Document Clustering and Social Networks

Edward J. Wegman
George Mason University, College of Science

Classification Society Annual Meeting
Washington University School of Medicine
June 12, 2009

Outline

- Overview of Text Mining
- Vector Space Text Models
  - Latent Semantic Indexing
- Social Networks
  - Graph and Matrix Duality
  - Two Mode Networks
  - Block Models and Clustering
- Document Clustering with Mixture Models
- Conclusions and Acknowledgements

Text Mining

- Synthesis of …
  - – Information Retrieval
    - Focuses on retrieving documents from a fixed database
    - Bag-of-words methods
    - May be multimedia including text, images, video, audio
  - – Natural Language Processing
    - Usually more challenging questions
    - Vector space models
    - Linguistics: morphology, syntax, semantics, lexicon
  - – Statistical Data Mining
    - Pattern recognition, classification, clustering

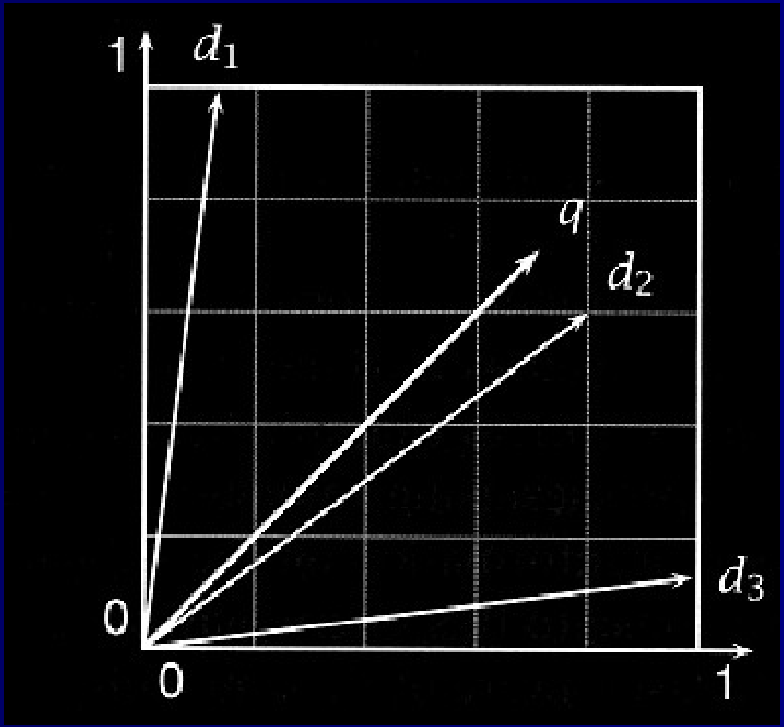
Text Mining Tasks

- Text Classification
  - Assigning a document to one of several pre-specified classes
- Text Clustering
  - Unsupervised learning – discovering cluster structure
- Text Summarization
  - Extracting a summary for a document
  - Based on syntax and semantics
- Author Identification/Determination
  - Based on stylistics, syntax, and semantics
- Automatic Translation
  - Based on morphology, syntax, semantics, and lexicon
- Cross Corpus Discovery
  - Also known as Literature Based Discovery

Text Preprocessing

- Denoising
  - Means removing stopper words … words with little semantic meaning such as *the, an, and, of, by, that* and so on.
  - Stopper words may be context dependent, e.g. *Theorem* and *Proof* in a mathematics document
- Stemming
  - Means removal suffixes, prefixes and infixes to root
  - An example: *wake, waking, awake, woke* → *wake*

Vector Space Model



- Documents and queries are represented in a high-dimensional vector space in which each dimension in the space corresponds to a word (term) in the corpus (document collection).
- The entities represented in the figure are *q* for query and *d_1*, *d_2*, and *d_3* for the three documents.
- The term weights are derived from occurrence counts.

Vector Space Methods

- The classic structure in vector space text mining methods is a term-document matrix where
  - Rows correspond to terms, columns correspond to documents, and
  - Entries may be binary or frequency counts.
- A simple and obvious generalization is a bigram (multigram)-document matrix where
  - Rows correspond to bigrams, columns to documents, and again entries are either binary or frequency counts.

Vector Space Methods

- *Latent Semantic Indexing (LSI)* is a technique that projects queries and documents into a space with *latent* semantic dimensions.

- Co-occuring terms are projected into the same semantic dimensions and non-co-occuring terms onto different dimensions.

- In latent semantic space, a query and a document can have high cosine similarity even if they do not share any terms as long as their terms are semantically similar according to the co-occurence analysis.

Latent Semantic Indexing

- LSI is the application of Singular Value Decomposition (SVD) to the term-document matrix.

- SVD takes a matrix W and represents it as W in a lower dimensional space such that the two-norm is minimized, i.e. 1 . 1 □ .

- The SVD projects an • -dimensional space onto a • -dimensional space where The

Latent Semantic Indexing

- In our application to word-document matrices, • is the number of word types (terms) in the corpus (document collection).

- Typically • is chosen between 100 to 150.

- The SVD projection is computed by decomposing the term-document matrix t $_1$˜ into the product of three matrices

t $_{tt}$˜ h r e e ˜ m†

where – – ˜– . – – – – –

Latent Semantic Indexing

- These matrices have *orthonormal* columns. This means the column vectors are of unit length and are orthogonal to each other. In particular

a†r e (the identity matrix) o r †t h

- The diagonal matrix • contains the *singular values* of o in descending order. The f$^{\sigma}$ singular values indicates the amount of variation along the i^{in} axis.

- By restricting the matrices • By and to the first
- B columns, we obtain y r eand st ~ r†

