SAND2009-5518C

Persistent homology for parameter sensitivity in large-scale text-analysis (informatics) graphs

Danny Dunlavy
Computer Science and Informatics Department (1415)
Sandia National Laboratories

August 29, 2009

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

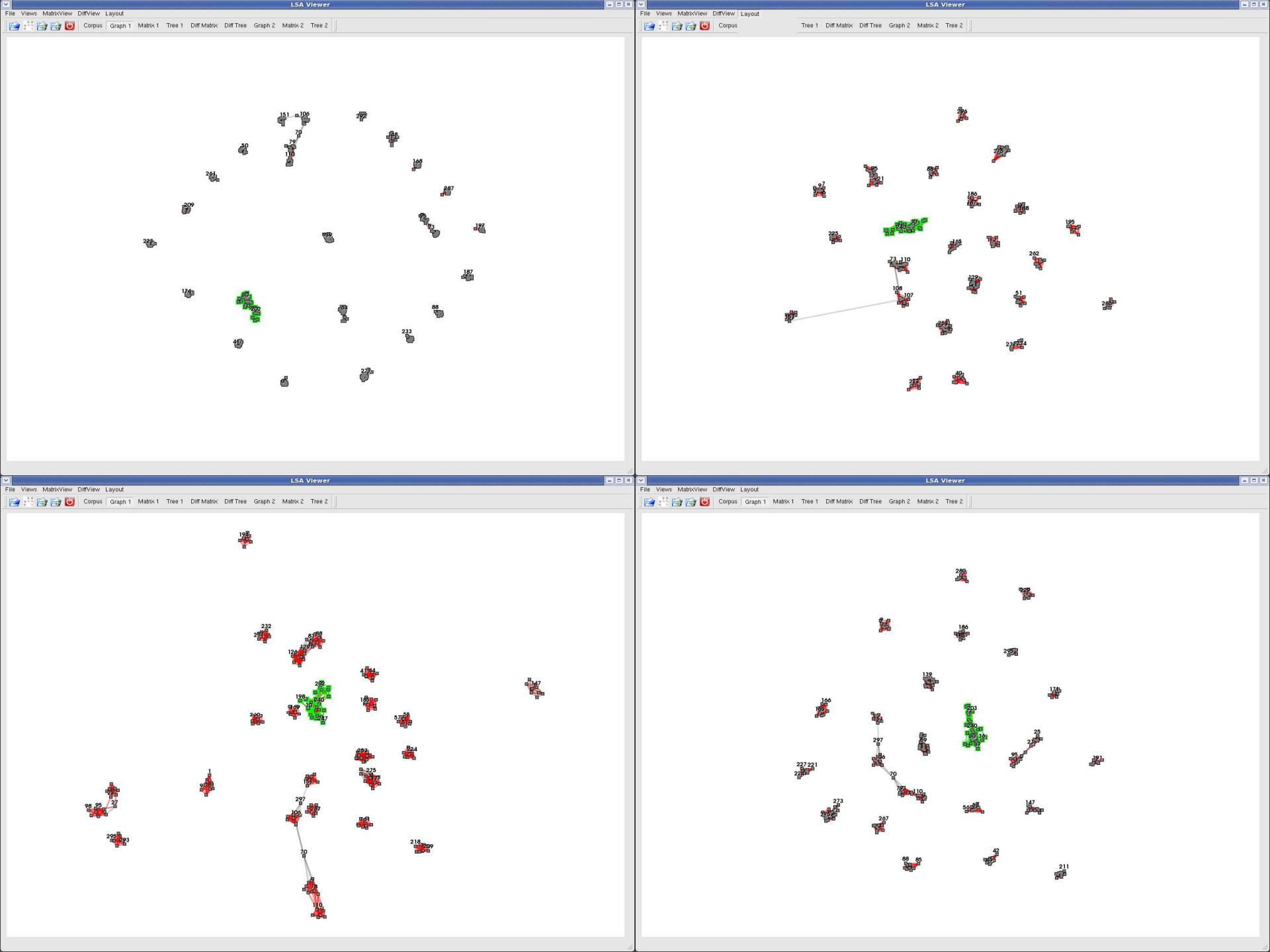
Sandia National Laboratories

Acknolwedgement

- **Text Analysis**
  - Pat Crossno, Tim Shead, Tammy Kolda, Philip Kegelmeyer, Brett Bader, Sean Gilpin, Tad Turpen

