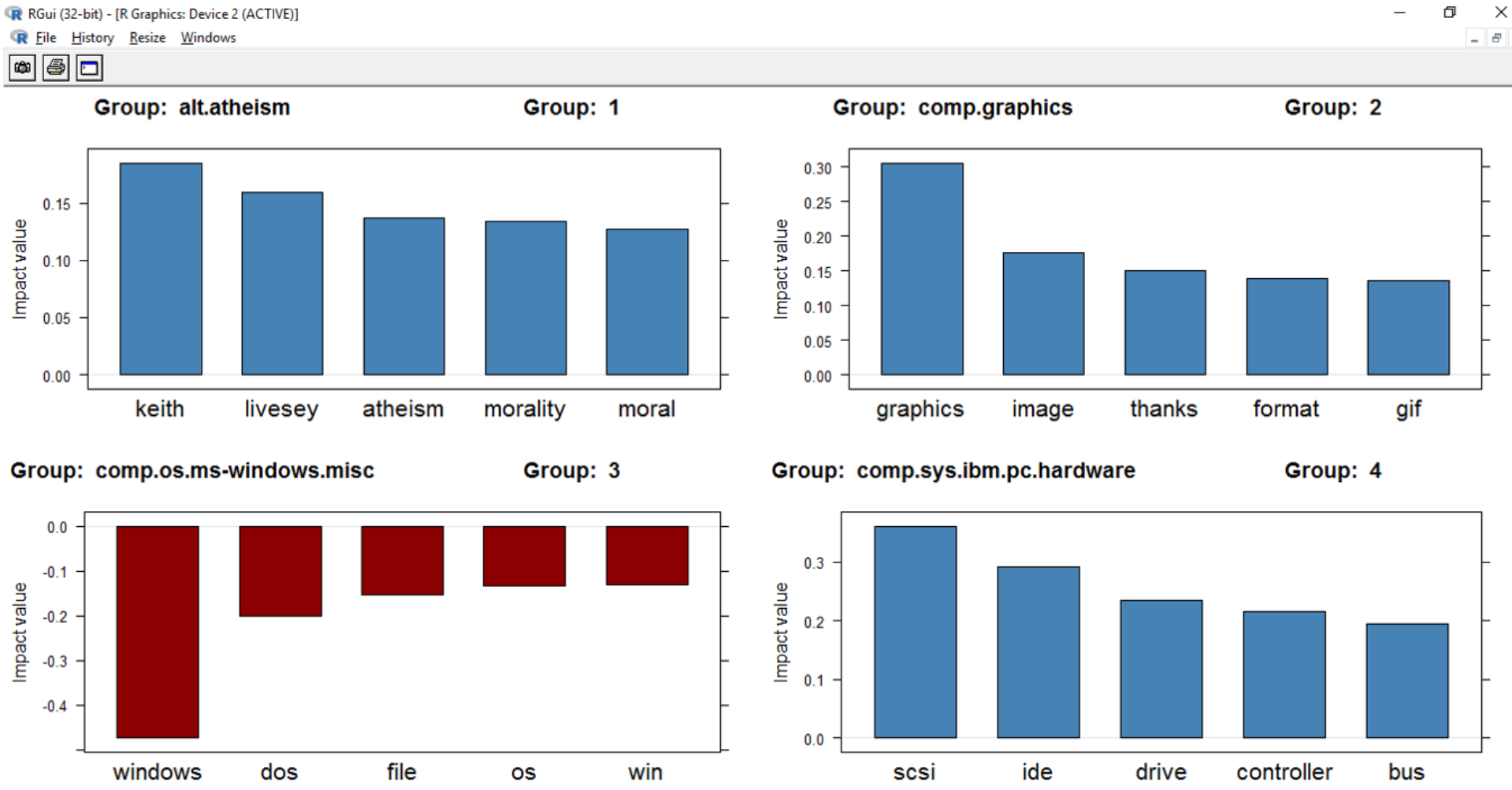
Clustering Insights - Nate

Insights

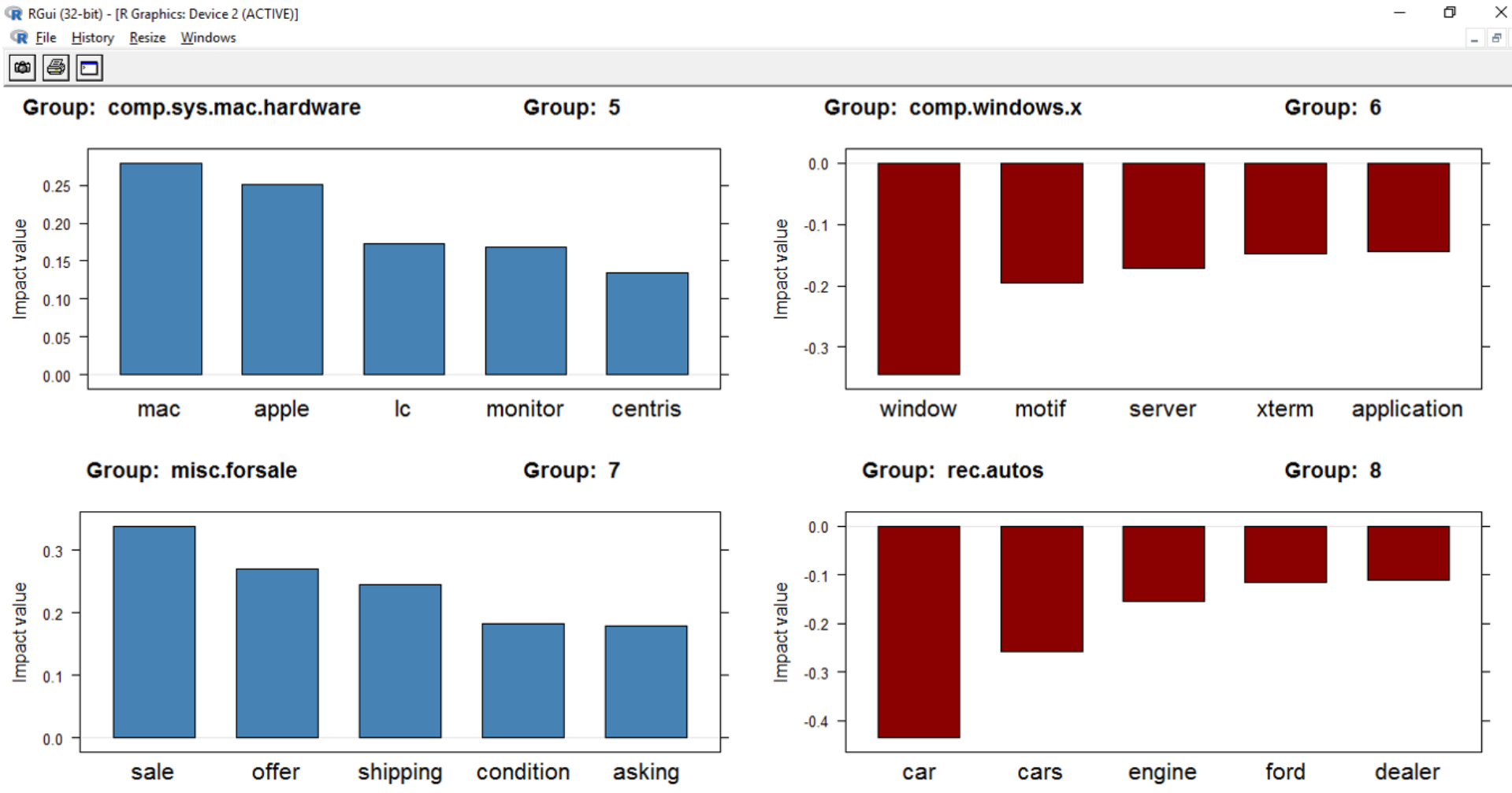
- We want to determine the columns that explain the principle "direction" of each cluster.
- This is done by finding the Principal Component vector after SVD