with

Wi t$_{www}$h$_{www}$um n†s

LSI - Some Basic Relations

- □†□ □ † † † † † †
  † †††† † † †† □†T□† † . †

- • ¡ † † o † † † † †
  ††††††† † o . †

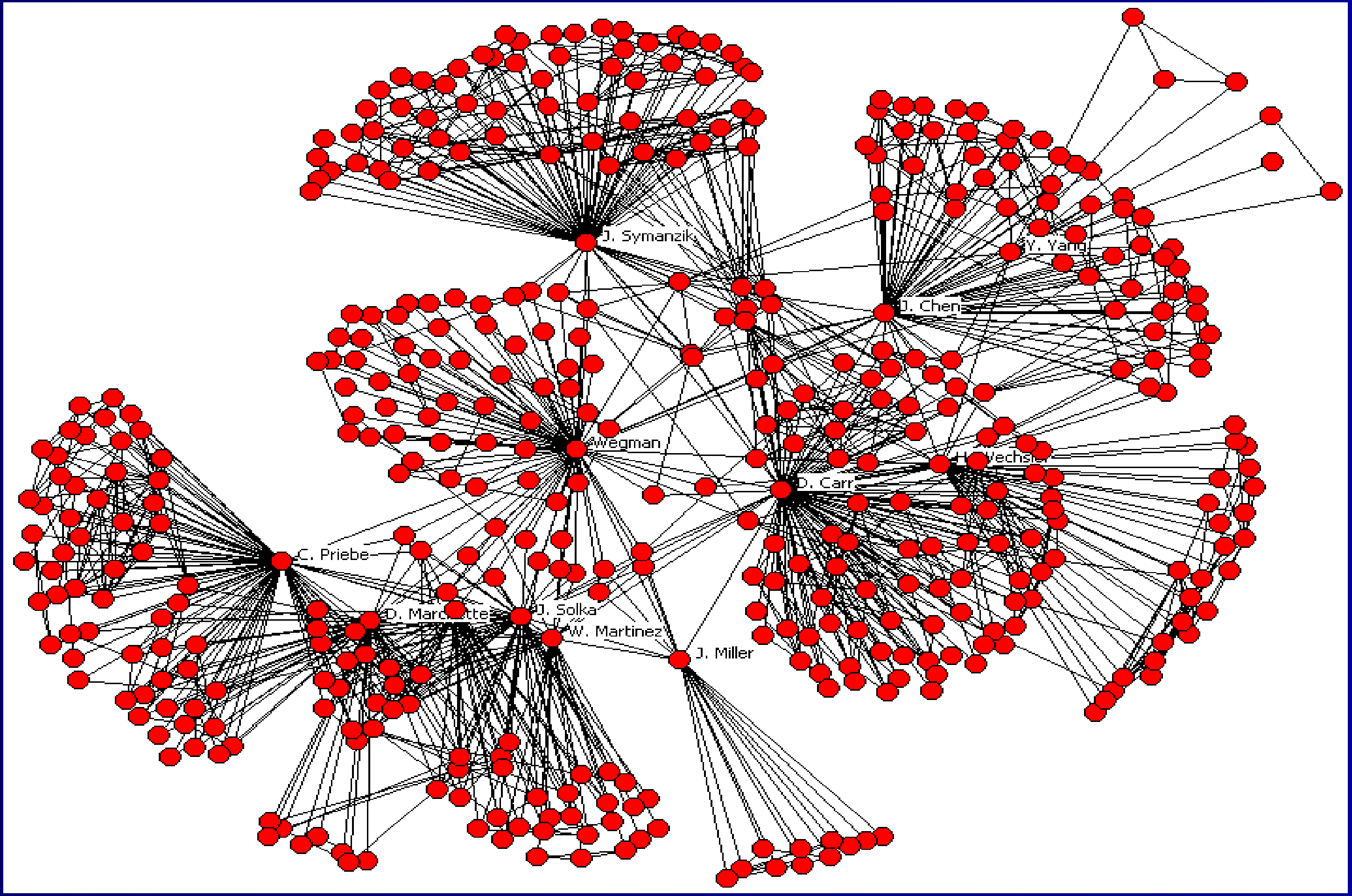
-

- † ††† † † †o † † † †

Social Networks

- Social networks can be represented as graphs
  - A graph G(V, E), is a set of vertices, V, and edges, E
  - The social network depicts actors (in classic social networks, these are humans) and their connections or ties
  - Actors are represented by vertices, ties between actors by edges
- There is one-to-one correspondence between graphs and so-called adjacency matrices
- Example: Author-Coauthor Networks

Graphs versus Matrices



Two-Mode Networks

- When there are two types of actors
  - Individuals and Institutions
  - Alcohol Outlets and Zip Codes
  - Paleoclimate Proxies and Papers
  - Authors and Documents
  - Words and Documents
  - Bigrams and Documents
- SNA refers to these as two-mode networks, graph theory as bi-partite graphs
  - Can convert from two-mode to one-mode

Two-Mode Computation

Consider a bipartite ***individual by institution*** social network. Let $A_{m\times n}$ be the individual by institution adjacency matrix with $m =$ the number of individuals and $n =$ the number of institutions. Then

$$C_{m\times m} = A_{m\times n}A^{T}{}_{n\times m} =$$

Individual-Individual social network adjacency matrix with $c_{ii} = \sum_j a_{ij} =$ the strength of ties to all individuals in i's social network and $c_{ij} =$ the tie strength between individual i and individual j.

Two-Mode Computation

Similarly,

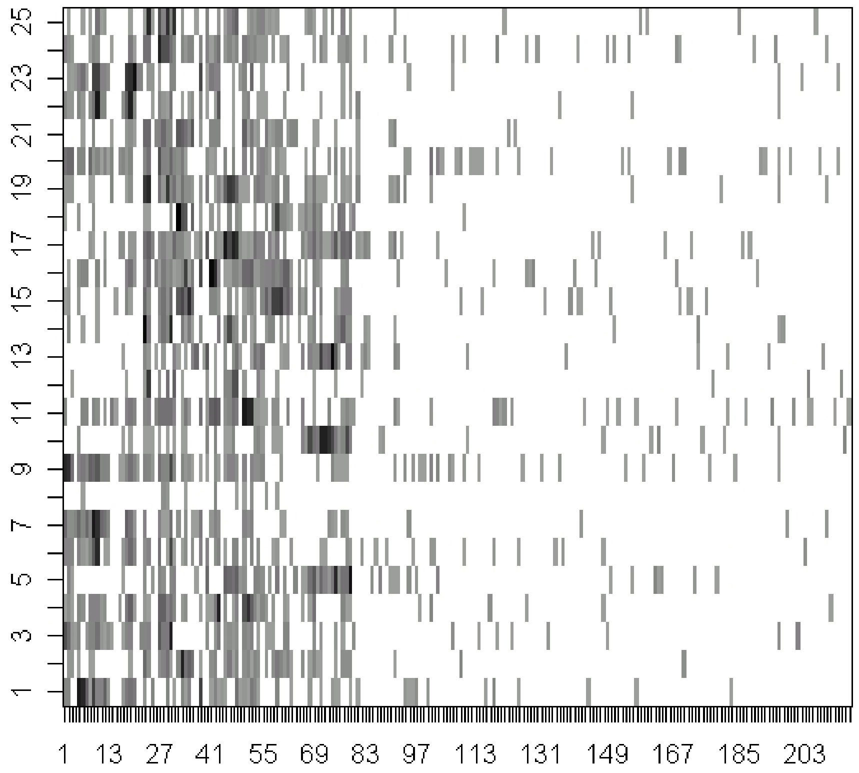
$$P_{n \times n} = A^{T}{}_{n \times m} A_{m \times n} =$$

Institution by Institution social network adjacency matrix with $p_{jj} = \sum_i a_{ij} =$ strength of ties to all institutions in i's social network with p_{ij} the tie strength between institution i and institution j.