- **Computational Topology**
  - David Day, Scott Mitchell, Shawn Martin

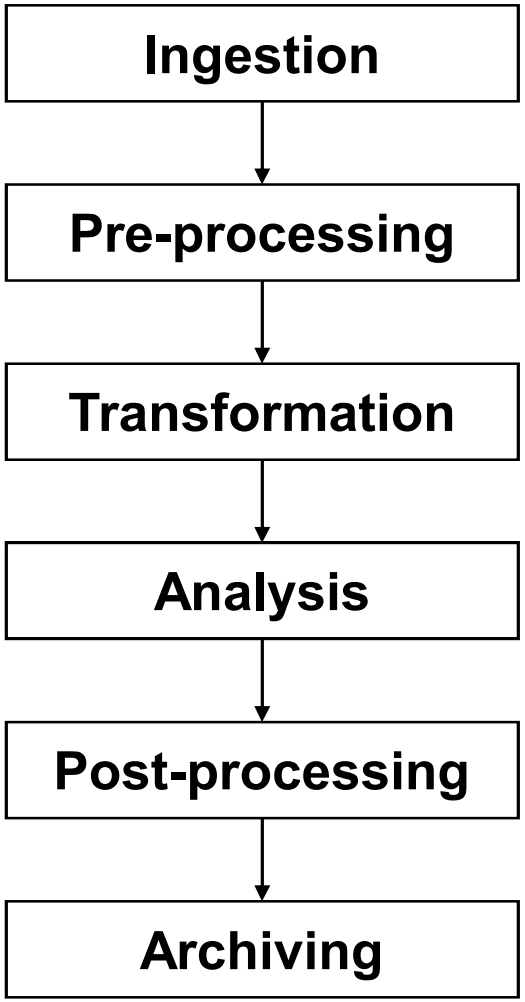
Sandia National Laboratories


LSA Viewer
File Views MatrixView DiffView Layout
Corpus Graph 1 Matrix 1 Tree 1 Diff Matrix Diff Tree Graph 2 Matrix 2 Tree 2
LSA Viewer
File Views MatrixView DiffView Layout
Corpus Tree 1 Diff Matrix Diff Tree Graph 2 Matrix 2 Tree 2
LSA Viewer
File Views MatrixView DiffView Layout
Corpus Graph 1 Matrix 1 Tree 1 Diff Matrix Diff Tree Graph 2 Matrix 2 Tree 2
LSA Viewer
File Views MatrixView DiffView Layout
Corpus Graph 1 Matrix 1 Tree 1 Diff Matrix Diff Tree Graph 2 Matrix 2 Tree 2

Main Goals for Text Analysis Research

- **Relationship discovery and understanding**
  - Document-document, term-term, term-document
  - Data clustering, classification, summarization
- **Understanding of sensitivities**
  - Statistical significance, hypothesis testing
  - Visual analysis
  - Surrogate data generation and model verification
  - Persistent homology
- **Incorporation of analyst knowledge**
  - Annotation and relevance feedback
  - Metric learning, priors
- **Applications**
  - Nuclear nonproliferation, intelligence analysis, technology assessment, sentiment analysis, cyber security

Text Analysis Pipeline



File readers (ASCII, UTF-8, XML, PDF, ...)

Tokenization, stemming, part-of-speech tagging
named entity extraction, sentence boundaries

Data model, dimensionality reduction,
feature weighting, feature extraction/selection

Information retrieval, clustering, summarization,
classification, pattern recognition, statistics

Visualization, filtering, summary statistics

Database, file, web site

Machine Learning and Text Analysis

- **User feedback**
  - Learn how users interact with analysis capabilities
  - Leverage annotation in future analyses
  - Encode knowledge and perspective

- **Example: labeling data as "relevant" / "not relevant"**
  - Posed as a classification problem
    - **Goal:** data instance ⇒ label (class, category)
    - **Method:** supervised learning ⇒ classification models
    - **Ensemble:** combined set of classification models
      - E.g., bagging, boosting, random forest

Vector Space Model

- **Vector Space Model for Text**
  - Terms (features): $\mathbf{t} \in \mathbb{R}^m$
  - Documents (objects): $\mathbf{d} \in \mathbb{R}^n$
  - Term-Document Matrix: A
  - a_{ij}: measure of importance of term i in document j

	d_1	$\cdots$	d_n
t_1	a_{11}	$\cdots$	a_{1n}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_m	a_{m1}	$\cdots$	a_{mn}

- **Term Examples**
  - *Sentence:* "Danny re-sent $1."
  - *Words:* danny, sent, re [# chars?], $ [sym?], 1 [#?], re-sent [-?]
  - *n-grams (n=3):* dan, ann, nny, ny_, _re, re-, e-s, sen, ent, nt_, …
  - *Named entities (people, orgs, money, etc.):* danny, $1
- **Document Examples**
  - Documents, paragraphs, sentences, fixed-size chunks

[G. Salton, A. Wong, and C. S. Yang (1975), *Comm. ACM*, 18(11), 613–620.]

