



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

CSRI Workshop on Combinatorial Algebraic Topology (CAT)

Danny Dunlavy

Computer Science and Informatics Department (1415)

Sandia National Laboratories

August 29, 2009

SAND2009-5518C

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.





Acknowledgments

- **Text Analysis**

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

- **Computational Topology**

- David Day, Scott Mitchell, Shawn Martin

- **JPlex in Matlab**

- Henry Adams

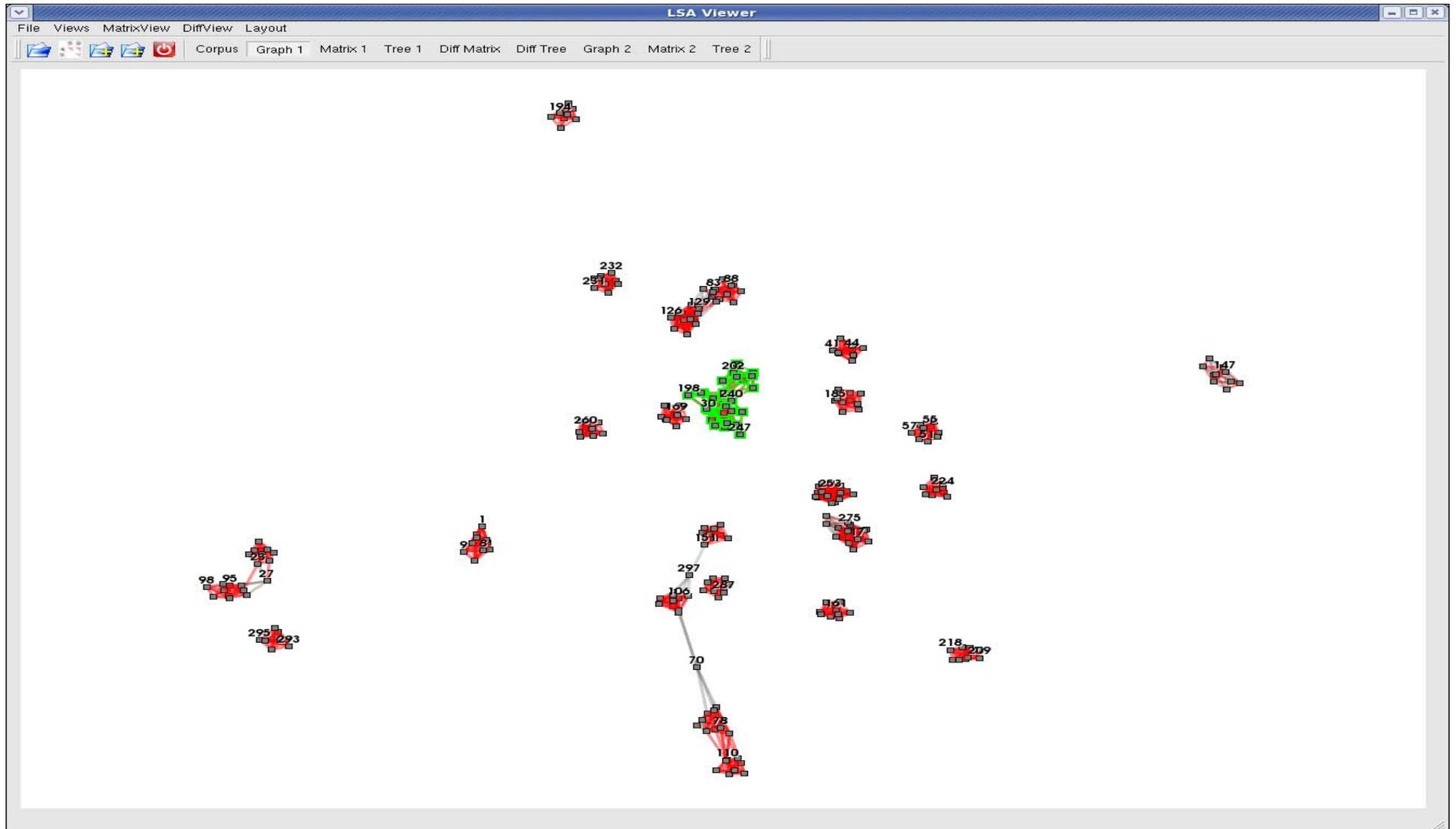


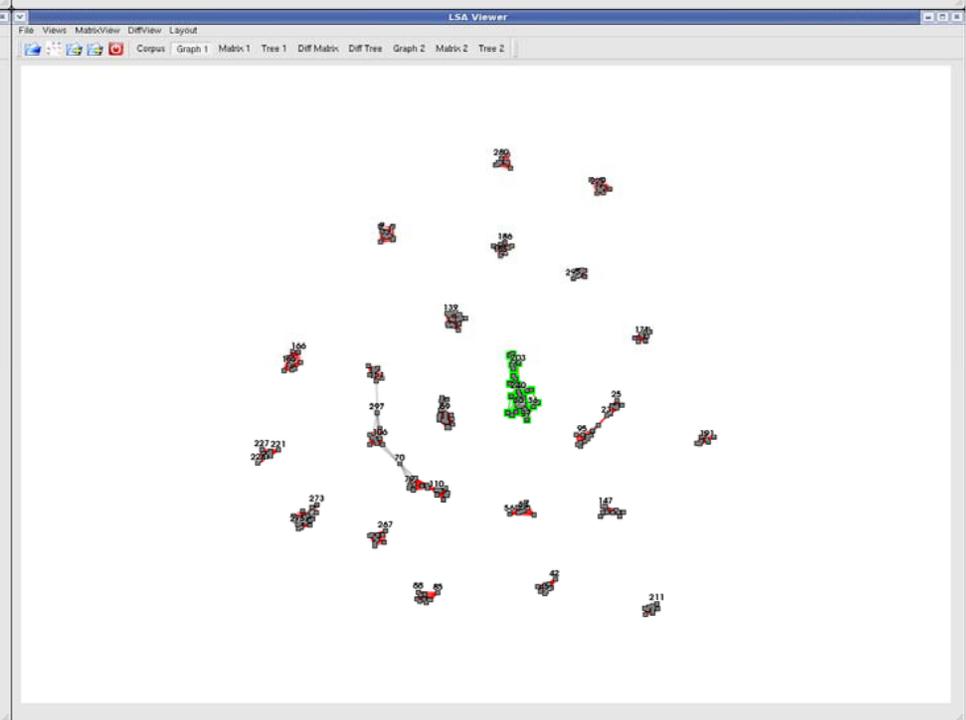
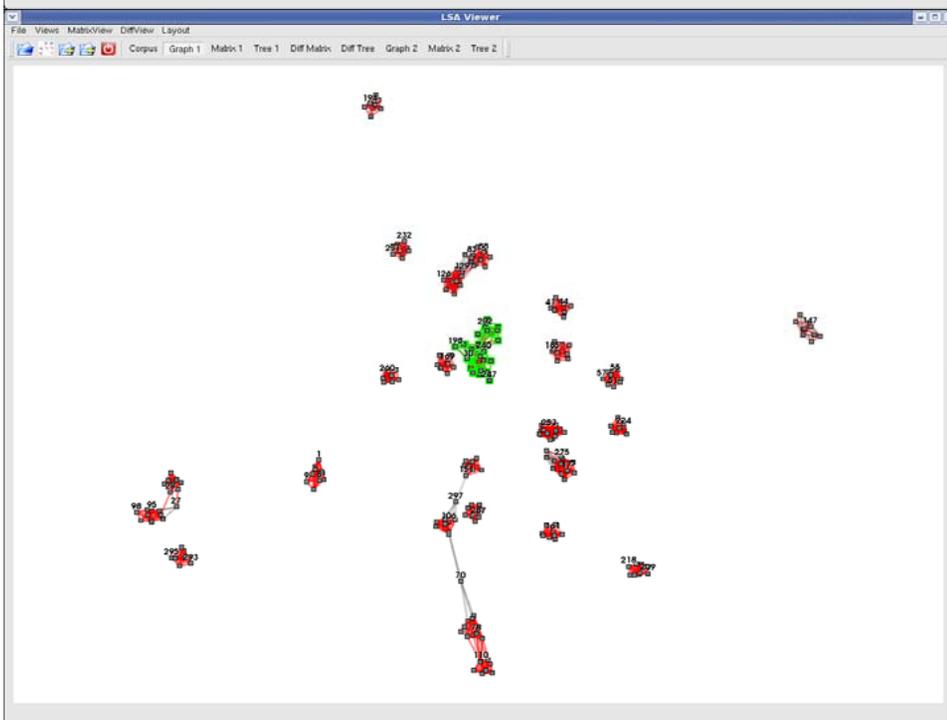
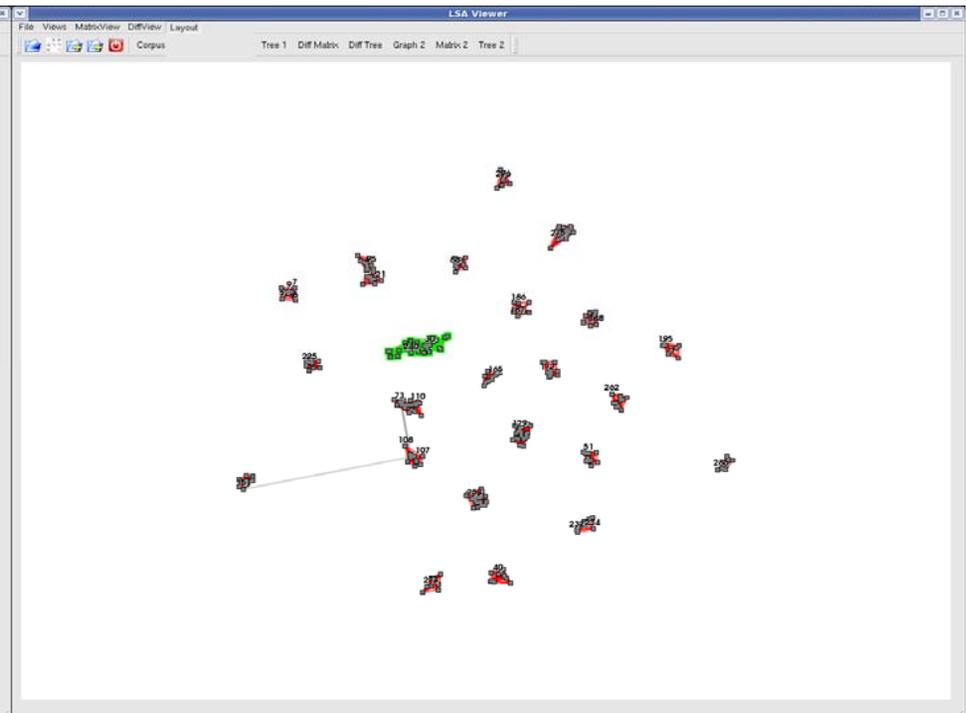
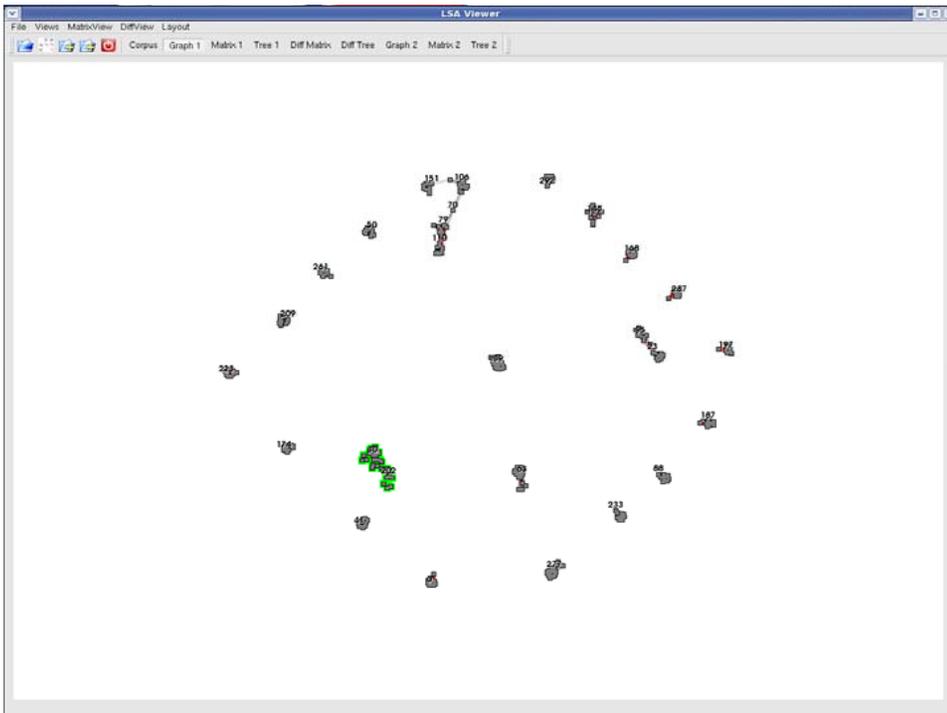
Start with text ...