Two-Mode Computation

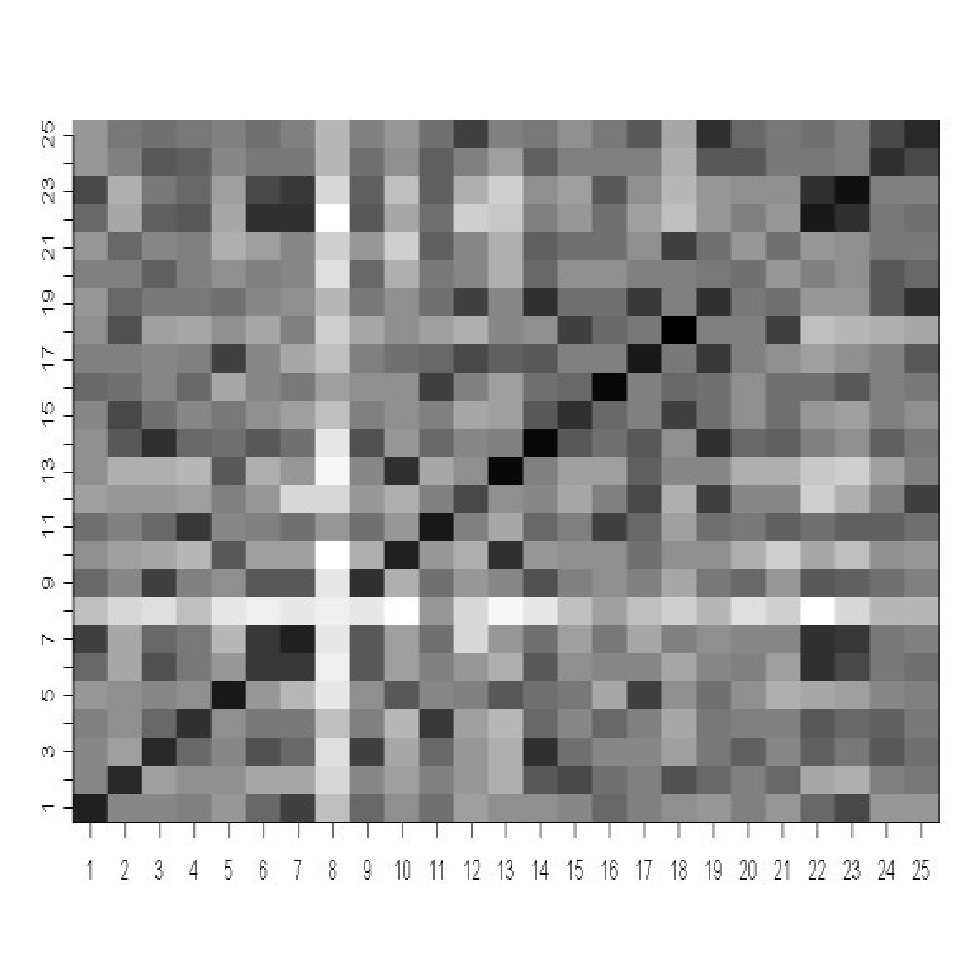
- Of course, this exactly resembles the computation for LSI.
- Viewed as a two-mode social network, this computation allows us:
  - to calculate strength of ties between terms relative to this document database (corpus)
  - And also to calculate strength of ties between documents relative to this lexicon
- If we can cluster these terms and these documents, we can discover:
  - similar sets of documents with respect to this lexicon
  - sets of words that are used the same way in this corpus

Example of a Two-Mode Network



Our *A* matrix

Example of a Two-Mode Network

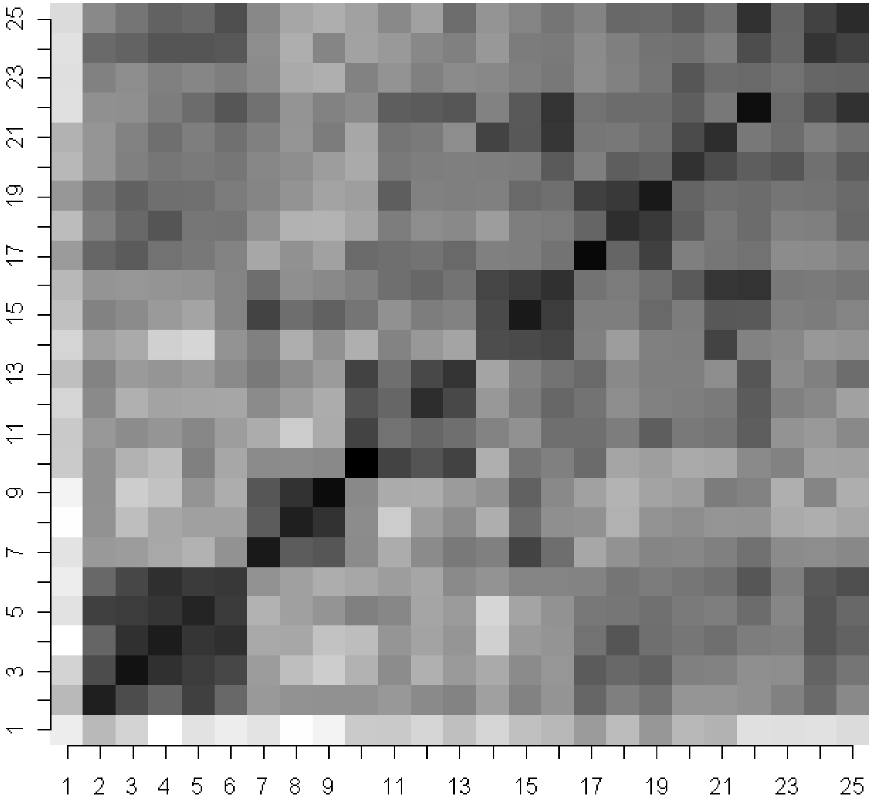


Our *P* matrix

Block Models

- A **partition of a network** is a clustering of the vertices in the network so that each vertex is assigned to exactly one class or cluster.
- Partitions may specify some property that depends on attributes of the vertices.
- Partitions divide the vertices of a network into a number of mutually exclusive subsets.
  - That is, a partition splits a network into parts.
- Partitions are also sometimes called **blocks or block models**.
  - These are essentially a way to cluster actors together in groups that behave in a similar way.