Feature Weighting

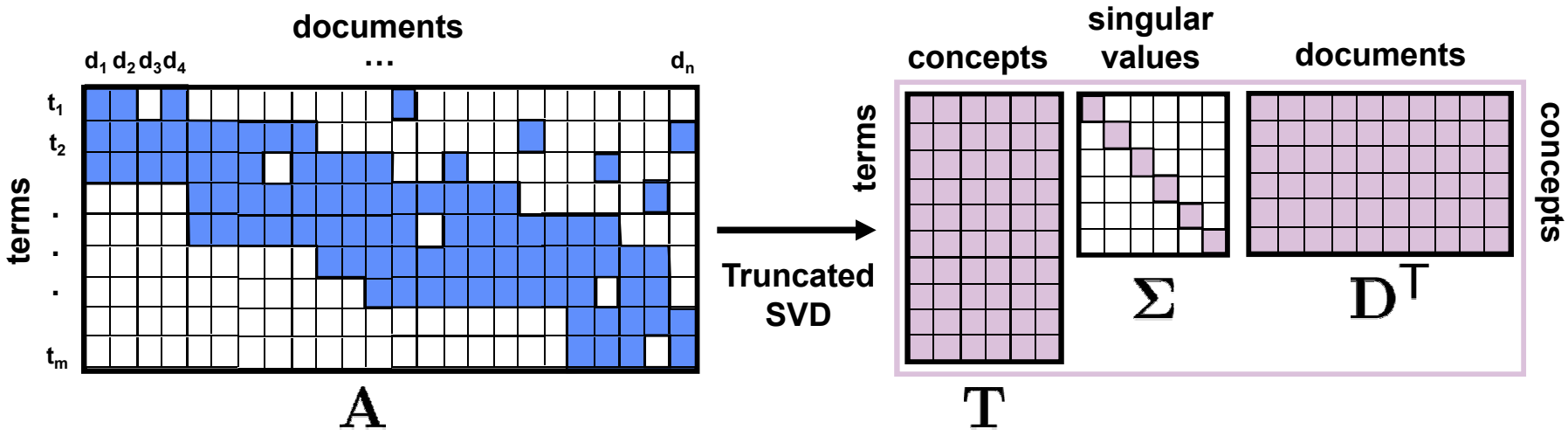
Term × Document Matrix Scaling: $a_{ij} = \tau_{ij} \cdot \gamma_i \cdot \delta_j$

Local Weights (τ_{ij})	
Term Frequency	f_{ij}
Binary	$\chi(f_{ij}) = \begin{cases} 0 & f_{ij} = 0 \\ 1 & f_{ij} > 0 \end{cases}$
Log	$\log(f_{ij} + 1)$
Global Weights (γ_i)	
None	1
Normalized	$\left(\sum_i f_{ij}^2\right)^{-1/2}$
Inverse Document Frequency (IDF)	$\log\left(n / \sum_j \chi(f_{ij})\right)$
IDF Squared(IDF2)	$\log\left(n / \sum_j (\chi(f_{ij}))^2\right)$
Entropy	$1 - \sum_j \frac{\left(f_{ij} / \sum_k f_{ik}\right) \log\left(f_{ij} / \sum_k f_{ik}\right)}{\log n}$
Normalization (δ_j)	
None	1
Normalized	$\left(\sum_i (\tau_{ij}\gamma_i)^2\right)^{-1/2}$

More about Features

- **Impact of data characteristics and extraction algorithms on features**
  - Natural language processing (NLP)
    - Stemming and lemmatization
    - Part-of speech tagging
    - Named entity extraction
    - Sentence boundary detection
  - Data imperfections
    - Encoding errors
    - Segmentation errors
    - Incomplete data

Latent Semantic Analysis (LSA)



- **SVD:** $\mathbf{A} = \mathbf{T}\Sigma\mathbf{D}^{\mathsf{T}}$
- **Truncated SVD:** $\mathbf{A} \approx \mathbf{A}_k = \mathbf{T}_k\Sigma_k\mathbf{D}_k^{\mathsf{T}} = \sum_{r=1}^{k} \sigma_r\, \mathbf{t}_r\mathbf{d}_r^T$
- **Query scores (query as new "doc"):** $q^{\mathsf{T}}\mathbf{A}$
- **LSA Ranking:** $q^{\mathsf{T}}\mathbf{A}_k$

[Deerwester, S. C., *et al.* (1990). *J. Am. Soc. Inform. Sci.* 41 (6), 391–407.]