Cluster Insight



Cluster Insight



20 Newsgroup Clustering Results - Shiou-Shiou

Data and Tasks

comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x

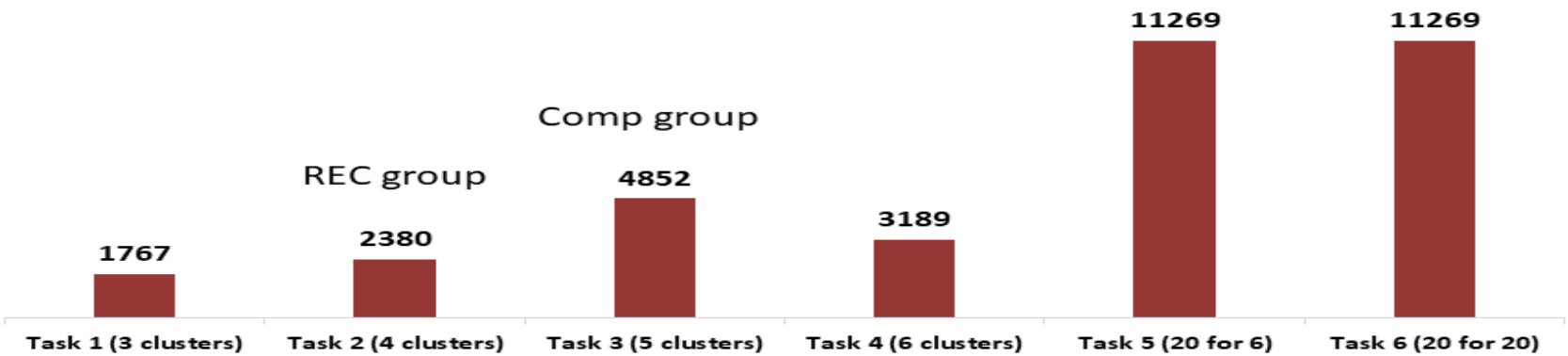
talk.**politics**.misc
talk.**politics**.guns
talk.**politics**.mideast

