SAND2008-2241P

Using Visualization for Relevancy Feedback Tuning of Text Analysis Algorithms

Patricia Crossno
Danny Dunlavy
Tim Shead

April 4, 2008

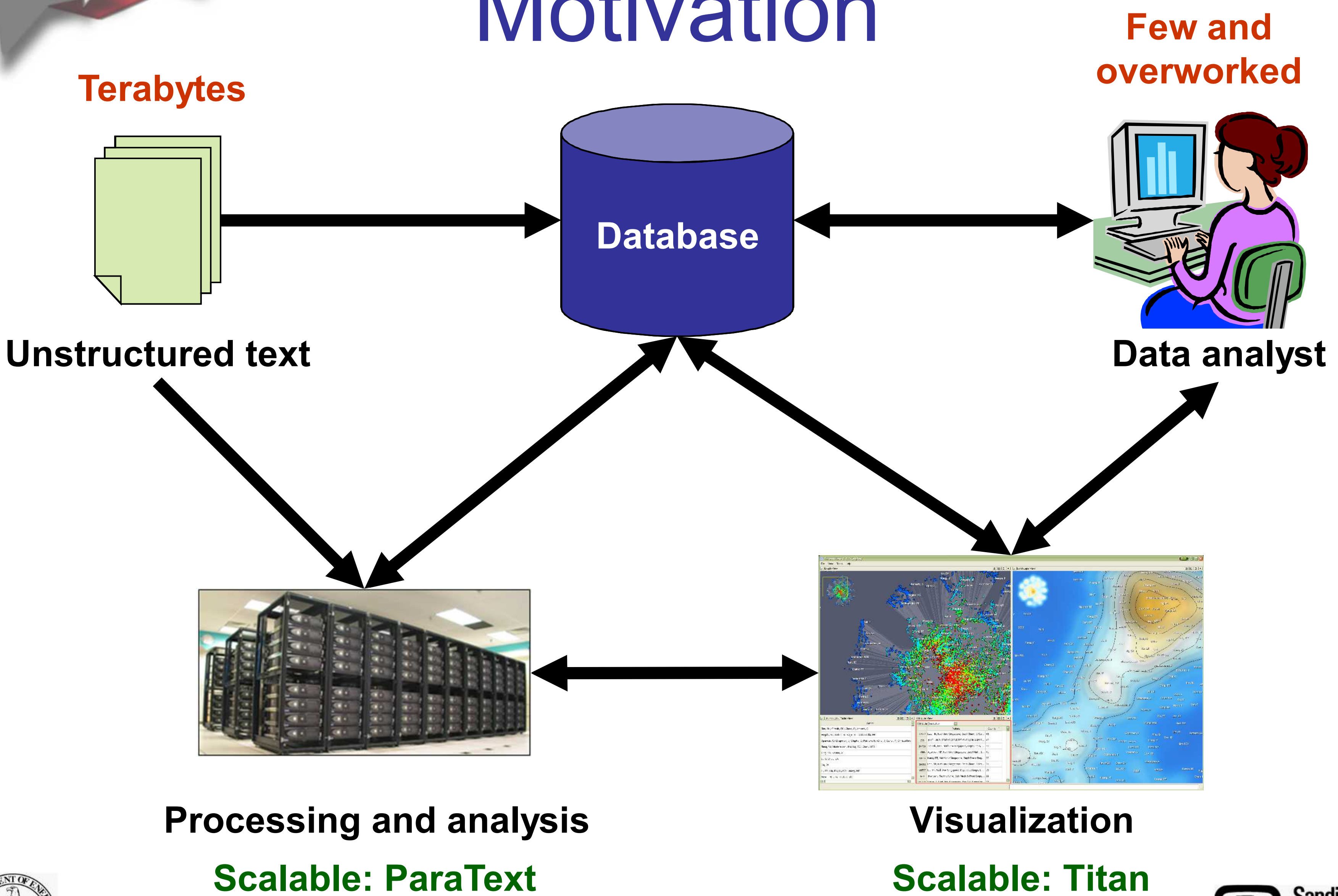
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

Sandia National Laboratories

Outline

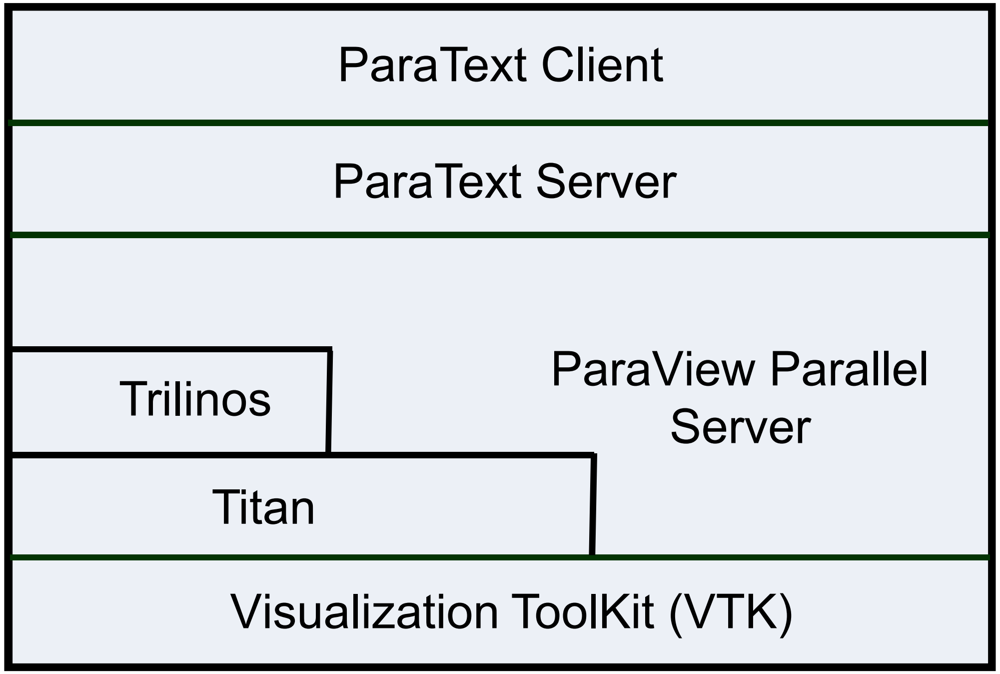
- Motivation
- ParaText
- Latent Semantic Analysis (LSA)
- LSALIB
- Sensitivity Analysis
- Relevancy Feedback