LSA Example

d_1 : Hurricane. A hurricane is a catastrophe.

d_2 : An example of a catastrophe is a hurricane.

d_3 : An earthquake is bad.

d_4 : Earthquake. An earthquake is a catastrophe.

Remove stopwords

	q
hurricane	**1**
earthquake	**0**
catastrophe	**0**

normalization only

A	d_1	d_2	d_3	d_4
hurricane	**.89**	**.71**	**0**	**0**
earthquake	**0**	**0**	**1**	**.89**
catastrophe	**.45**	**.71**	**0**	**.45**
q^TA	**.89**	**.71**	**0**	**0**

rank-2 approximation

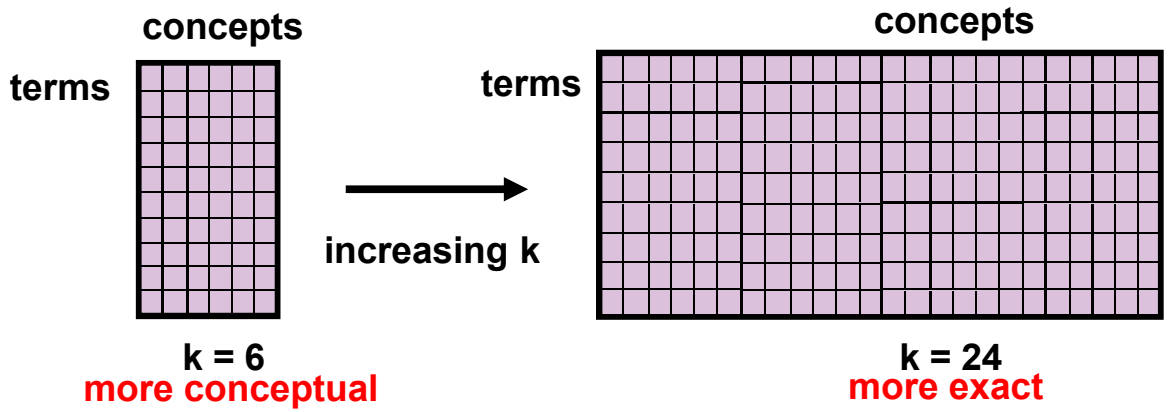
A_2	d_1	d_2	d_3	d_4
hurricane	**.78**	**.78**	**-.11**	**.11**
earthquake	**-.03**	**.02**	**.96**	**.92**
catastrophe	**.59**	**.60**	**.15**	**.30**
q^TA_2	**.78**	**.78**	**–**	**.11**

captures link to doc 4

Sandia National Laboratories

LSA: Rank Selection

- **Conceptual searching**
  - rank(k) ↑ : more exact data similarities
  - rank(k) ↓ : more conceptual data similarities
  - Compute larger rank, then use smaller rank



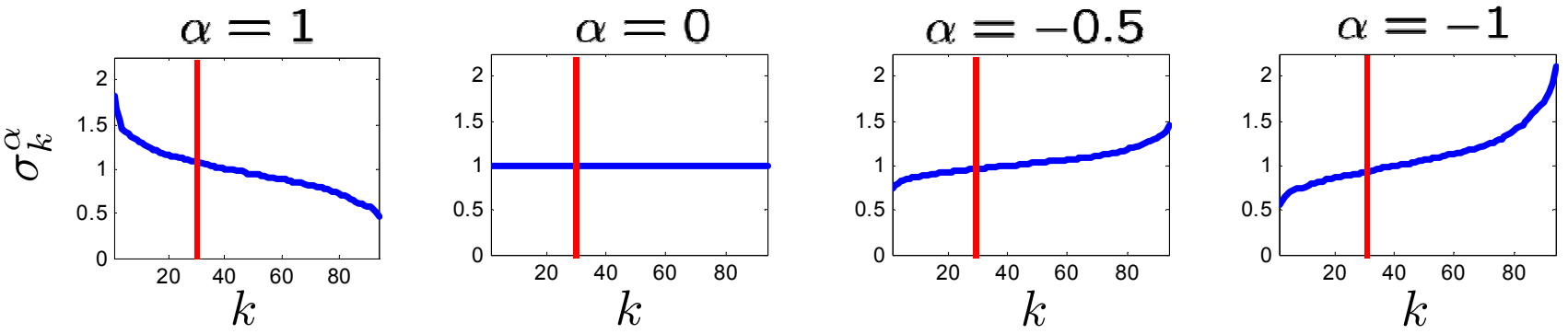
- **Determining useful values for rank**
  - Cross-validation, expectation maximization, Markov chain Monte Carlo, Bayesian inference