sci.crypt
sci.electronics
sci.med
sci.space

rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey

talk.**religion**.misc
alt.**atheism**
soc.**religion**.christian

misc.forsale

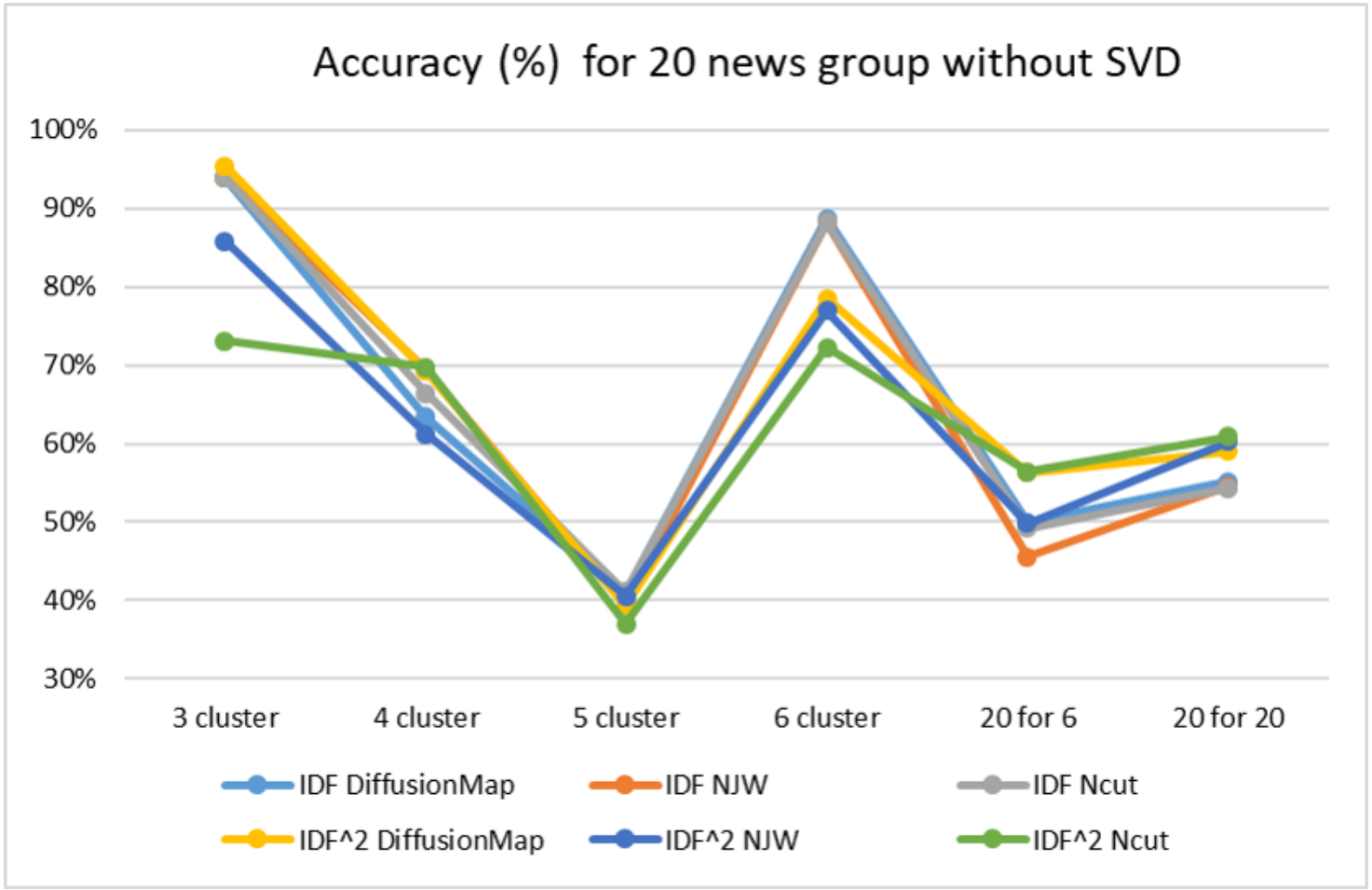