<DOC> <DOCNO> APW19981017.0151 </DOCNO> <DOCTYPE> NEWS STORY </DOCTYPE> <DATE_TIME> 10/17/1998 05:50:00 </DATE_TIME> <HEADER> w2985 &Cx1f; wstm- r i &Cx13; &Cx11; BC-Britain-Pinochet 10-17 0337 </HEADER> <BODY> <SLUG> BC-Britain-Pinochet </SLUG> <HEADLINE> Lawmaker: Britain must question Gen. Pinochet about killings, </HEADLINE> torture &QL; <TEXT> LONDON (AP) _ An influential lawmaker from the governing Labor Party on Saturday backed Spanish requests to question former Chilean dictator Gen. Augusto Pinochet, in London for back surgery, on allegations of genocide and terrorism. "There've been something like a total of 3,197 cases" of murder and torture during Pinochet's 1973-90 regime, Ann Clywd said in an interview with British Broadcasting Corp. radio. "In the course of a long search by his relatives, the remains of some of the disappeared have been discovered in unmarked graves," said Ms. Clywd, who chairs an all-party committee of lawmakers on human rights. "And hundreds of former detainees have made statements confirming that the disappeared were held in detention centers." "Clearly, the Spanish authorities have every right to try and investigate these matters further." Pinochet, 82, who underwent surgery for a herniated disc a week ago, was reported to have left London Bridge Hospital earlier Wednesday. British authorities have declined to comment on his whereabouts or the Spanish requests. Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday. Another judge, Manuel Garcia Castellon, filed a request to question Pinochet a day earlier. In 1996, Castellon opened his probe into murder, torture and disappearances in Chile during Pinochet's regime. Garzon is also investigating the disappearance of hundreds of Spanish citizens in Argentina during the 1976-83 military dictatorships. Pinochet, who ousted elected President Salvador Allende in a bloody 1973 coup, remained commander-in-chief of the Chilean army until March, when he was sworn in as a senator-for-life, a post established for him in a constitution drafted by his regime. (scl) </TEXT> (PROFILE (WS SL:BC-Britain-Pinochet; CT:i; (REG:EURO; (REG:BRIT; (REG:SCAN; (REG:MEST; (REG:AFRI; (REG:INDI; (REG:ENGL; (REG:ASIA; (LANG:ENGLISH;)))) </BODY> <TRAILER> AP-NY-10-17-98 0550EDT </TRAILER> </DOC> <DOC> <DOCNO> APW19981017.0306 </DOCNO> <DOCTYPE> NEWS STORY </DOCTYPE> <DATE_TIME> 10/17/1998 09:20:00 </DATE_TIME> <HEADER> w1148 &Cx1f; wstm- u i &Cx13; &Cx11; BC-Britain-Pinochet 4thLd-Writethru 10-17 0503 </HEADER> <BODY> <SLUG> BC-Britain-Pinochet, 4th Ld-Writethru </SLUG> <HEADLINE> British police arrest Pinochet on murder charges </HEADLINE> &UR; Eds: UPDATES with reaction, fresh details; adds byline &QL; &UR; By SUE LEEMAN &QC; &UR; Associated Press Writer &QC; <TEXT> LONDON (AP) _ British police said Saturday they have arrested former Chilean dictator Gen. Augusto Pinochet on allegations of murdering Spanish citizens during his years in power. Pinochet, 82, in London for surgery, was held Friday night after British authorities received a Spanish extradition warrant, a Scotland Yard spokeswoman said. The warrant charges that between Sept. 11 1973, the year he seized power, and Dec. 31, 1983, Pinochet murdered Spanish citizens in Chile, the spokeswoman said, speaking anonymously. The spokeswoman refused to confirm Pinochet's whereabouts, but Pinochet's press secretary in Santiago said he being held in the London clinic where he underwent surgery for a herniated disc on Oct. 9. No hearing date has been set. Spanish Foreign Minister Abel Matutes, attending the Ibero American summit in Porto, Portugal, said his government "respects the decisions taken by courts." But Chilean Ambassador to London, Mario Artaza, said he will seek the release of Pinochet "what we must do is make it clear that Mr. Pinochet is a senator, who travels with a diplomatic passport," Artaza said. The government has previously said that Chile does not recognize the authority of foreign courts over situations which occurred in Chile. Pinochet, who turns 83 next month, was reported to have left London Bridge Hospital on Wednesday. But his press secretary Fernando Martinez said he was at the clinic Friday night when police came for him. Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday. Another judge, Manuel Garcia Castellon, filed a request to question Pinochet a day earlier. Castellon's probe into murder, torture and disappearances in Chile during Pinochet's regime began in 1996. Garzon is also investigating the disappearance of hundreds of Spanish citizens in Argentina during the 1976-83 military dictatorships. Pinochet is implicated in Garzon's probe through his involvement in "Operation Condor," in which military regimes in Chile, Argentina and Uruguay coordinated anti-leftist campaigns The judges' petitions are based on the European Convention on Terrorism which requires signatories to cooperate with each others' judicial processes in cases of terrorism, according to Juan Garces, a lawyer involved in the Spanish investigation into human rights violations in Chile. Pinochet, who ousted elected President Salvador Allende in a bloody 1973 coup, remained commander-in-chief