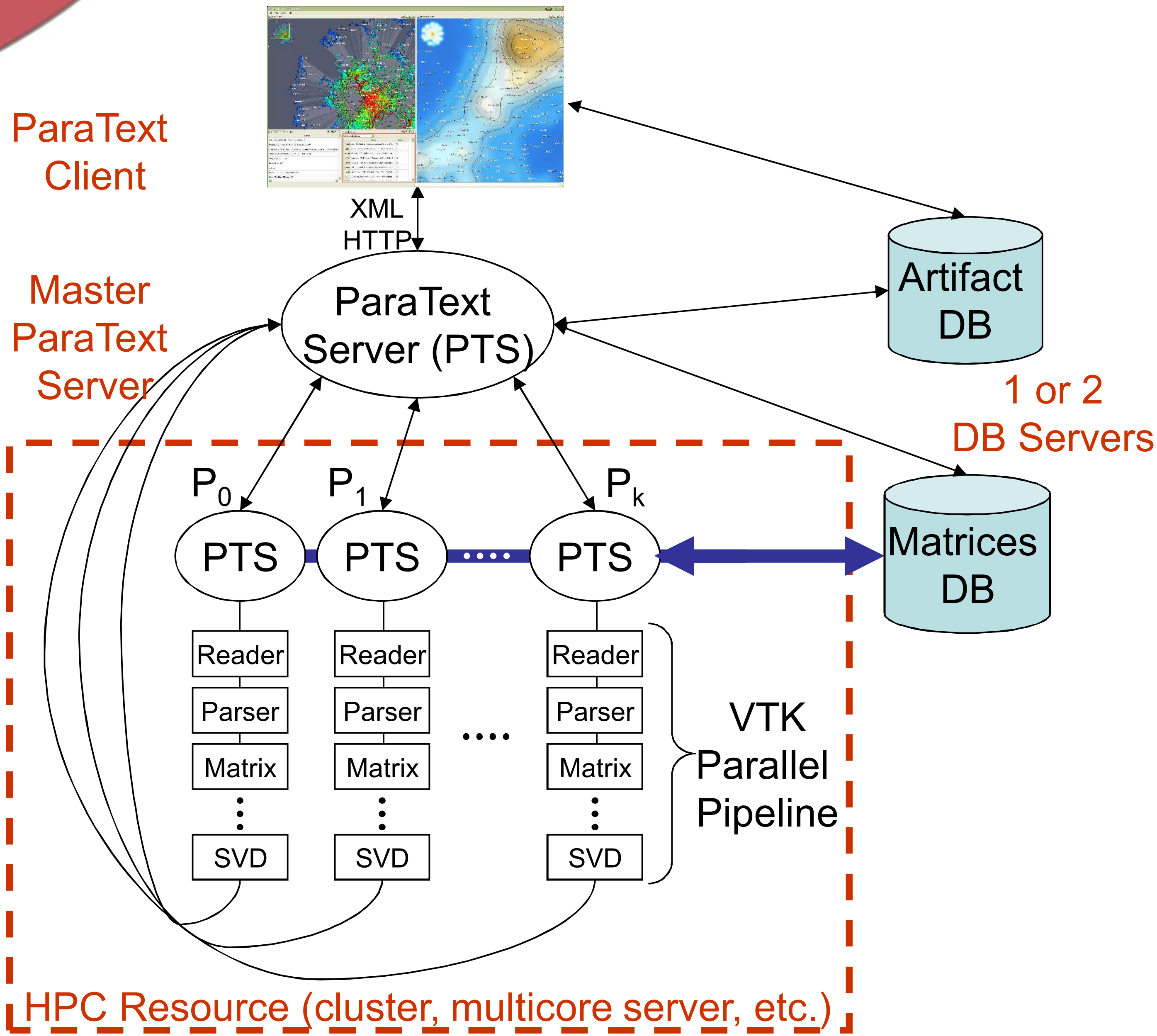
Motivation



Terabytes
Few and overworked
Database
Unstructured text
Data analyst
Processing and analysis
Scalable: ParaText
Visualization
Scalable: Titan

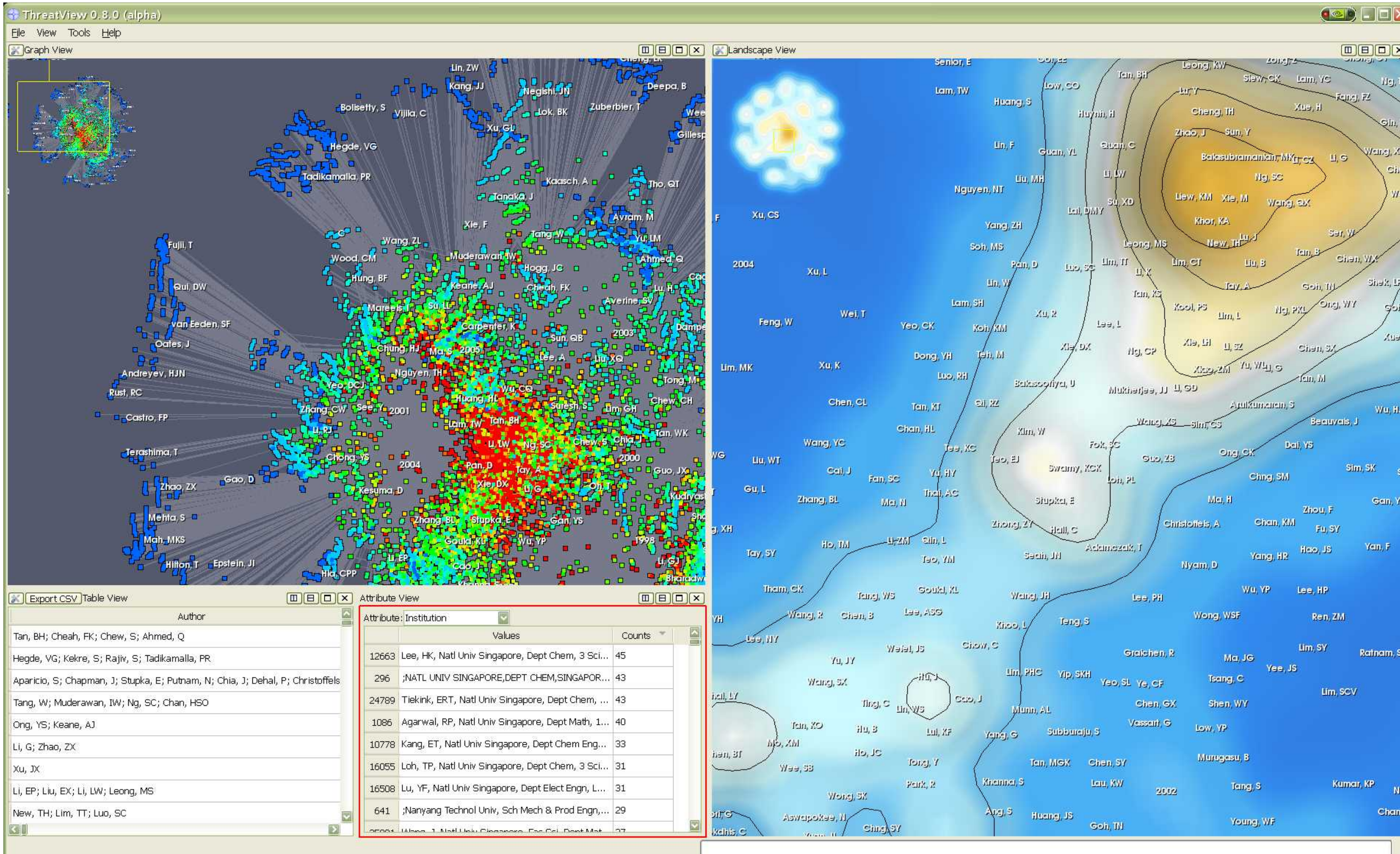
Scalable Solutions for Processing and Searching Very Large Document Collections (ParaText)





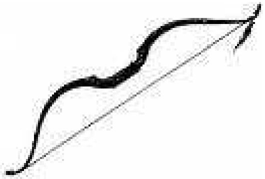
ParaText Client
XML
HTTP
Master ParaText Server
ParaText Server (PTS)
Artifact DB
1 or 2 DB Servers
P_0
P_1
P_k
PTS
PTS
PTS
Matrices DB
Reader
Parser
Matrix
SVD
Reader
Parser
Matrix
SVD
Reader
Parser
Matrix
SVD
VTK Parallel Pipeline
HPC Resource (cluster, multicore server, etc.)