Measurements

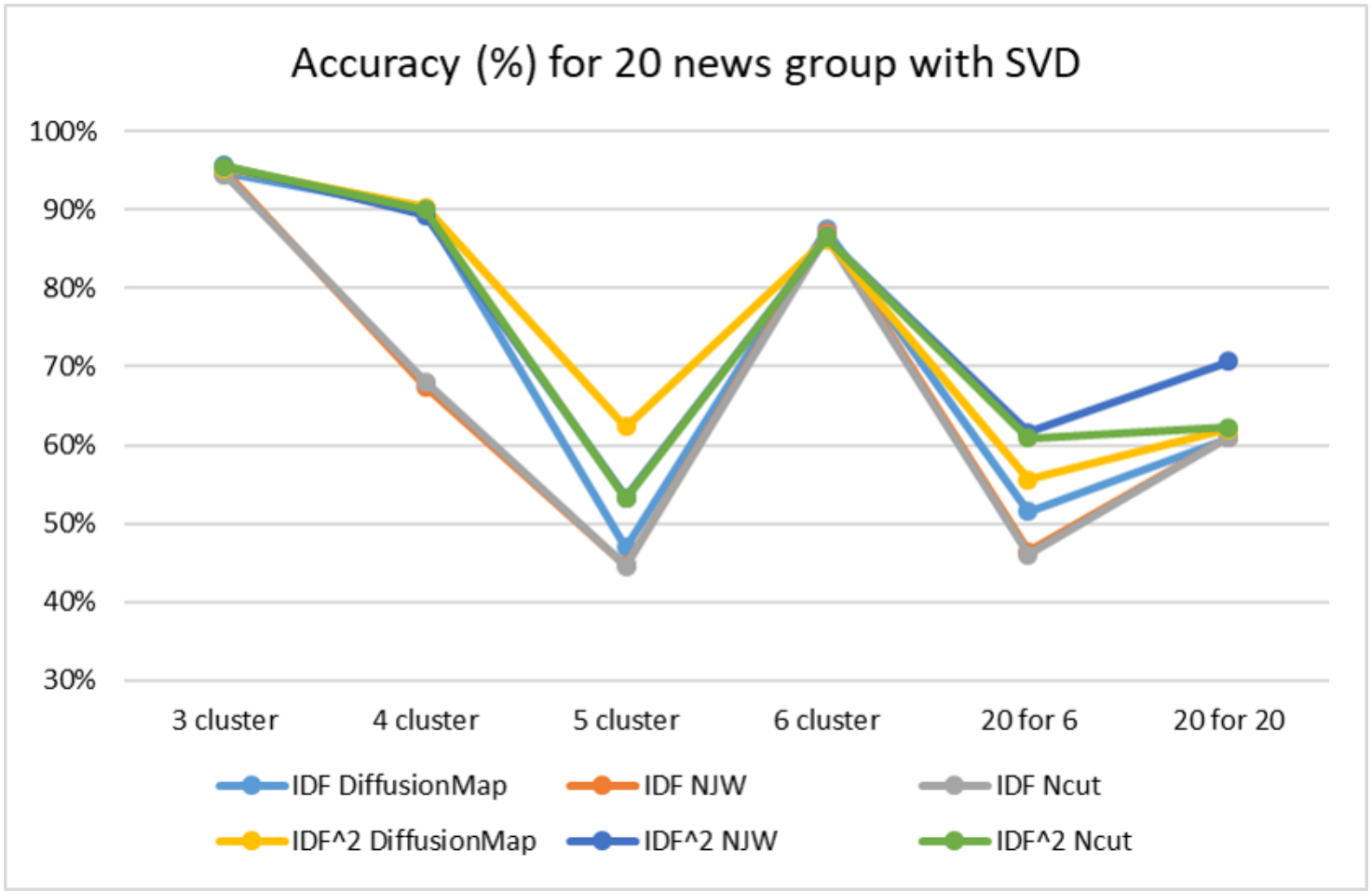
- Accuracy: The percentage of data points that are truly in the same cluster are predicted to be in the same cluster.

- Adjusted Rand Index (ARI), F-measure,..