of the Chilean army until March, when he was sworn in as a senator-for-life, a post established for him in a constitution drafted by his regime. (scl) </TEXT> (PROFILE (WS SL:BC-Britain-Pinochet, 4th Ld-Writethru; CT:i; (REG:EURO; (REG:BRIT; (REG:SCAN; (REG:MEST; (REG:AFRI; (REG:INDI; (REG:ENGL; (REG:ASIA; (LANG:ENGLISH;)))) </BODY> <TRAILER> AP-NY-10-17-98 0920EDT </TRAILER> </DOC> <DOC> <DOCNO> APW19981017.0477 </DOCNO> <DOCTYPE> NEWS STORY </DOCTYPE> <DATE_TIME> 10/17/1998 12:15:00 </DATE_TIME> <HEADER> w1339 &Cx1f; wstm- u i &Cx13; &Cx11; BC-Britain-Pinochet 5thLd-Writethru 10-17 0733 </HEADER> <BODY> <SLUG> BC-Britain-Pinochet, 5th Ld-Writethru </SLUG> <HEADLINE> British police arrest Pinochet on murder charges </HEADLINE> &UR; Eds: AMS. UPDATES, recasts &QL; &UR; By SUE LEEMAN &QC; &UR; Associated Press Writer &QC; <TEXT> LONDON (AP) _ Eight years after his turbulent regime ended, former Chilean strongman Gen. Augusto Pinochet is being called to account by Spanish authorities for the deaths, detention and torture of political opponents. Responding to a Spanish extradition warrant, British police announced Saturday they have arrested Pinochet on allegations that he murdered an unidentified number of Spaniards in Chile between Sept. 11, 1973, the year he seized power, and Dec. 31, 1983. No reason for the dates was given. Chile said it would protest to British authorities, arguing that the 82-year-old senator-for-life has diplomatic immunity. But Britain said Pinochet did not have diplomatic immunity. Prime Minister Tony Blair's office said it was "a matter for the magistrates and the police." Pinochet, whose ruthless regime was widely criticized for its human rights record, was recovering from surgery in a London clinic when he was held Friday night. No hearing date has been set. Scotland Yard refused to confirm Pinochet's whereabouts, but his Santiago spokesman Fernando Martinez said he was in a London clinic when police came for him. A regular visitor to Britain, Pinochet underwent surgery Oct. 9 for a herniated disc, a spinal disorder which has given him pain and hampered his walking in recent months. In a statement issued in Porto, Portugal, where President Eduardo Frei is attending the Ibero American summit, the Chilean government said it is "filing a formal protest with the British government for what it considers a violation of the diplomatic immunity which Sen. Pinochet enjoys." The statement, read by acting foreign minister Mariano Fernandez, demanded "that steps be taken to allow an early end of this situation." Chile has previously said it does not recognize the authority of foreign courts over incidents within Chile. Spanish Foreign Minister Abel Matutes, also attending the Ibero American summit, said his government "respects the decisions taken by courts." British law recognizes two types of immunity: state immunity, which covers heads of state and government members on official visits, and diplomatic immunity for persons accredited as diplomats. Jeremy Corbyn, a lawmaker from Britain's governing Labor Party, applauded the arrest. "It will be the first time this ghastly dictator has faced questions," he told Sky television. "He is one of the great murderers of this century." Richard Bunting of the human rights group Amnesty International, which has frequently criticized Pinochet, said the British government was "under obligation to take legal action" against him. It was not immediately clear which clinic was treating Pinochet, who is 83 next week. Staff at the London Bridge Hospital, where he reportedly had surgery, refused comment. He has a pacemaker and hearing aid, but is generally in good health. Baltasar Garzon, one of two Spanish magistrates handling probes into human rights violations in Chile and Argentina, filed a request to question Pinochet on Wednesday, a day after another judge, Manuel Garcia Castellon, filed a similar petition. Castellon's probe into murder, torture and disappearances in Chile during Pinochet's regime began in 1996. Garzon is also investigating the disappearance of hundreds of Spanish citizens in Argentina during the 1976-83 military dictatorships. Pinochet is implicated in Garzon's probe through his involvement in "Operation Condor," in which military regimes in Chile, Argentina and Uruguay coordinated anti-leftist campaigns Pinochet, the son of a customs clerk who ousted elected President Salvador Allende in a bloody 1973 coup, remained commander-in-chief of the Chilean army until March, when he was sworn in as a senator-for-life, a post established for him in a constitution drafted by his regime. While in power he also pushed through an amnesty covering crimes committed before 1978, when most of his human rights abuses allegedly took place. One official report says 3,197 political opponents died during his term and 1,102 people remain unaccounted for after being detained by his security agents. (scl-rb) </TEXT> (PROFILE (WS SL:BC-Britain-Pinochet, 5th Ld-Writethru; CT:i; (REG:EURO; (REG:BRIT; (REG:SCAN; (REG:MEST; (REG:AFRI; (REG:INDI; (REG:ENGL; (REG:ASIA; (LANG:ENGLISH;)))) </BODY> <TRAILER> AP-NY-10-17-98 1215EDT </TRAILER> </DOC> <DOC> <DOCNO>