Landscape Metaphor



Text Analysis Issues

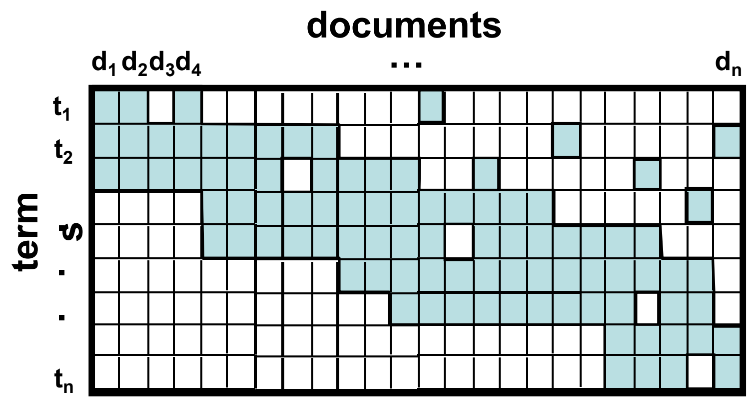
- keyword searching does not work well
  - miss relevant information
  - retrieve irrelevant information
- words with multiple meanings

- different words with the same meaning
  - baby and infant
  - sick and ill
- word relationships can distinguish both
- Latent Semantic Analysis [Dumais et al., 1988]

Modeling Text

- model documents as a linear equation
  - meaning (document) = Σ_j meaning ($term_j$)
- ignore term order and syntax
- discard non-differentiating words (stop list)
  - articles, prepositions, conjunctions, pronouns
  - common verbs, common adjectives
- remove common endings, like ‘ing’ (stemming)
- create term-document (occurrence) matrix



Matrix Size

- English growing - no upper bound
- How many words do we use?
  - Count lemmas (base words)
  - Based on Oxford English Corpus (OEC)
  - # Lemmas % of content in OEC

# Lemmas	% of content in OEC
10	25%
100	50%
1000	75%
7000	90%
50,000	95%
>1M	99%

  - last few % consists of rare or highly technical terms
    - *chrondrogenesis* or *dicarboxylate*
- Term dimension dominates until document count exceeds lemmas used

Term Weighting

Want to weight by information content

- weighting within a document
  - common words more meaningful
- weighting across documents
  - uncommon words differentiate
- normalization by document size
  - prevents large documents from dominating

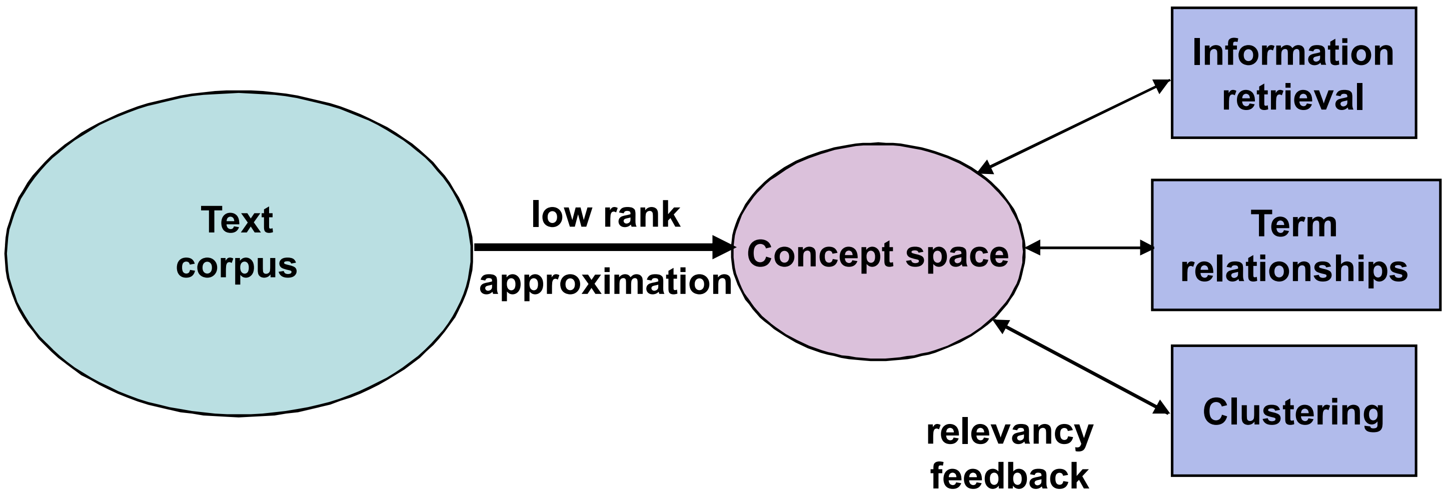
Multiply the three factors together

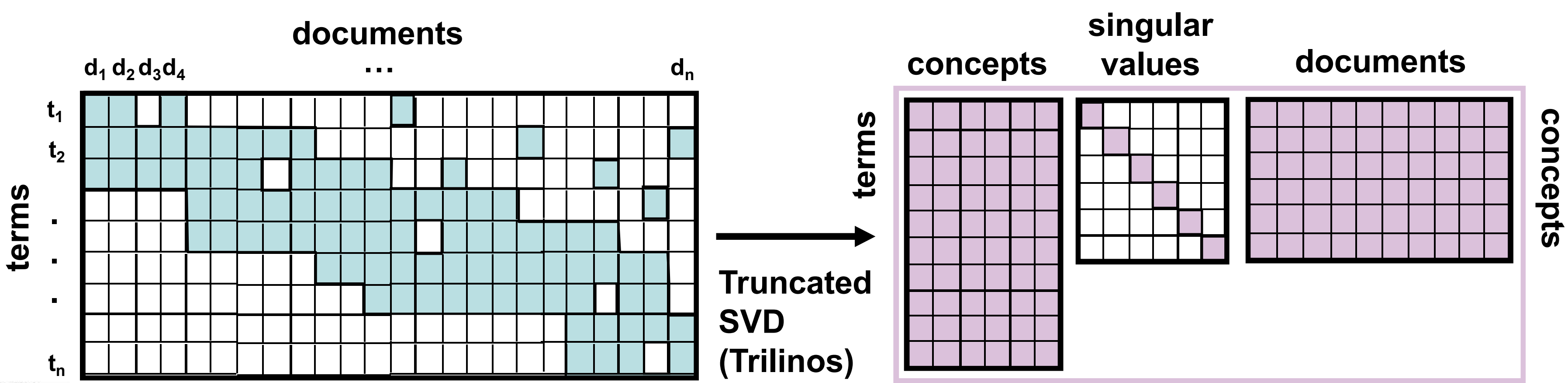
Term Weighting Options

$$\boxed{a_{ij} = \tau_{ij} \cdot \gamma_i \cdot \delta_j}$$

	Local Weights (τ_{ij})	
Term Frequency	**individual documents (columns)**	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	**over all documents (rows)**	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	**individual documents**	1
Normalized		$\left(\sum_i (\tau_{ij}\gamma_i)^2\right)^{-1/2}$