[Crossno, P.J., Dunlavy, D.M., Shead, T.M. (2009). IEEE VAST, Atlantic City, NJ.]

LSA: Singular Value (Re)Scaling

- **Document similarities:** $\mathbf{A}_k^{\mathsf{T}}\mathbf{A}_k = \mathbf{D}_k\boldsymbol{\Sigma}_k^2\mathbf{D}_k^{\mathsf{T}}$
- **Inner product view:** $(\mathbf{D}_k\boldsymbol{\Sigma}_k)\,(\mathbf{D}_k\boldsymbol{\Sigma}_k)^{\mathsf{T}}$
- **Scaled inner product view:** $\left(\mathbf{D}_k\boldsymbol{\Sigma}_k^{\alpha}\right)\left(\mathbf{D}_k\boldsymbol{\Sigma}_k^{\alpha}\right)^{\mathsf{T}}$

What is the best scaling for document similarity graph generation? [Data: 97 documents, 335 terms]



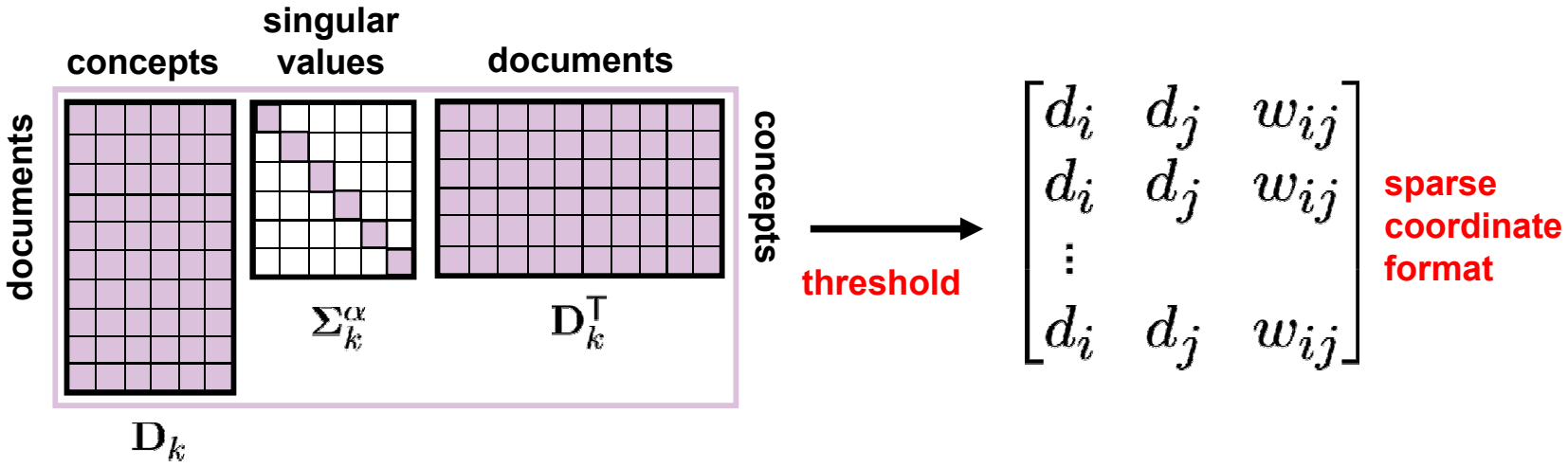
original scaling **no scaling** **inverse sqrt** **inverse**

[Crossno, P.J., Dunlavy, D.M., Shead, T.M. (2009). IEEE VAST, Atlantic City, NJ.]