... and transform it into a graph of relations





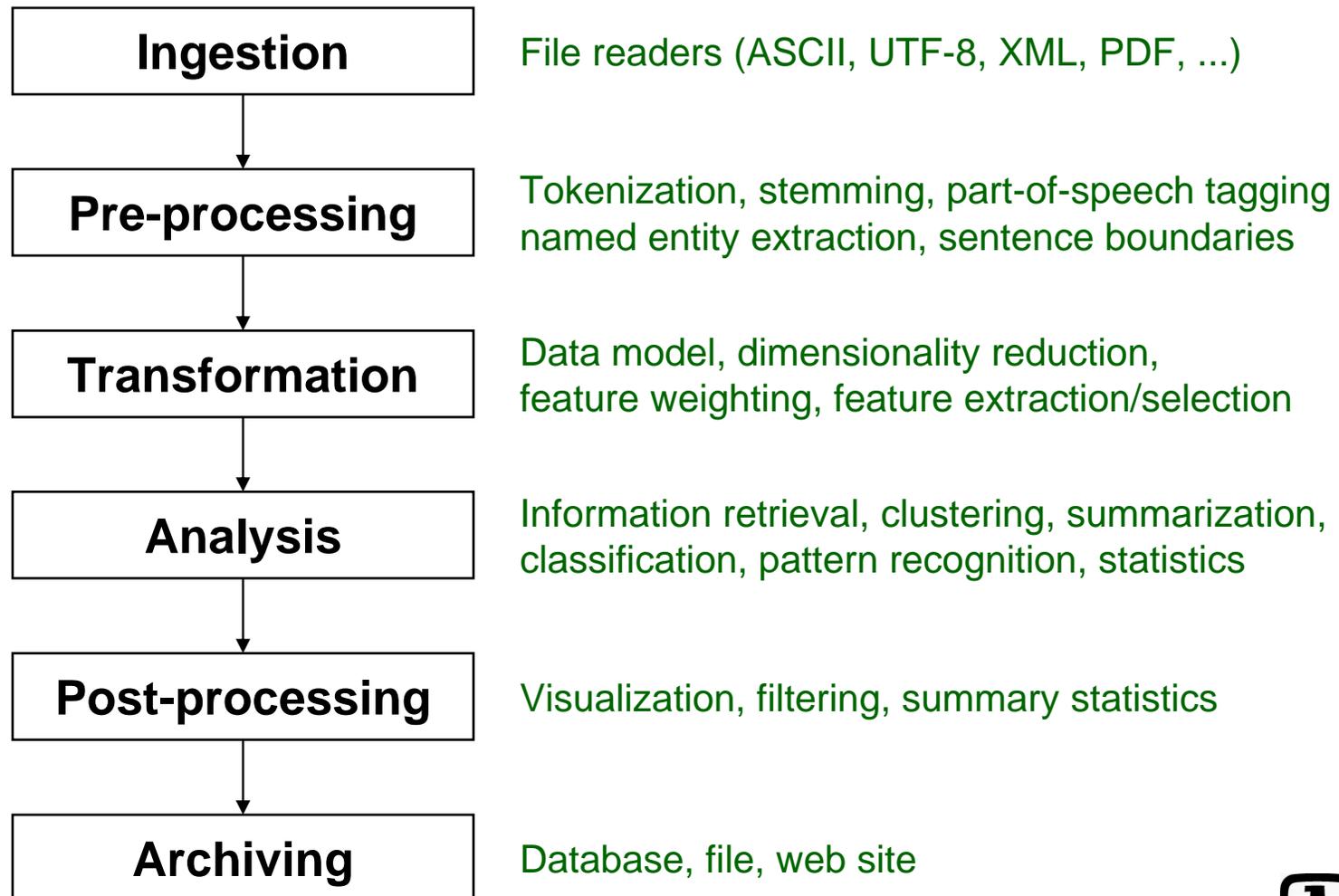


Main Goals for Text Analysis Research

- **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
- **Analyze large-scale data sets**
 - Small data: 700K documents with 600K features
- **National Security Applications**



Text Analysis Pipeline





Vector Space Model

- **Vector Space Model for Text**

- Terms (features): $t \in \mathbb{R}^m$
- Documents (objects): $d \in \mathbb{R}^n$
- 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 \times Document Matrix Scaling: $a_{ij} = \tau_{ij} \cdot \gamma_i \cdot \delta_j$

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

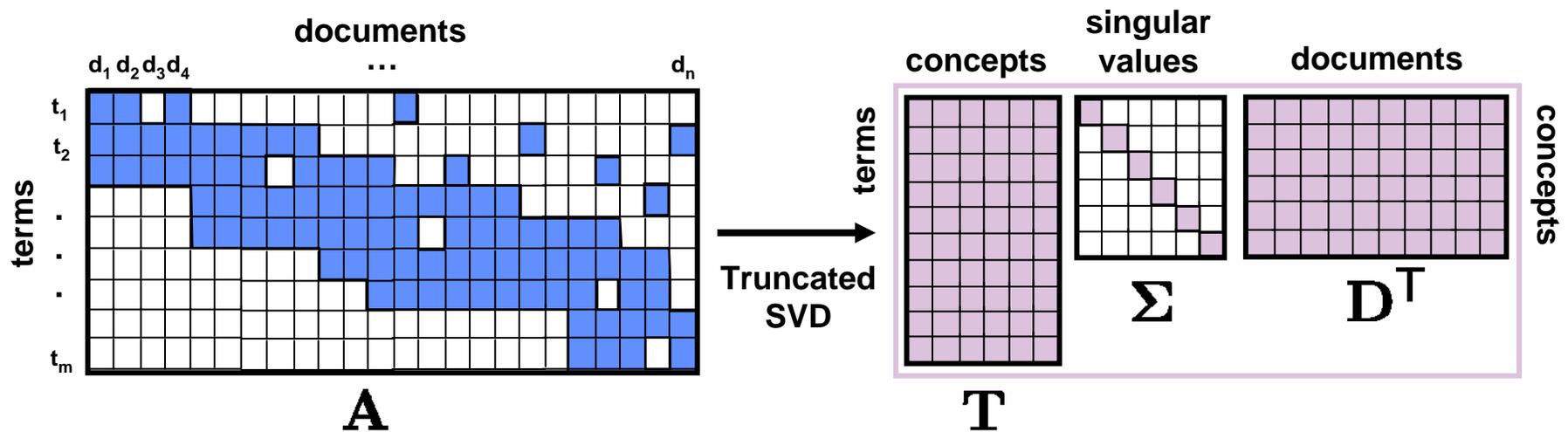


More about Features

- **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:** $A = T\Sigma D^T$
- **Truncated SVD:** $A \approx A_k = T_k \Sigma_k D_k^T = \sum_{r=1}^k \sigma_r t_r d_r^T$
- **Query scores (query as new “doc”):** $q^T A$
- **LSA Ranking:** $q^T A_k$