Latent Semantic Analysis





Concept Space

- high-dimensional (50-D to 1500-D)
- documents are points
- similarity = relationship in concept space
  - geometrically close = conceptually close
  - geometrically distant = conceptually distance
  - no exact keyword matching
- truncated SVD reduces dimensionality, removing noise through a low-rank approximation
- truncation level determines number of concepts
- query
  - project query text into concept space
  - return nearby documents

LSALIB: 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
$q^T A$	.89	.71	0	0

rank-2 approximation

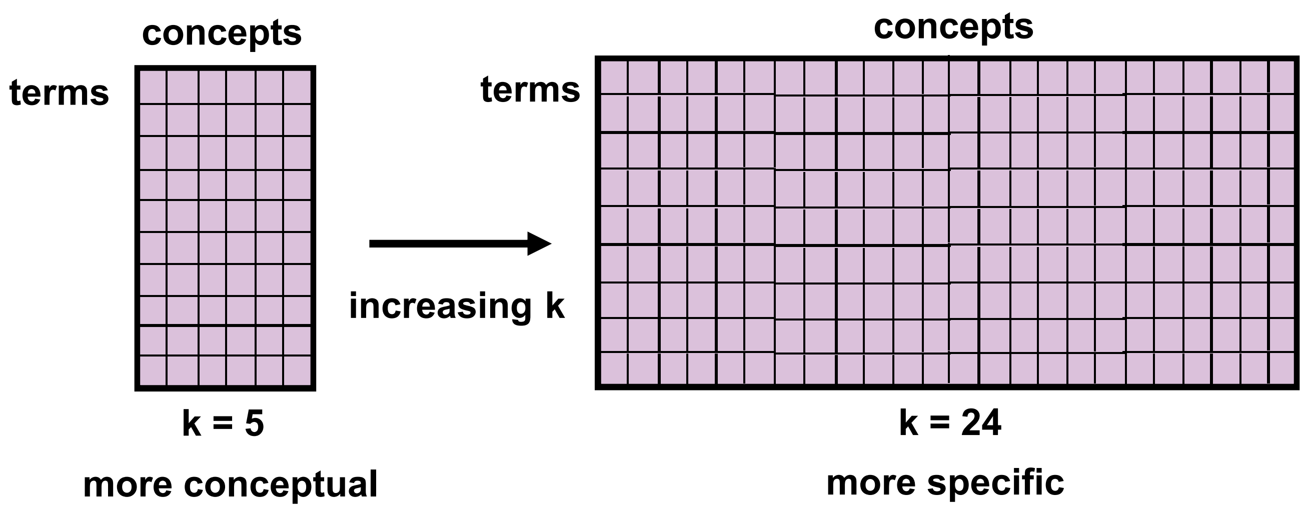
A_2	d_1	d_2	d_3	d_4
hurricane	.78	.78	-.11	.11
earthquake	-.03	.02	.96	.92
catastrophe	.59	.60	.15	.30
$q^T A_2$	.78	.78	–	.11

captures link to doc 4

Sandia National Laboratories

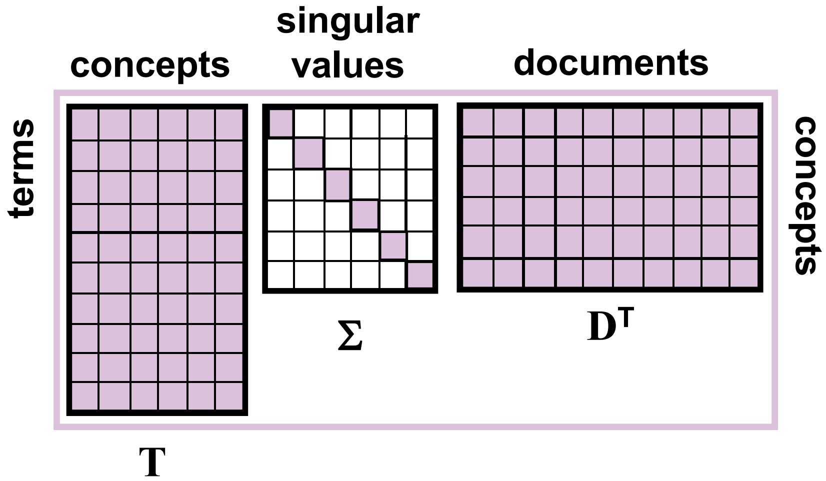
LSALIB

- Implements latent semantic analysis
  - Conceptual searching
    - rank(k) ↑ : more exact matches
    - rank(k) ↓ : more conceptual matches
    - Can compute larger rank and use smaller rank



LSALIB: Matrix Operations

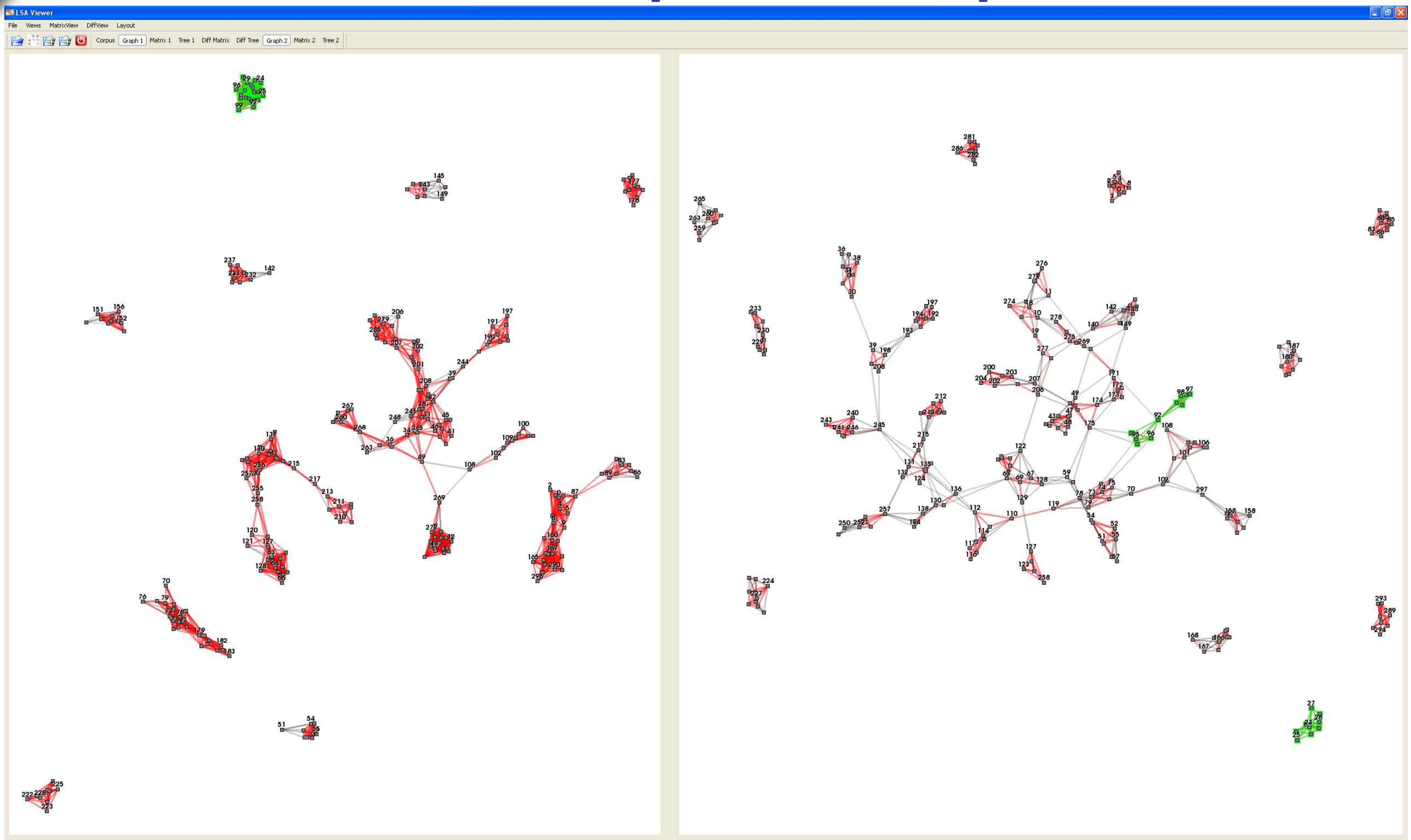
- SVD: $\mathbf{A} = \mathbf{T}\mathbf{\Sigma}\mathbf{D}^\mathsf{T}$
- Truncated: $\mathbf{A} \approx \mathbf{A}_k = \mathbf{T}_k\mathbf{\Sigma}_k\mathbf{D}_k^\mathsf{T} = \sum_{r=1}^{k} \sigma_r\, \mathbf{t}_r\mathbf{d}_r^T$
- Query scores (query as new “doc”): $q^\mathsf{T}\mathbf{A}$
- LSA Ranking: $q^\mathsf{T}\mathbf{A}_k$
- Document similarities: $\mathbf{D}_k\mathbf{\Sigma}_k^2\mathbf{D}_k^\mathsf{T}$ (want sparse output)
- Term Similarities: $\mathbf{T}_k\mathbf{\Sigma}_k^2\mathbf{T}_k^\mathsf{T}$ (want sparse output)