Sandia National Laboratories

LSA: Document Similarity Graphs

- **Document similarity matrix**



- **Document similarity graph**
  - Each document (or term, entity, etc.) is a vertex
  - Each row defines an edge

LSA: Graph Similarities

- **Statistics on edges**
  - One graph: one-sample t statistic

$$t_{ij} = \frac{\frac{1}{n_s+1}\left\{\sum_{r=K-n_s/2}^{K+n_s/2}\left[\mathbf{D}_r^{\mathsf{T}}\boldsymbol{\Sigma}_r^{\alpha}\mathbf{D}_r\right]_{ij}\right\} - \left[\mathbf{D}_K^{\mathsf{T}}\boldsymbol{\Sigma}_K^{\alpha}\mathbf{D}_K\right]_{ij}}{s/\sqrt{n_s+1}}$$

  - Two graphs: two-sample t statistic

$$t_{ij} = \frac{\frac{1}{n_1+1}\left\{\sum_{r=K_1-n_1/2}^{K_1+n_1/2}\left[\mathbf{D}_r^{\mathsf{T}}\boldsymbol{\Sigma}_r^{\alpha}\mathbf{D}_r\right]_{ij}\right\} - \frac{1}{n_2+1}\left\{\sum_{r=K_2-n_2/2}^{K_2+n_2/2}\left[\mathbf{D}_r^{\mathsf{T}}\boldsymbol{\Sigma}_r^{\alpha}\mathbf{D}_r\right]_{ij}\right\}}{\sqrt{\frac{s_1}{n_1+1}+\frac{s_2}{n_2+1}}}$$

Edges from graph 1

Edges from graph 2

Sandia National Laboratories

LSA: Relevance Feedback

query (t_1, t_2, t_4,…)

terms

concepts

singular values

documents

concepts

statistical bias

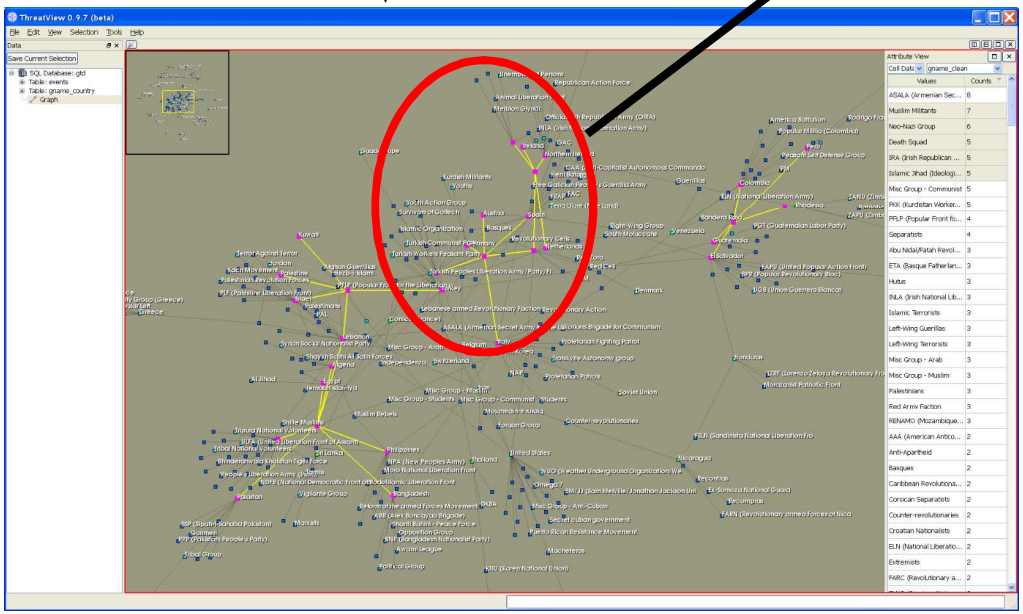
(t_1, t_2, t_4,…) ->

(t_1, t_3, t_7,…)

Database

sought (t_1, t_3, t_7,…)