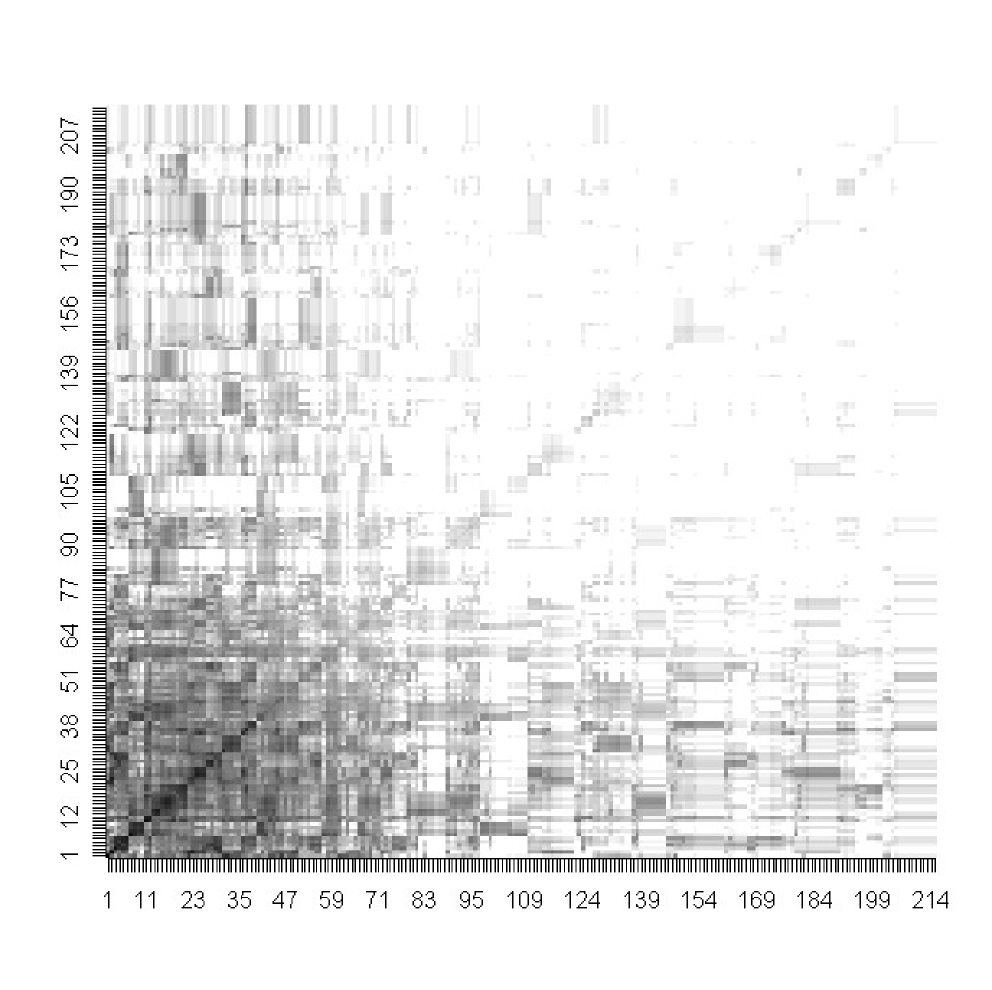
Example of a Two-Mode Network



Block Model

***P* Matrix - Clustered**

Example of a Two-Mode Network



Block Model Matrix – Our *C* Matrix Clustered

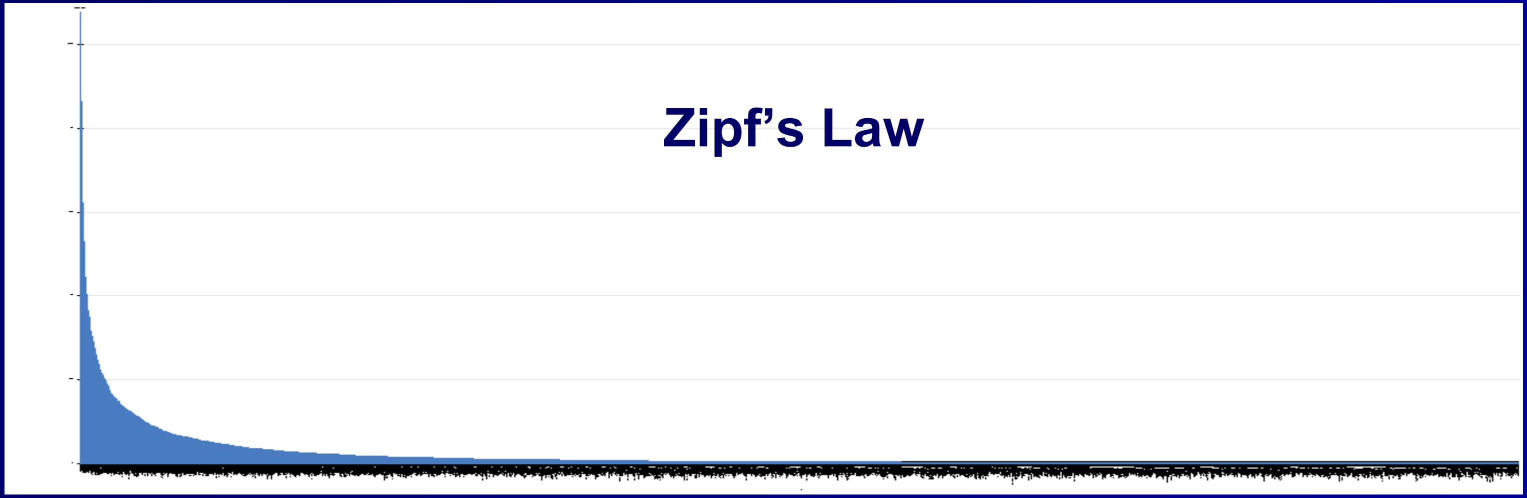
Example Data

- The text data were collected by the Linguistic Data Consortium in 1997 and were originally used in Martinez (2002)
  - The data consisted of 15,863 news reports collected from Reuters and CNN from July 1, 1994 to June 30, 1995
  - The full lexicon for the text database included 68,354 distinct words
    - In all 313 stopper words are removed
    - after denoising and stemming, there remain 45,021 words in the lexicon
  - In the examples that I report here, there are 503 documents only

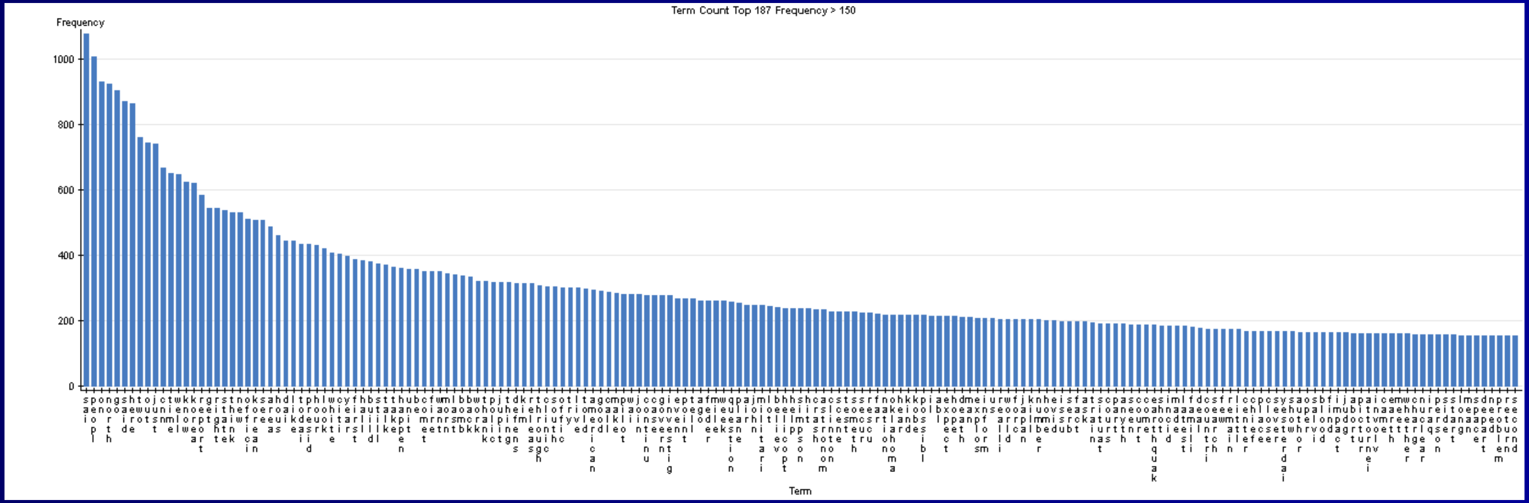
Example Data

- A simple 503 document corpus we have worked with has 7,143 denoised and stemmed entries in its lexicon and 91,709 bigrams.
  - Thus the TDM is 7,143 by 503 and the BDM is 91,709 by 503.
  - The term vector is 7,143 dimensional and the bigram vector is 91,709 dimensional.
  - The BPM for each document is 91,709 by 91,709 and, of course, very sparse.
- A corpus can easily reach 20,000 documents or more.