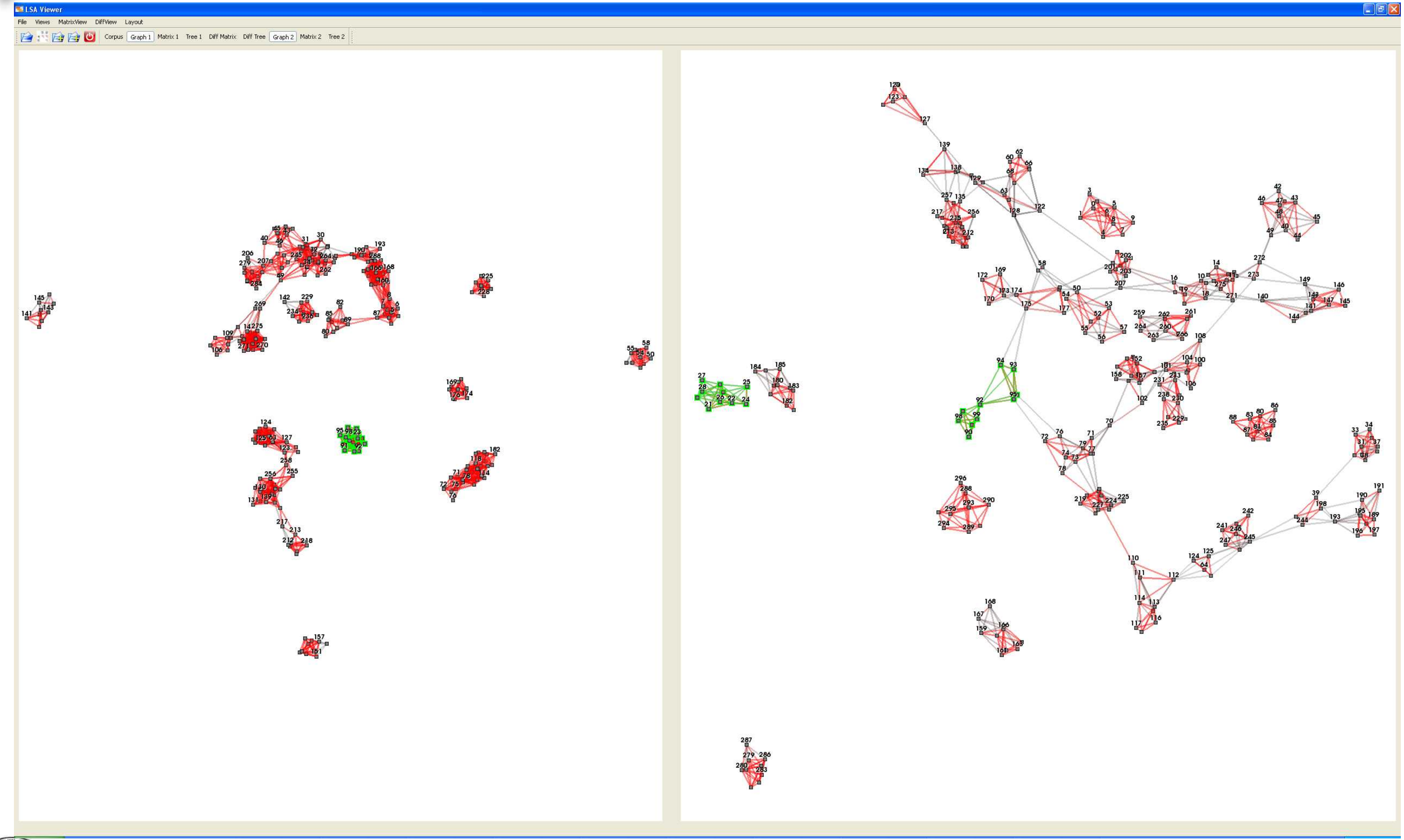
Sensitivity Analysis

– What is the sensitivity of LSA to different parameter choices?
– How does conceptual clustering change with rank?
– Does the layout algorithm change our view of the conceptual cluster?
– Is a change in document similarity edge weighting significant?
– How do different weighting choices impact all of the above?