retrieve selection



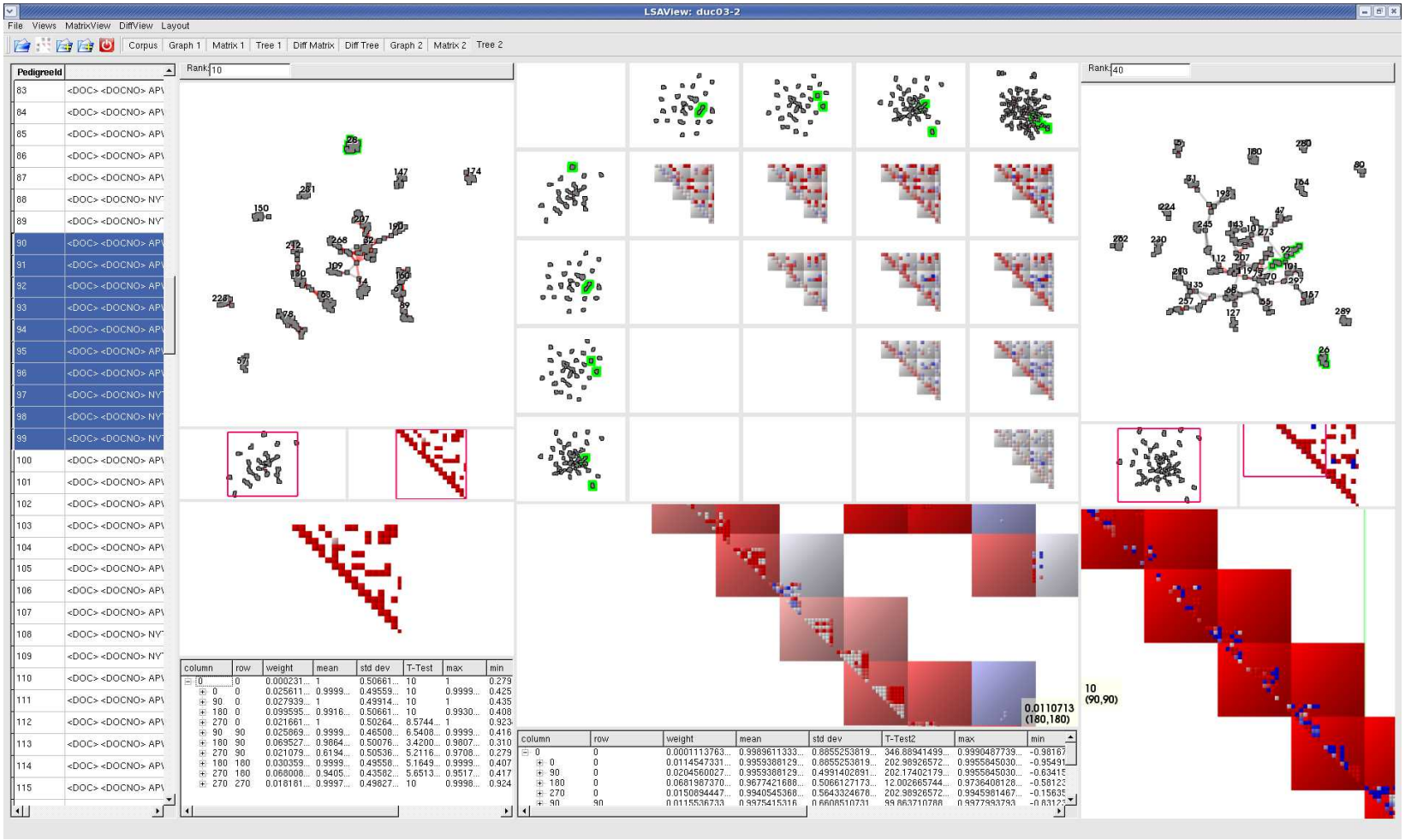
Sandia National Laboratories

Machine Learning and Text Analysis

- **User feedback**
  - Learn how users interact with analysis capabilities
  - Leverage annotation in future analyses
  - Encode knowledge and perspective

- **Example: labeling data as "relevant" / "not relevant"**
  - Posed as a classification problem
    - **Goal:** data instance ⇒ label (class, category)
    - **Method:** supervised learning ⇒ classification models
    - **Ensemble:** combined set of classification models
      - E.g., bagging, boosting, random forest