[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

normalization only

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

rank-2 approximation

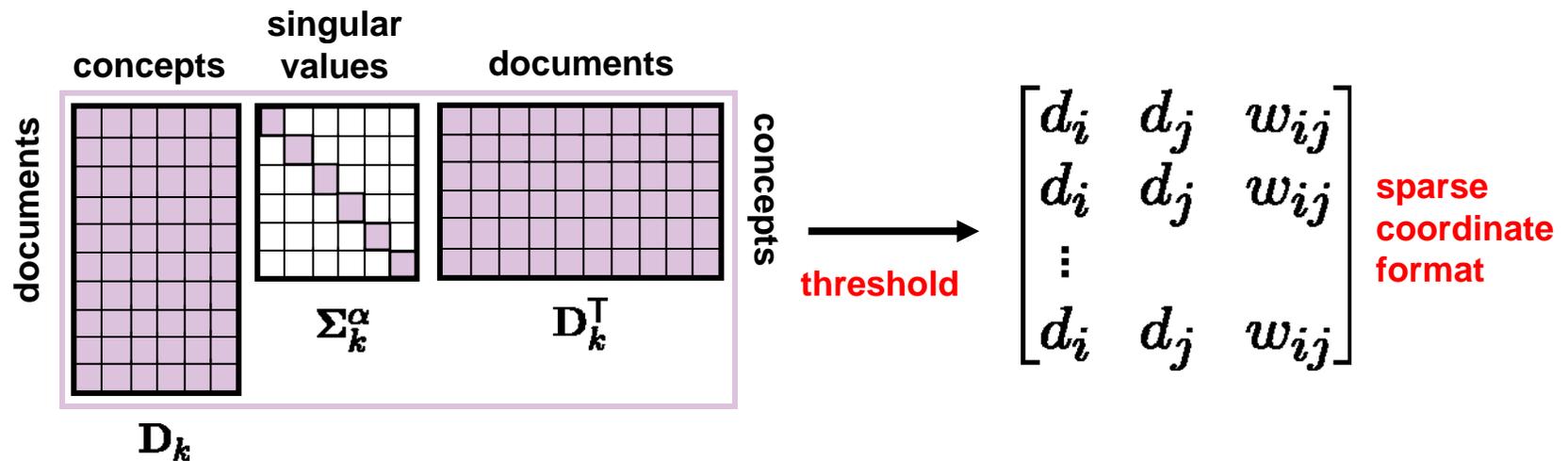
	A_2				
	d_1	d_2	d_3	d_4	
<i>hurricane</i>	.78	.78	-.11	.11	
<i>earthquake</i>	-.03	.02	.96	.92	
<i>catastrophe</i>	.59	.60	.15	.30	
$q^T A_2$.78	.78	-	.11