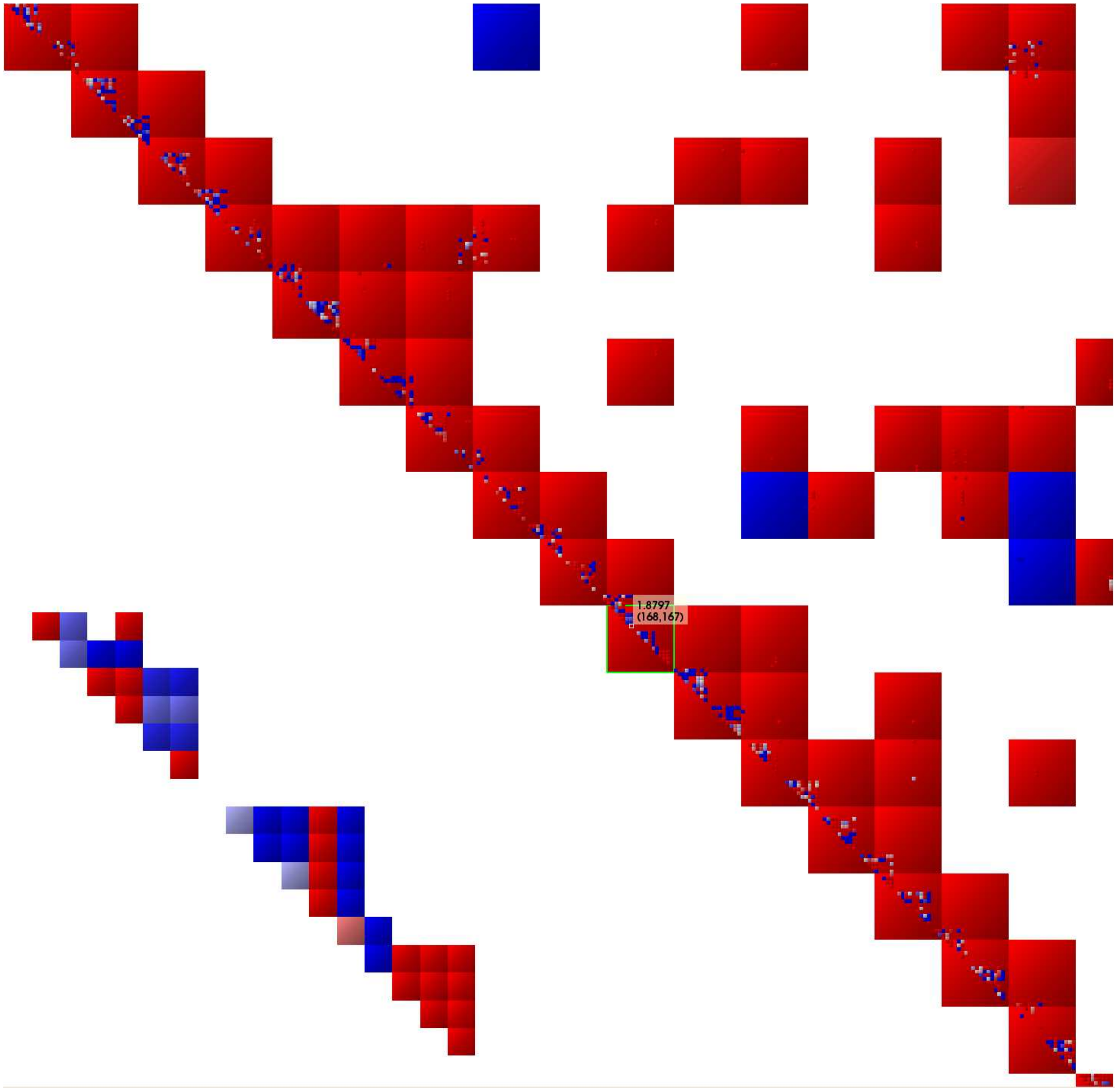
Doc Sim Graph Comparison



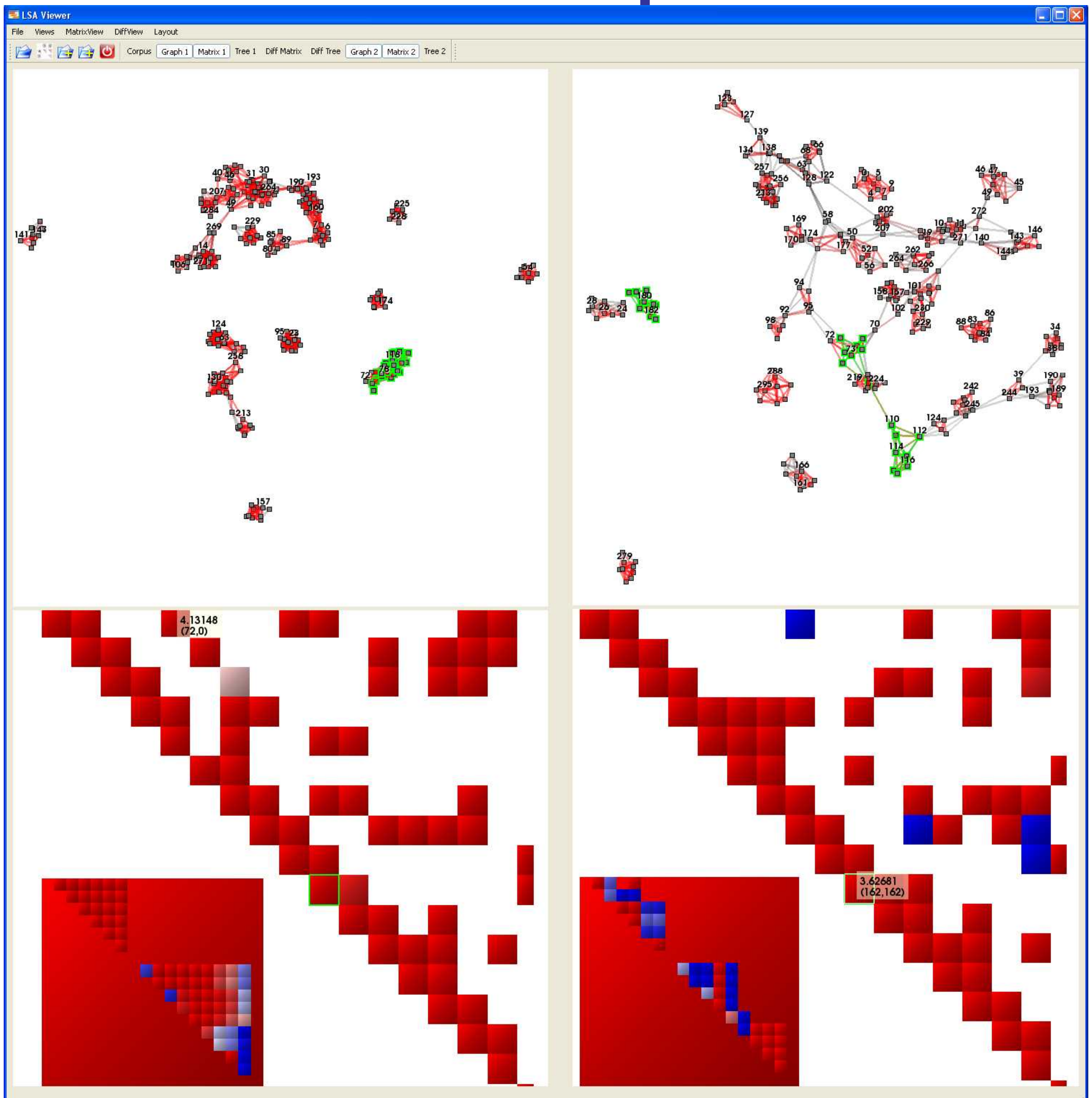
Layout Comparison



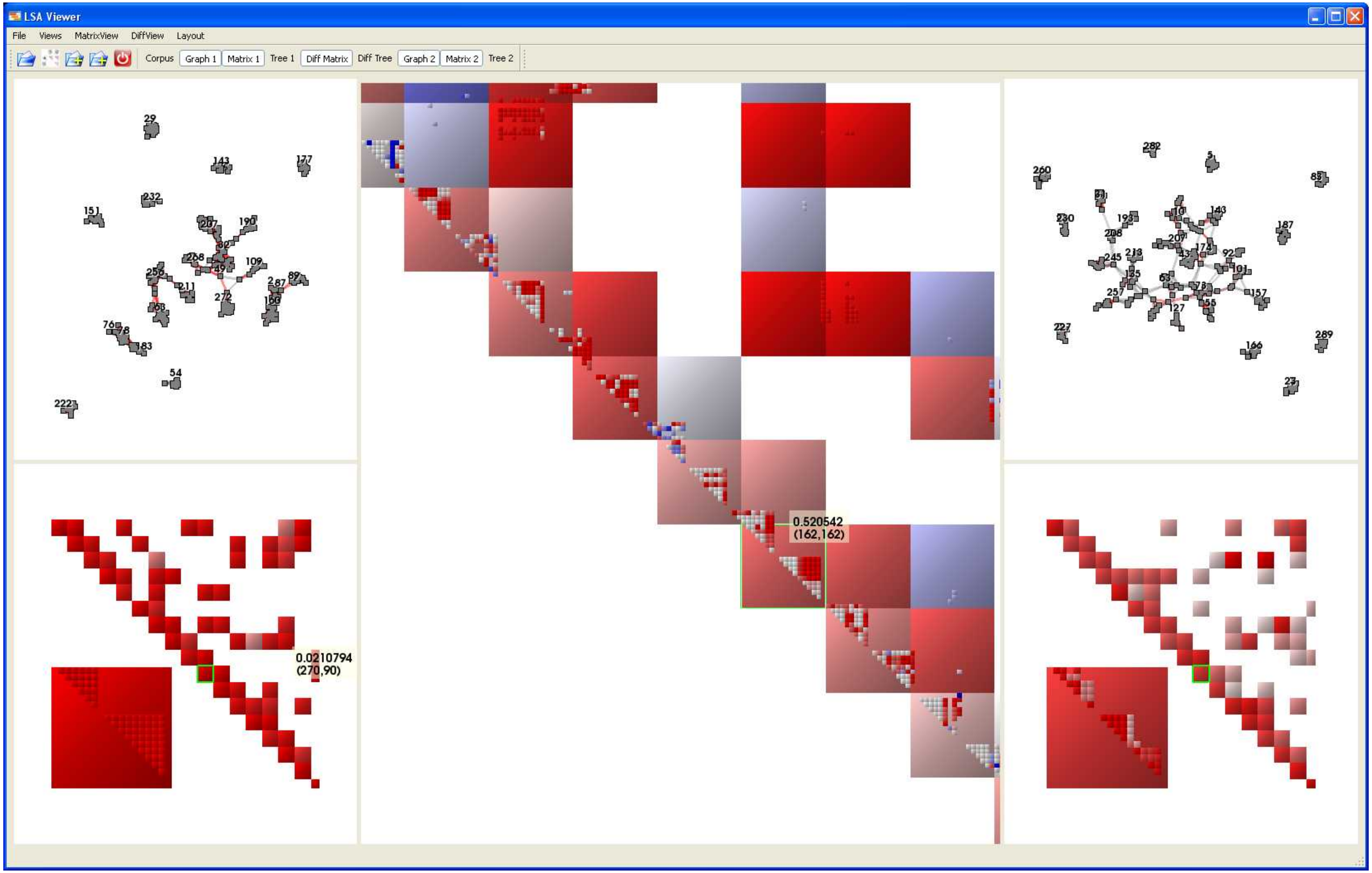