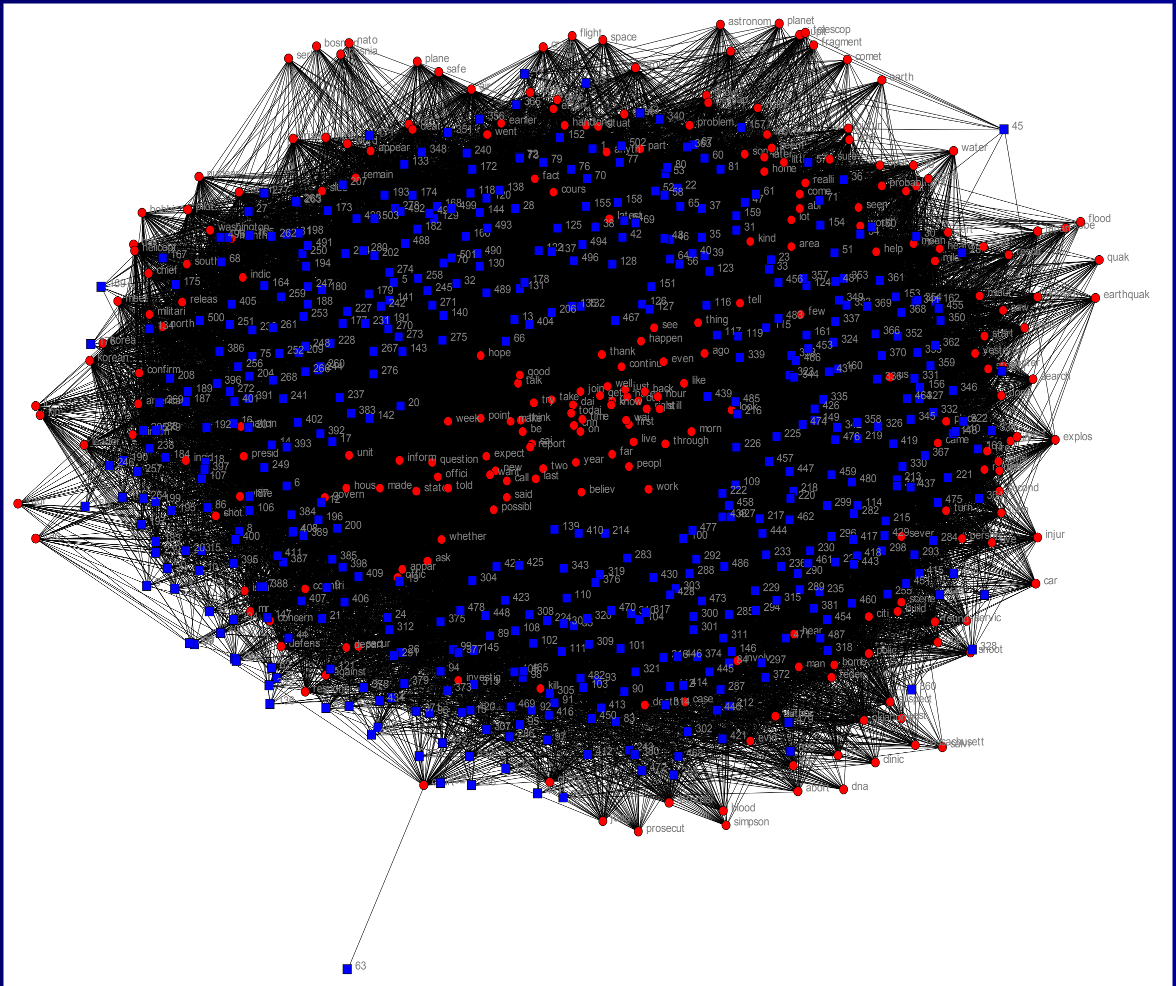
Term-Document Matrix Analysis



Zipf's Law



Term-Document Matrix Analysis



Mixture Models for Clustering

- Mixture models fit a mixture of (normal) distributions
- We can use the means as centroids of clusters
- Assign observations to the “closest” centroid
- Possible improvement in computational complexity

Our Proposed Algorithm

- Choose the number of desired clusters.
- Using a normal mixtures model, calculate the mean vector for each of the document proto-clusters.
- Assign each document (vector) to a proto-cluster anchored by the closest mean vector.
    - This is a Voronoi tessellation of the 7143-dimensional term vector space. The Voronoi tiles correspond to topics for the documents.
- Or assign documents based on maximum posterior probability.

Normal Mixtures

~ . . □, . . . , □ □

where w 1 h er e 1 i s1 t is taken as the multivariate normal density, 1 1 are the mixing coefficients, is the number of mixing terms, and . ~ ~ ~ v v is the mean vector and covariance matrix. The sample size we denote by m, in our case m . . . . The dimension, ~ , of the vector is . . . . a

EM Algorithm for Normal Mixtures

Mi 1 1 □ 1 □w̃i x ~ ; 1 x i 1 □ / 1

1 1 □ĩ x~

□

~ ~ ~ :1 / 1 ; : : :õ □ : :~ o ~□ †□ / : 1 :

†† is the estimated posterior probability that belongs to component , 1 1 is the estimated mixing coefficient, ~ ~ and are the estimated mean and covariance matrix respectively.

Notation

- 1 . . . . ; the number of documents.

- • •. the desired number of clusters

- ˜ . . . . the dimension of the term vector the size of the lexicon for this corpus

Considerations about the Normal Density

Because the dimensionality of the term vectors is so large, there are some considerations about the EM algorithm to be made. Recall

aa□ $\frac{\tilde{}}{□1\ .\bar{}\ □.\bar{}}$. – v ˜ ˜ . ˜ †

† tends to be singular, certainly ill-conditioned. In our experience just used as a raw estimate roundoff error causes e to have a zero determinant. Morover, . 1 $\tilde{}$.¯ also rounds to zero.