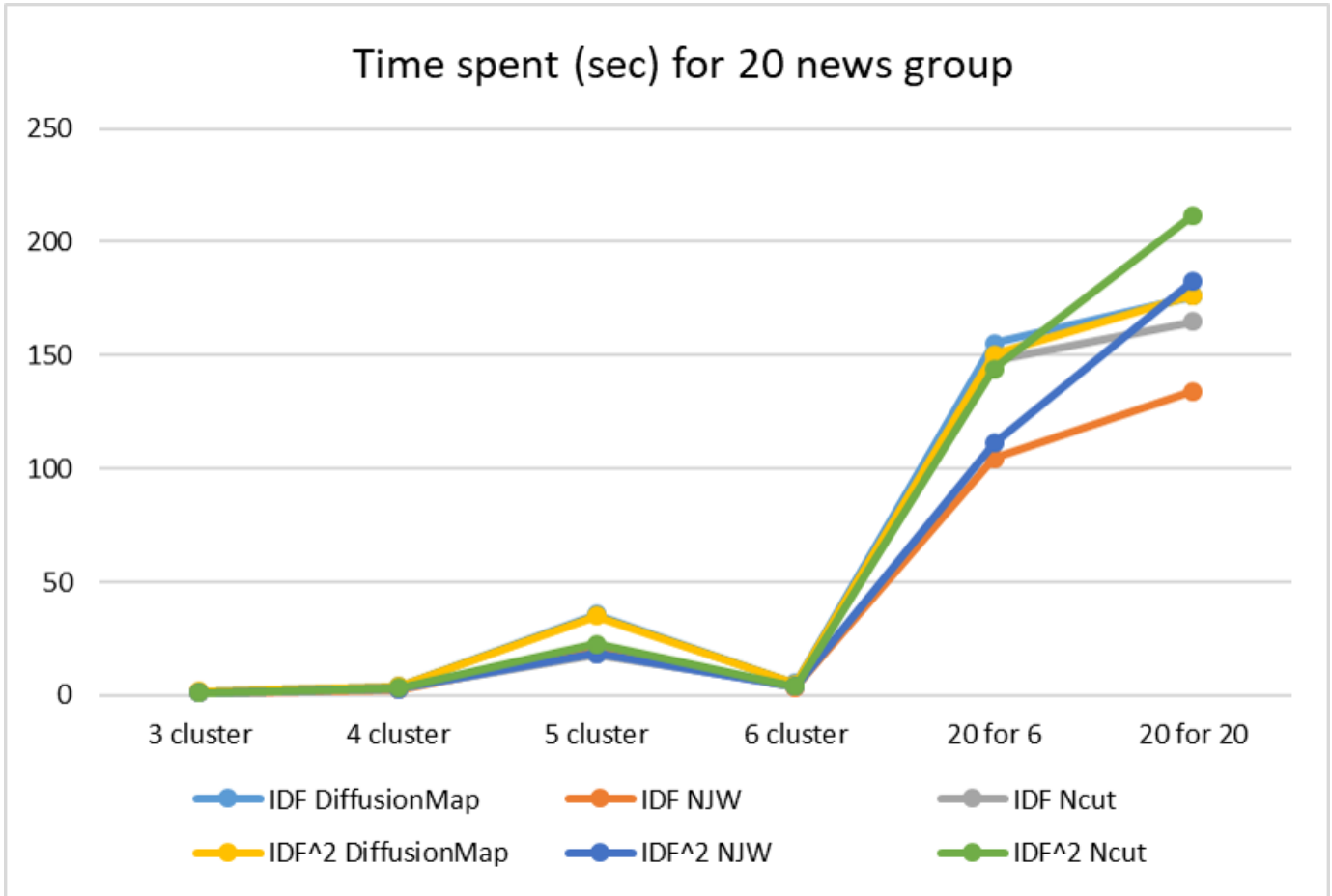
20 Newsgroup Clustering Results



20 Newsgroup Clustering Results



20 Newsgroup Clustering Results



Future Work - Joey

Efficiency Improvements: Landmark Centers

Sample Similarity Matrix

1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1	1

Randomly Selected "Centers"

1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	1	1	1	0	0	0
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	1	1	1

Subset of Similarity Matrix

1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Efficiency Improvements: SVD Eigen-Estimation

- Normalized Similarity $= D^{-1/2} W D^{-1/2}$

$$= D^{-1/2}\Big(XX^{\mathsf{T}} - I_n\Big)D^{-1/2}$$

$$= \Big(D^{-1/2}X\Big)\Big(D^{-1/2}X\Big)^{\mathsf{T}} - D^{-1}$$

where $D = \text{RowSums}(W) = XX^{\mathsf{T}} \cdot \vec{1} = X\Big(X^{\mathsf{T}} \cdot \vec{1}\Big)$