Sparse Matrix View



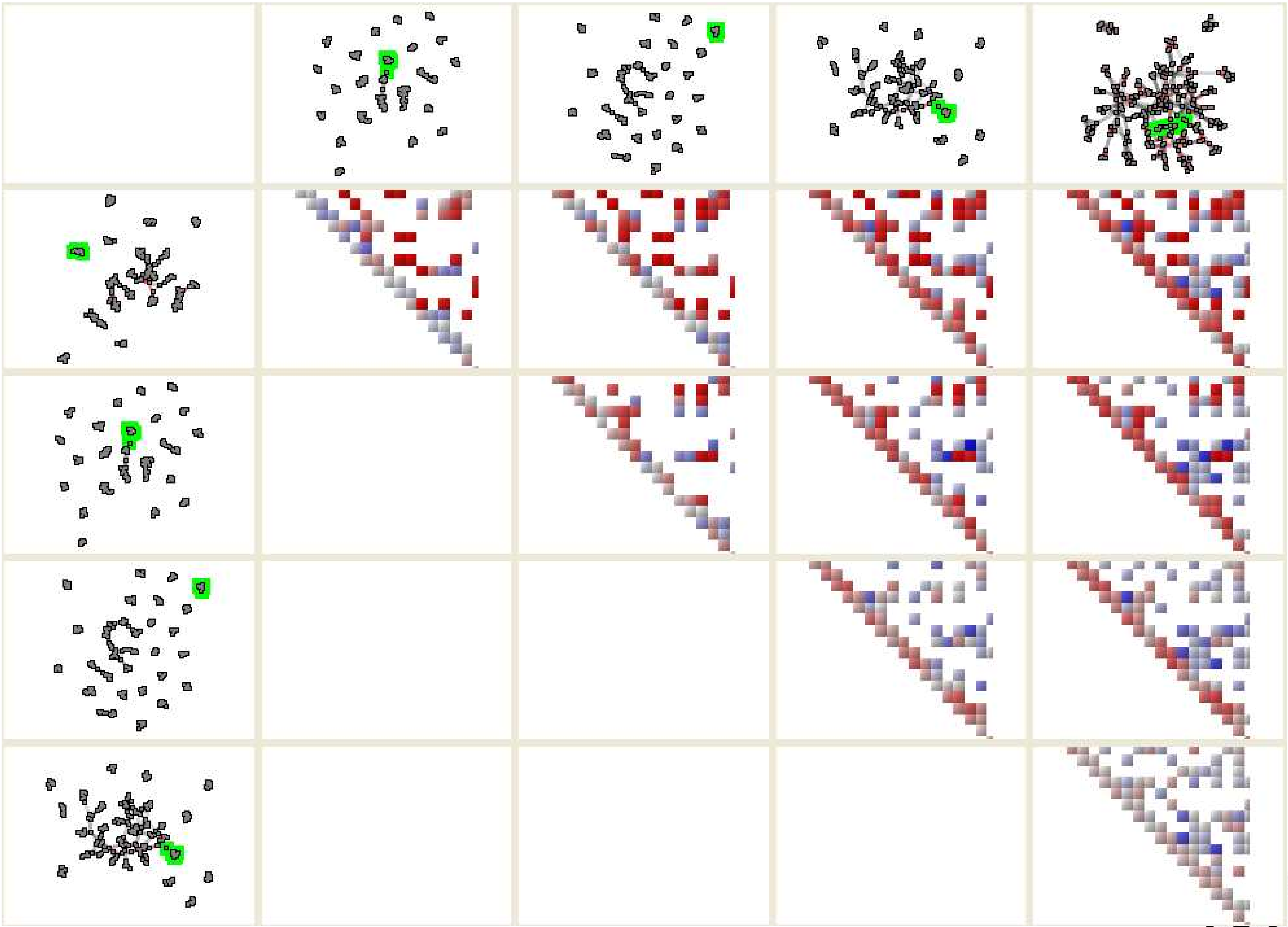
Rank Comparison



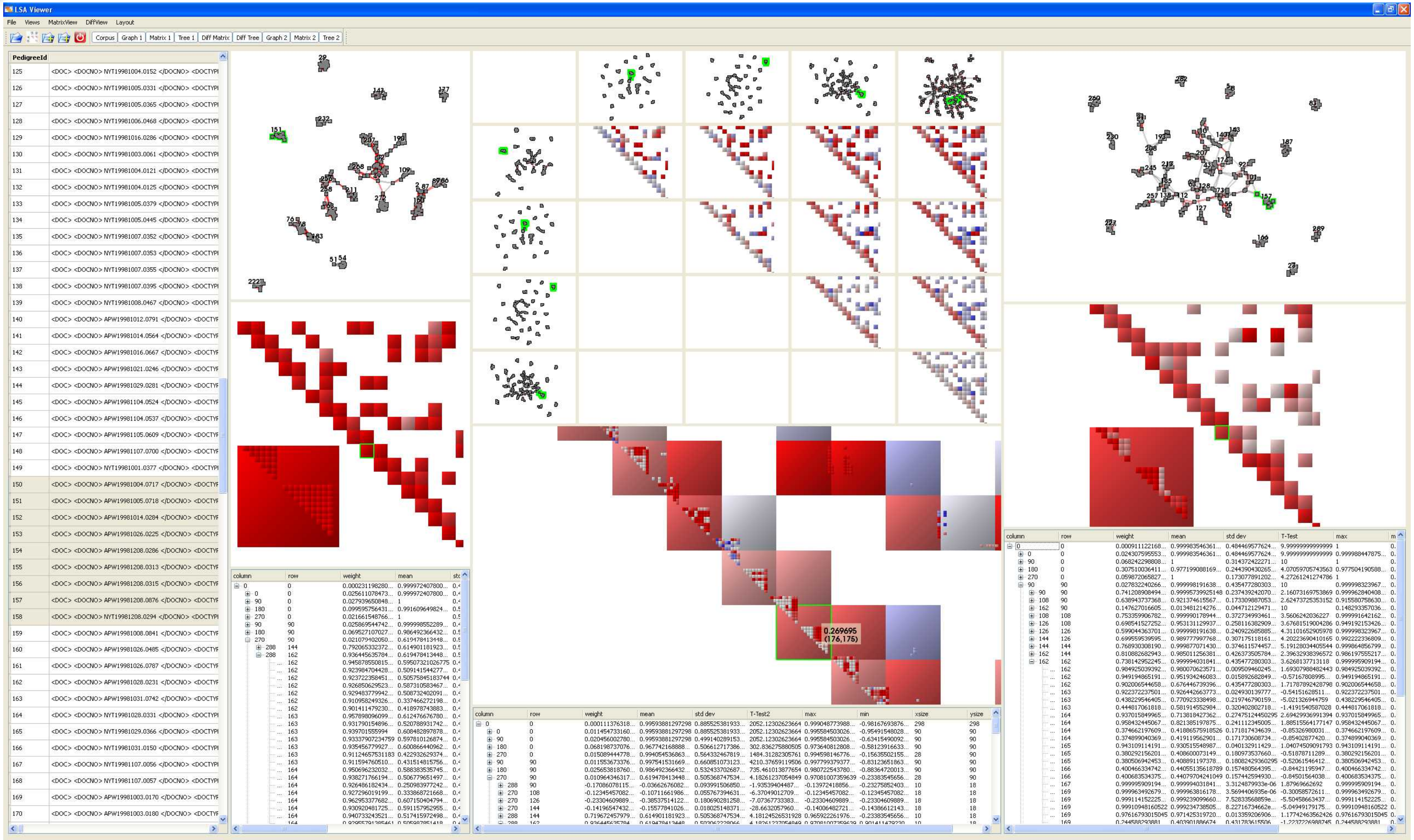
Matrix Differences