Revised EM Algorithm

In order to regularize the computation, we take I $_{l}$ I n, the identity matrix. Then the EM algorithm becomes

$$b_{\text{œ}}\, b \; \frac{1^{\sim}_{1} \quad \cdot - . \; . \quad \sim \; {}_{e}\, e \; \sim \quad \dagger}{. \quad 1^{\sim}_{.} \quad \cdot - . \; . \quad \sim \; {}_{e}\, e \; \sim \quad \dagger}$$

$$11.$$

$$\cdot \quad . \quad . \; . \; :$$

$$1 \; {}_{1} \; 1 \quad \frac{\cdot\, \cdot \;\cdot}{1}$$

$$\cdot \quad . \quad . \; . \; :1$$

$$\sim \; {}_{\sim} \sim \quad \frac{\cdot\, \cdot \;\cdot \sim}{1_{.}} \, :$$

And of course we no longer estimate **A**. We are really only interested in estimating the means.

Comuptational Complexity

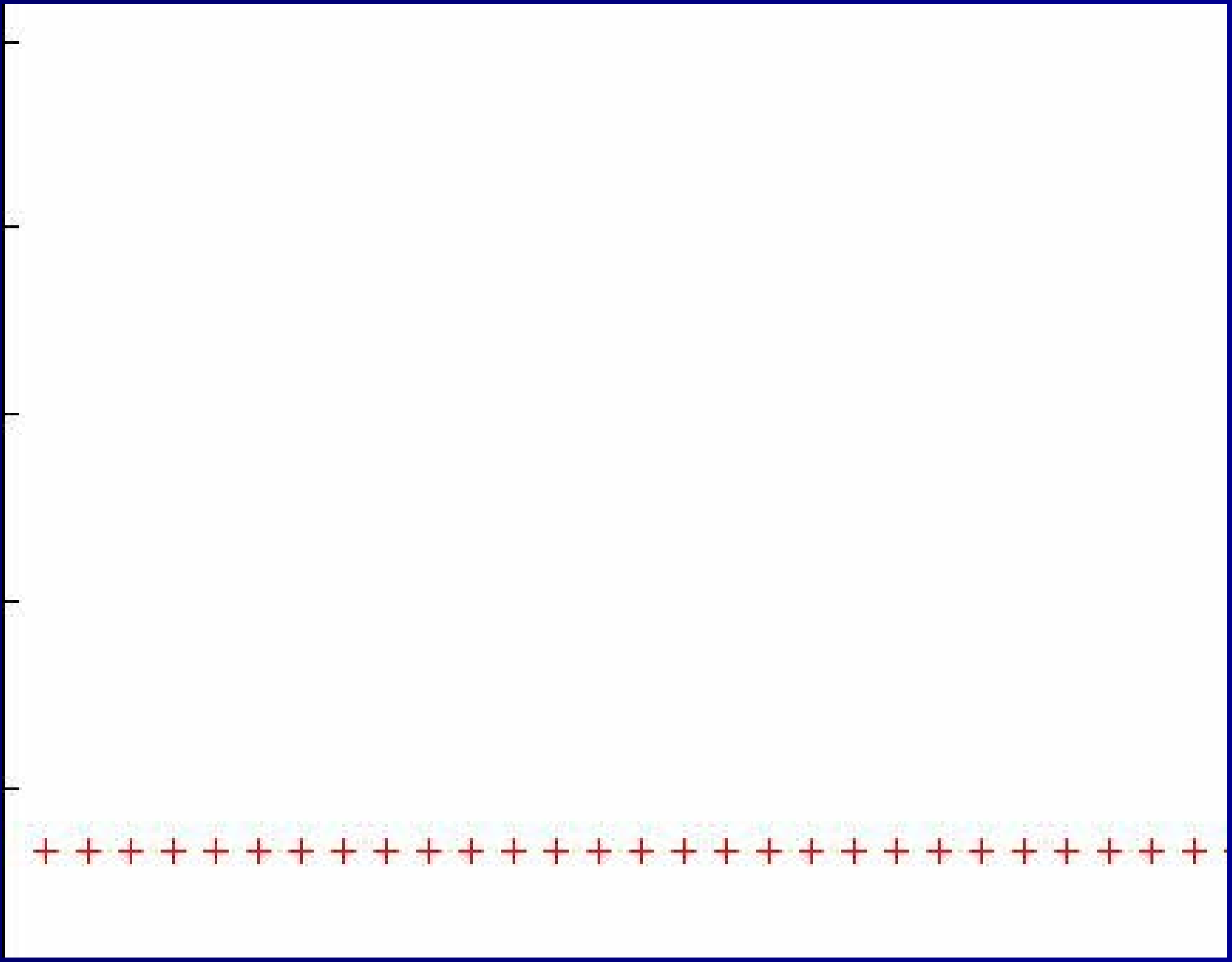
The computation of TTh has complexity ˜ ,
the computation of 1 1 has complexity 1 he c
and the computation of ˜ ˜ has complexity
˜ ˜˜ The EM algorithm is a recursive
algorithm. The number of recursions can be
determined by a stopping algorithm or fixed by
the user. In either case, if the number of
recursions is r , then the overall complexity of
the EM phase ist t˜ . It is linear in all the
key size variables.