captures link to doc 4



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

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

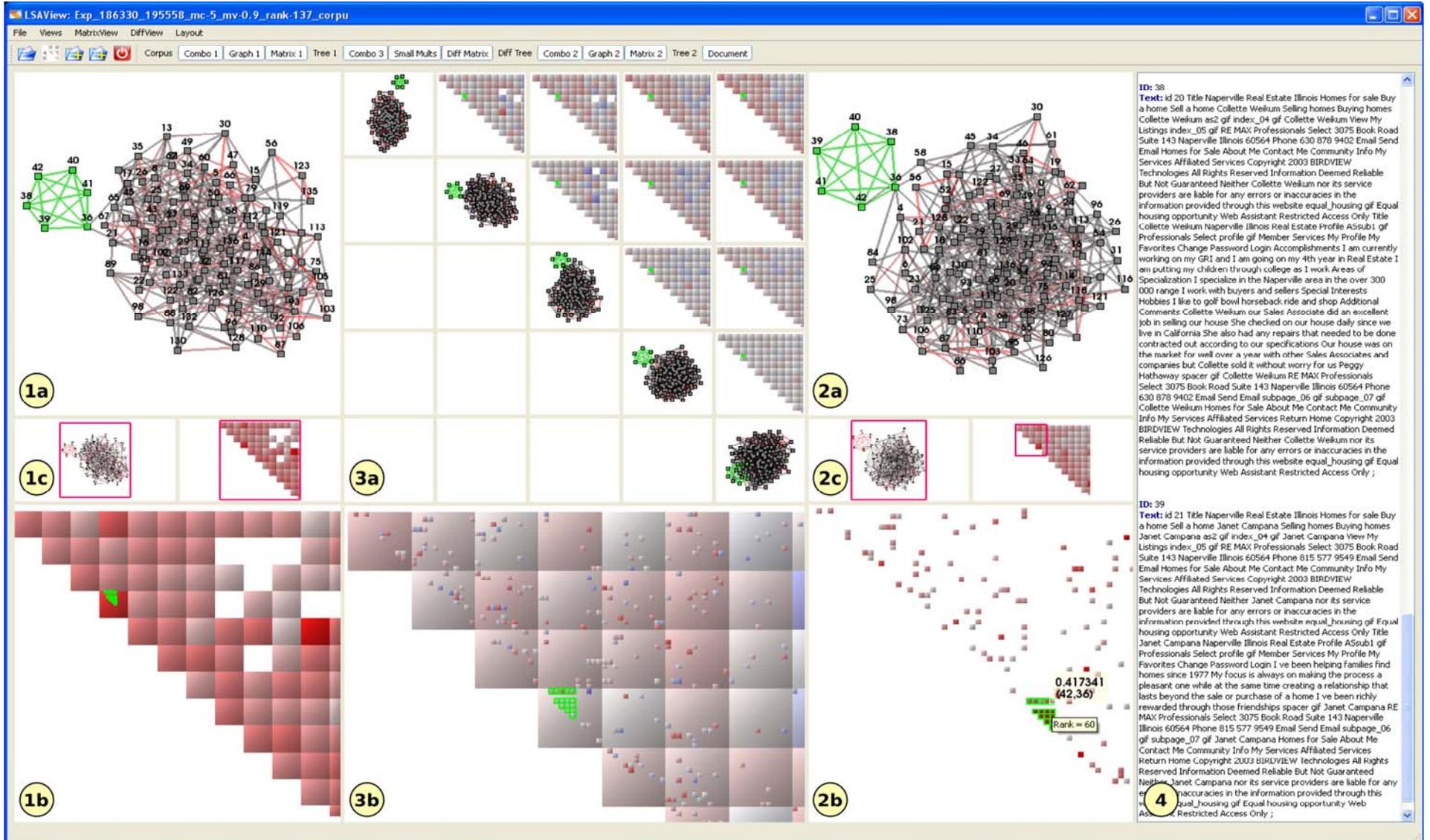
- Two graphs: two-sample t statistic

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

↑
Edges from graph 1

↑
Edges from graph 2

LSAView

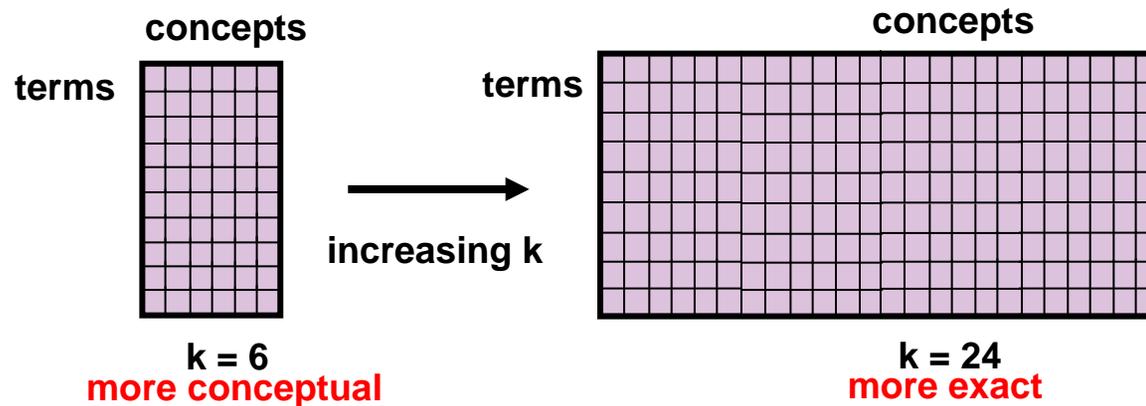




LSA: Rank Selection

- **Conceptual searching**

- rank(k) \uparrow : more exact data similarities
- rank(k) \downarrow : more conceptual data similarities
- Compute larger rank, then use smaller rank



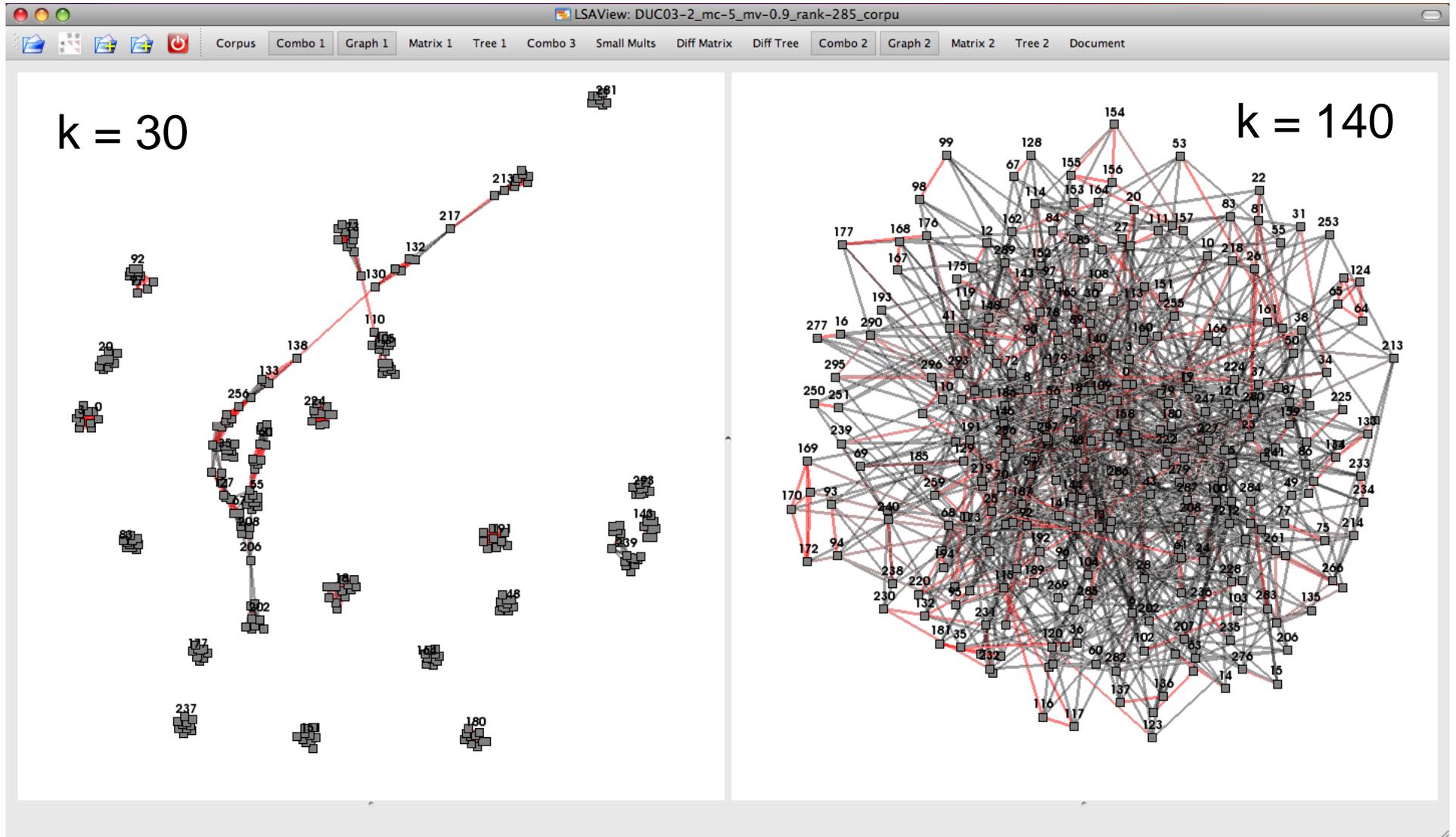
- **Determining useful values for rank**

- 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.]