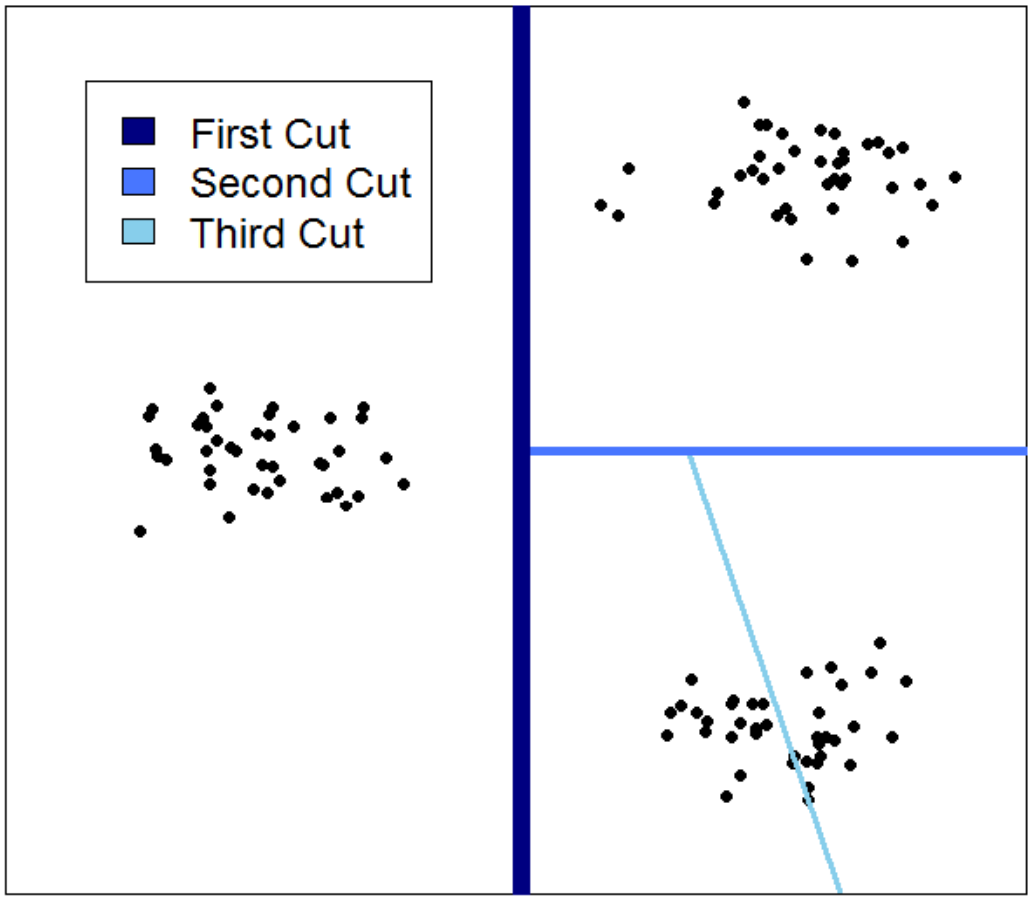
- $D \approx d \cdot I_n$:

$$\text{Eigenvectors}\left(D^{-1/2}XX^{\mathsf{T}}D^{-1/2} - D^{-1}\right) \Longleftrightarrow \text{S.V.D.}\Big(D^{-1/2}X\Big)$$