The Voronoi computation is T he Vo

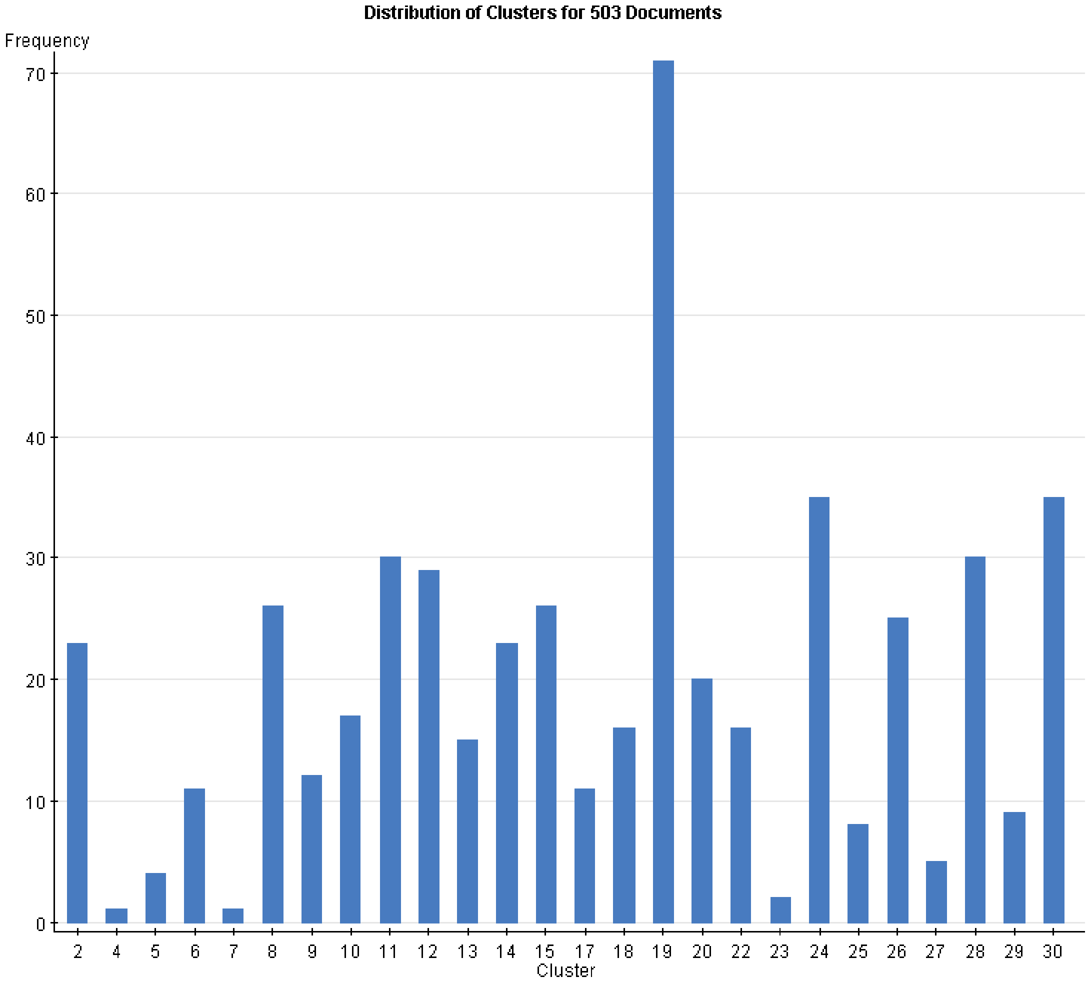
Results

In the present data set, ˜ .. . . ,
. , and Time in seconds from loading file to membership computation is . . . seconds. This computation was done on an Intel Centrino Dual Core processor running at 1.6 gigahertz.

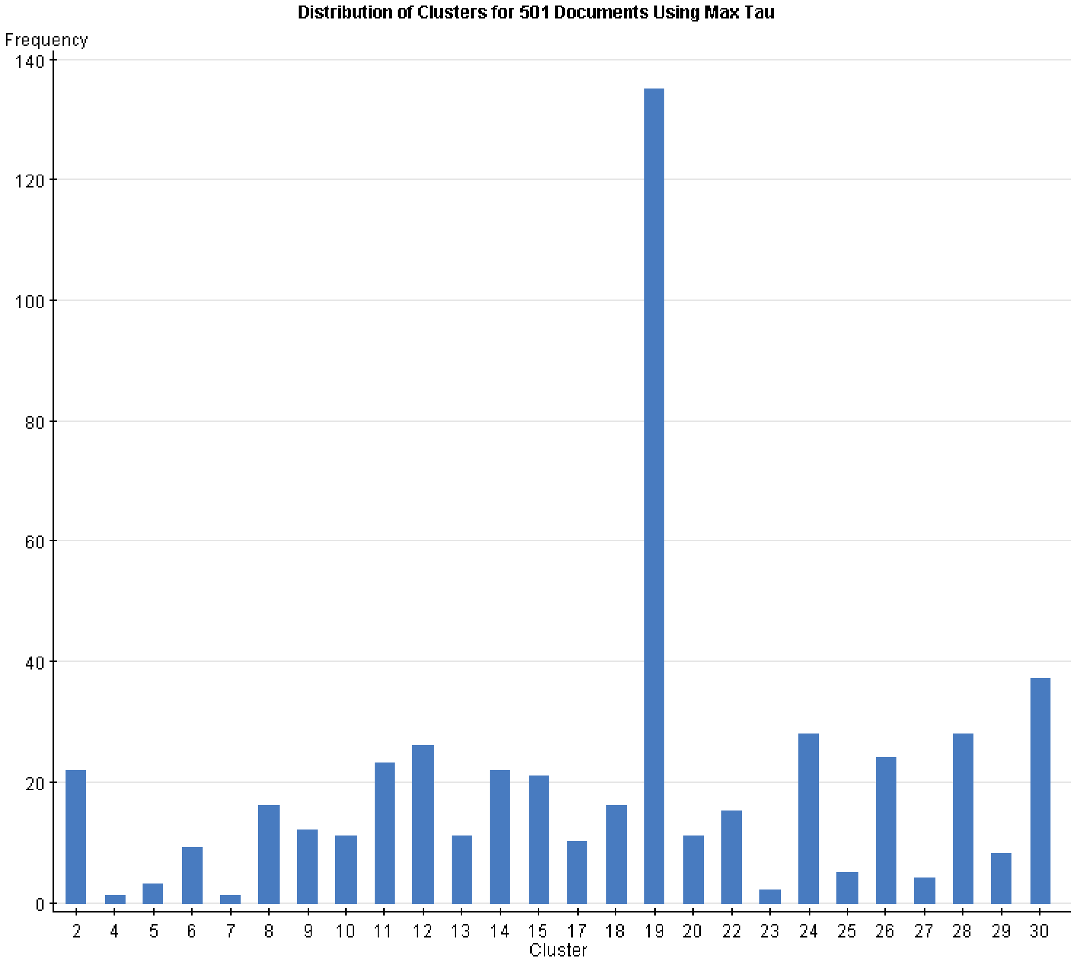
πWeights



Cluster Size Distribution (Based on Voronoi Tessellation)