Small Multiples

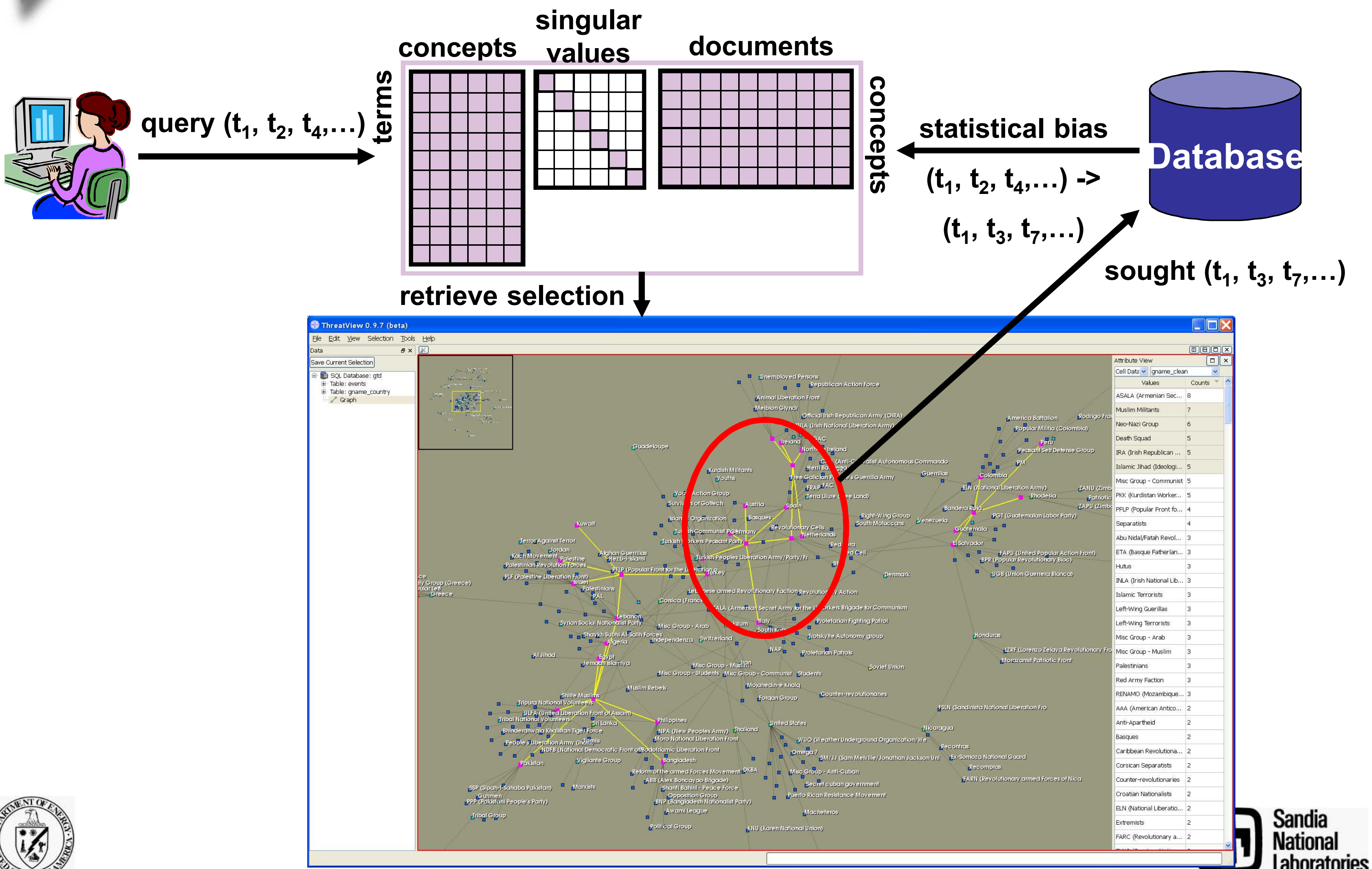


LSAView



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

Relevancy Feedback



Questions?