LSAView



Sandia National Laboratories

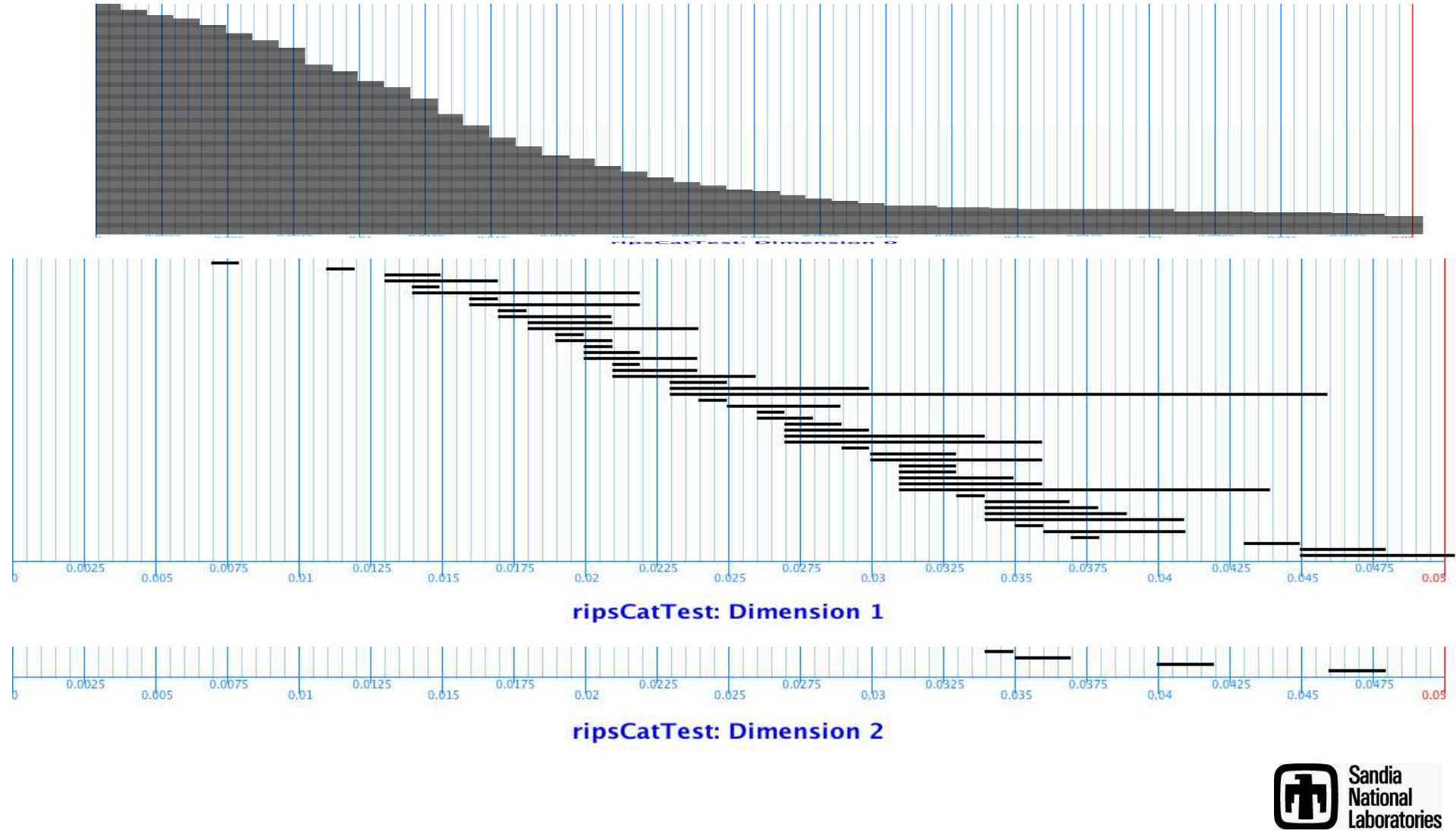
Alternative Approaches for LSA

- **SVD alternatives**
  - Semi-discrete decomposition (SDD),
  - Non-negative matrix factorization (NMF),
  - Matrix subset selection (e.g., CUR)
- **Probabilistic modeling**
  - Probabilistic LSA
  - Latent Dirichlet allocation (LDA)
- **Multiway modeling, semantic graphs**
  - Examples: term-document-author, term-document-time
  - Data is modeled as a multidimensional array (tensor)
  - Tensor decompositions
    - PARAFAC, Tucker, DEDICOM, …

Research questions

- **Is there useful structure in the graphs generated using LSA?**
- **Can we find the structure and map it to analysts' knowledge?**
- **Is the structure persistent with respect to LSA parameters?**
- **Can understanding of persistent homology lead to improved algorithms for knowledge discovery?**
- **How sensitive are structures with respect to …**
  - LSA parameters? Data outliers? Data noise? Changes over time?
- **How do we communicate structure and persistence to users?**
- **Is it possible to detect persistence for dynamic data?**
- **Is it possible to detect persistence for streaming data?**
- **What does structural persistence mean for semantic graphs and can it be computed?**

JPlex Example



Sandia National Laboratories

Thank You

Persistent homology for parameter sensitivity in large-scale text-analysis (informatics) graphs

Danny Dunlavy

dmdunla@sandia.gov

http://www.cs.sandia.gov/~dmdunla

Sandia National Laboratories