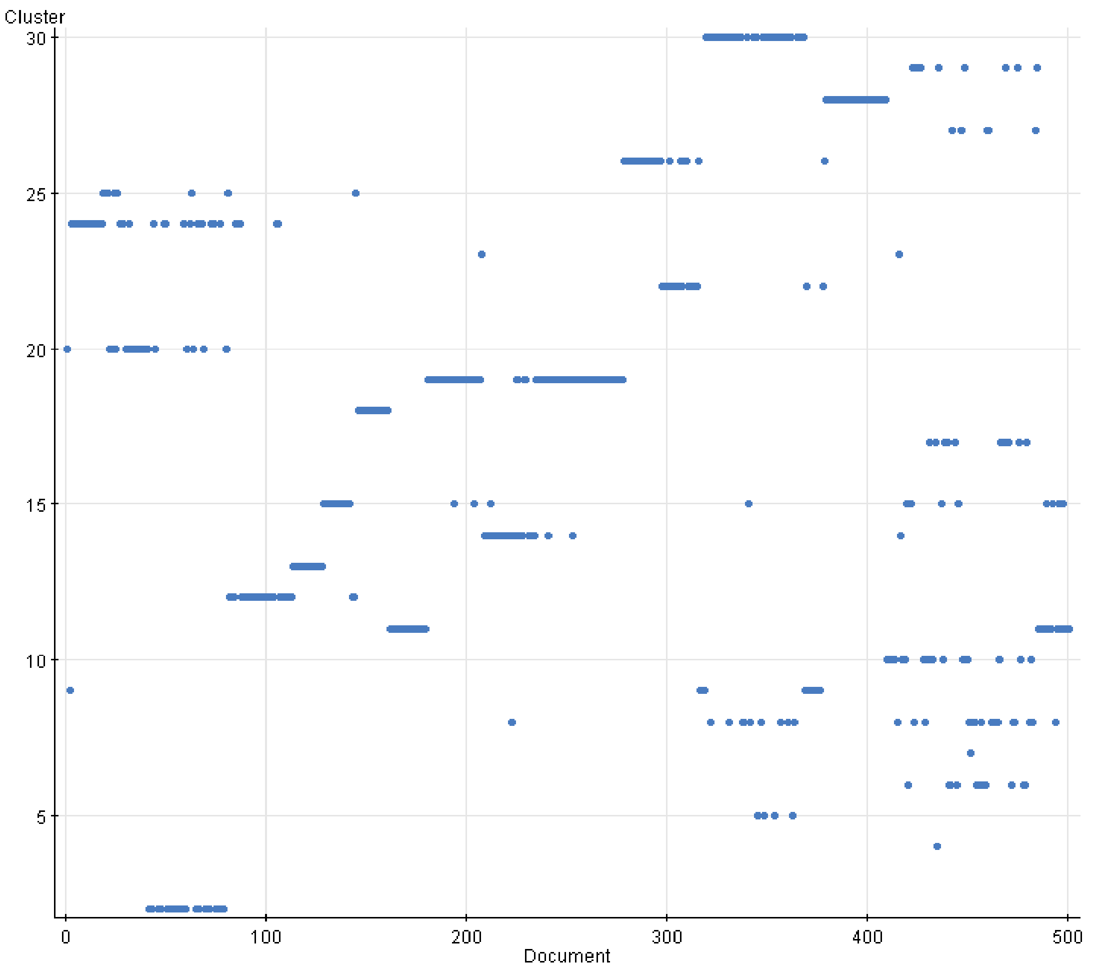


Cluster Size Distribution (Based on Maximum Estimated Posterior Probability, τ_{ij})



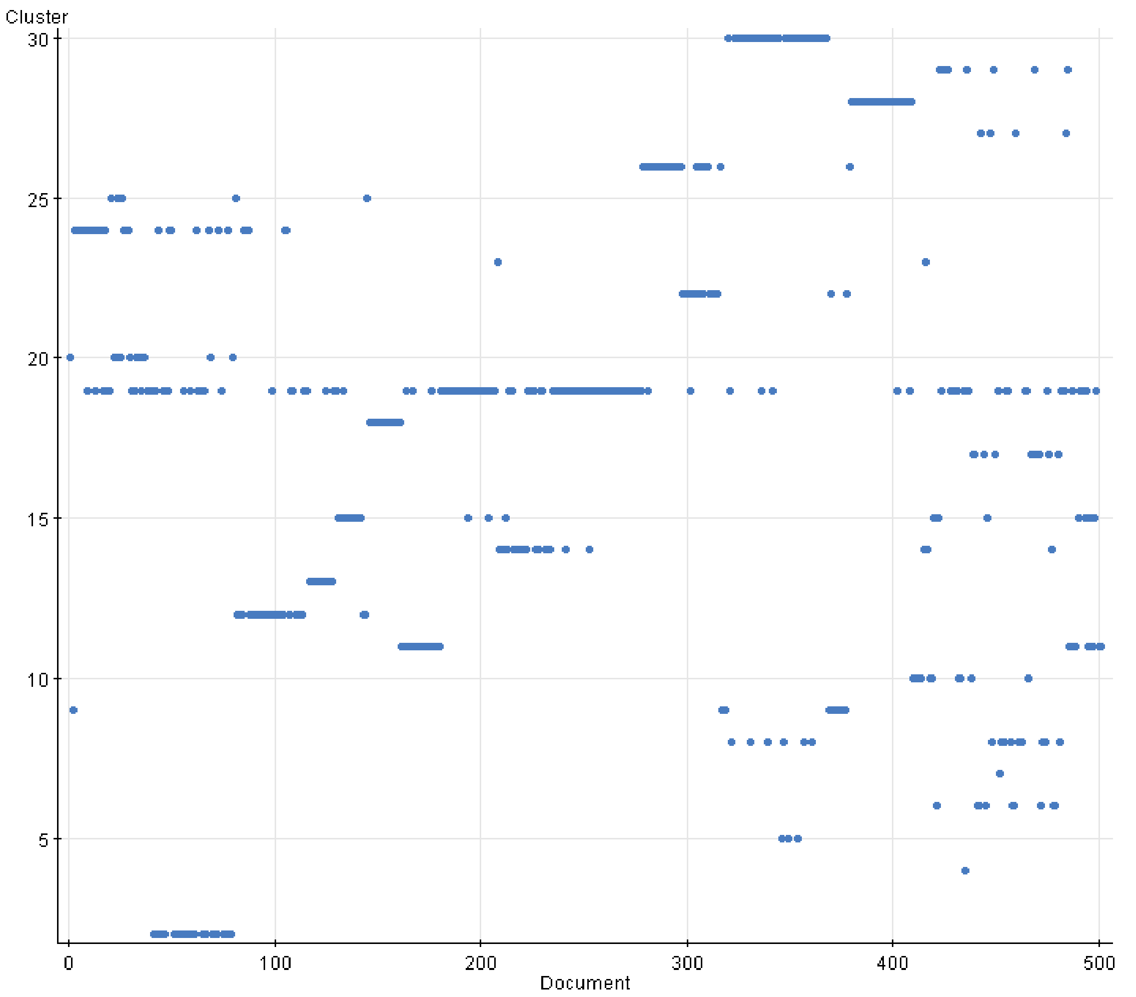
Document by Cluster Plot

(Voronoi)



Document by Cluster Plot

(Maximum Posterior Probability)



Cluster Identities

- Cluster 02: Comet Shoemaker Levy Crashing into Jupiter.
- Cluster 08: Oklahoma City Bombing.
- Cluster 11: Bosnian-Serb Conflict.
- Cluster 12: Court-Law, O.J. Simpson Case.
- Cluster 15: Cessna Plane Crashed onto South Lawn White House.
- Cluster 19: American Army Helicopter Emergency Landing in North Korea.
- Cluster 24: Death of North Korean Leader (Kim il Sung) and North Korea's Nuclear Ambitions.
- Cluster 26: Shootings at Abortion Clinics in Boston.
- Cluster 28: Two Americans Detained in Iraq.
- Cluster 30: Earthquake that Hit Japan.

Acknowledgments

- This is joint work with Dr. Yasmin Said and Dr. Walid Sharabati.
- Dr. Angel Martinez
- Army Research Office (Contract W911NF-04-1-0447)
- Army Research Laboratory (Contract W911NF-07-1-0059)
- National Institute On Alcohol Abuse And Alcoholism (Grant Number F32AA015876)
- Isaac Newton Institute
- Patent Pending

Contact Information

Edward J. Wegman
Department of Computational and Data Sciences
MS 6A2, George Mason University
4400 University Drive
Fairfax, VA 22030-4444 USA

Email: ewegman@gmail.com

Phone: (703) 993-1691