Adaptive Cluster Selection

- Divisive k-means

- Iterative Two-way Normalized Cuts



Feature Clustering

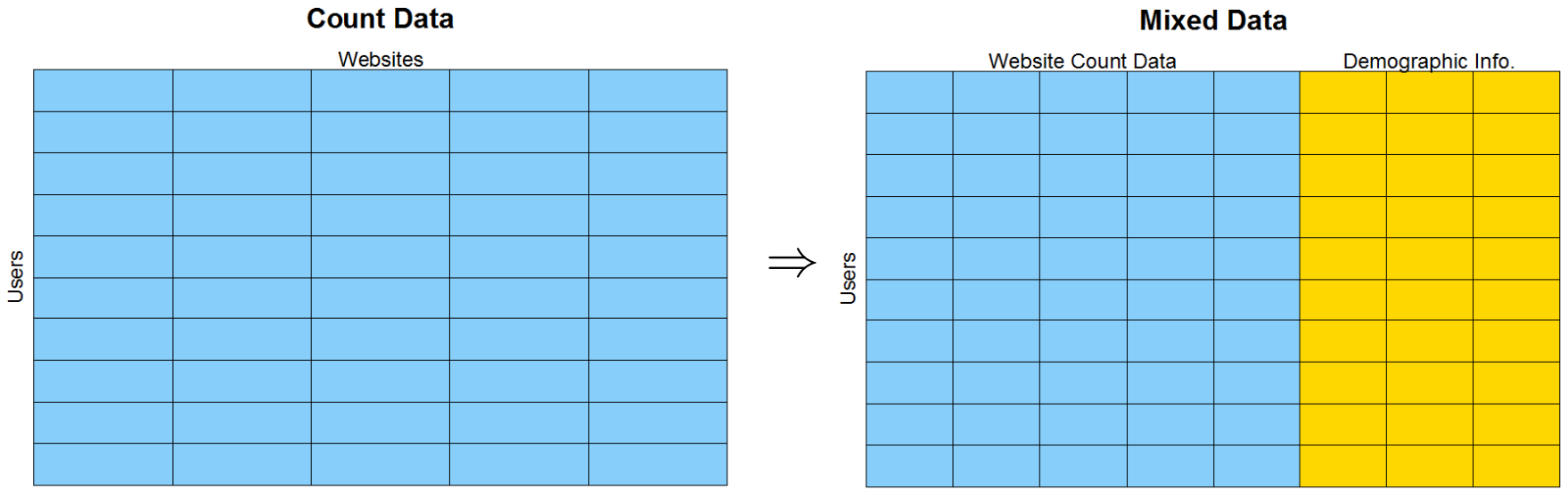
Plain Data Matrix

Documents \ Words									
	9	8	9	0	0	0	0	0	0
	8	5	8	0	0	0	0	0	0
	0	0	0	6	10	9	0	0	0
	0	0	0	7	8	6	0	0	0
	0	0	0	0	0	0	10	8	8
	0	0	0	0	0	0	7	6	10

Reduced Data Matrix

Documents	Word Group 1	Word Group 2	Word Group 3
	26	0	0
	21	0	0
	0	25	0
	0	21	0
	0	0	26
	0	0	23

Categorical Data



References

- Twenty Newsgroups Data Set, UCI Machine Learning Repository, http://qwone.com/ jason/20Newsgroups/
- Aggarwal, Charu C., and ChengXiang Zhai. "A survey of text clustering algorithms." Mining text data. Springer US, 2012. 77-128.
- Cai, Deng, Xiaofei He, and Jiawei Han. "Document clustering using locality preserving indexing." IEEE Transactions on Knowledge and Data Engineering 17.12 (2005): 1624-1637.
- Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation." IEEE Transactions on pattern analysis and machine intelligence 22.8 (2000): 888-905.
- Ng, Andrew Y., Michael I. Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm." NIPS. Vol. 14. No. 2. 2001.
- Zelnik-Manor, Lihi, and Pietro Perona. "Self-tuning spectral clustering." NIPS. Vol. 17. No. 1601-1608. 2004.
- Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.
- Nadler, Boaz, et al. "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators." NIPS. 2005.
- Coifman, Ronald R., and Stephane Lafon. "Diffusion maps." Applied and computational harmonic analysis 21.1 (2006): 5-30.
- Coifman, Ronald R., et al. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps." Proceedings of the National Academy of Sciences of the United States of America 102.21 (2005): 7426-7431.

Acknowledgements

- We specially thank Prof. Chen and Prof. Simic for the opportunity and guidance.

- Thank Verizon for sponsoring this project, and thank Irina Pragin, Yong Liu, Santanu Das, and Debasish Das for the cooperation and help.

Questions?