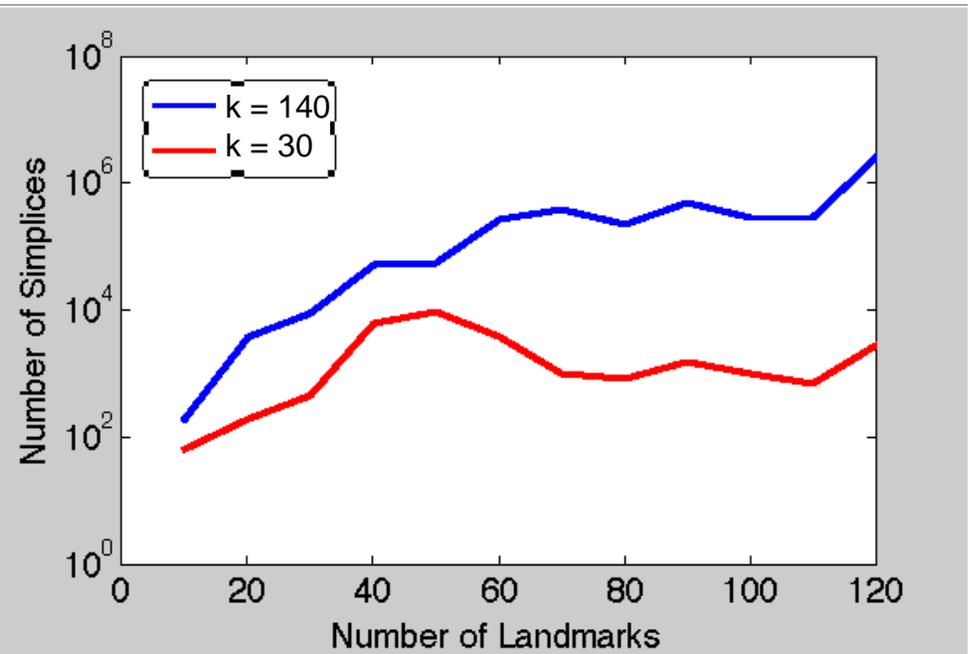
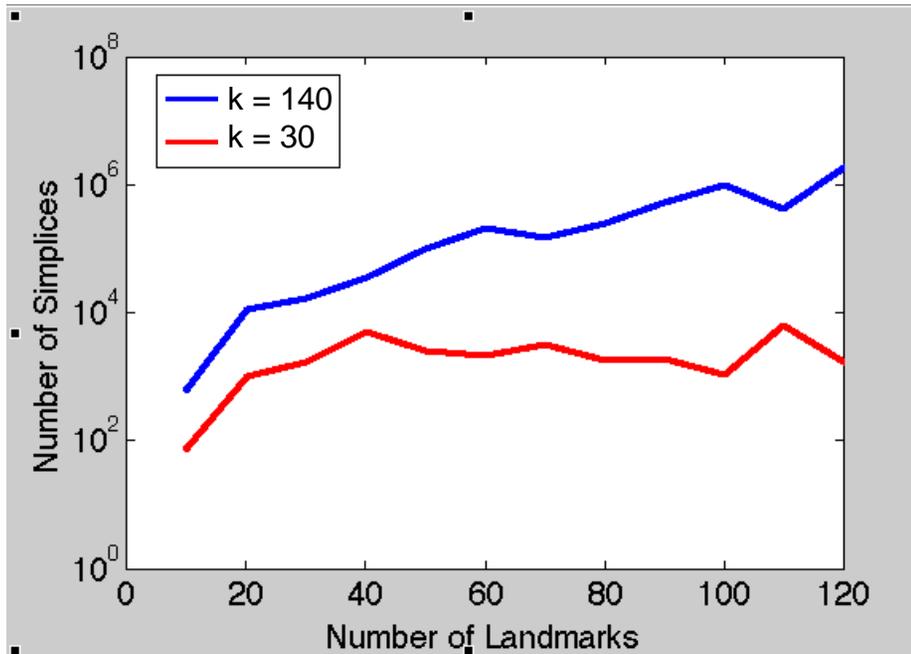


LSAView: Rank Selection



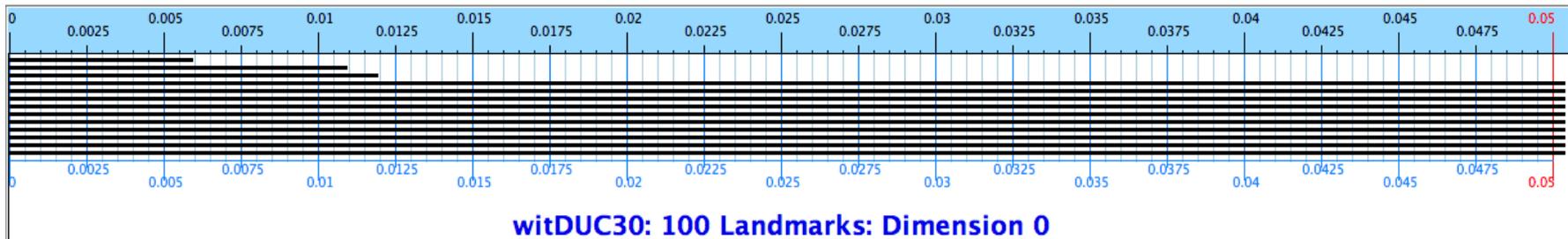
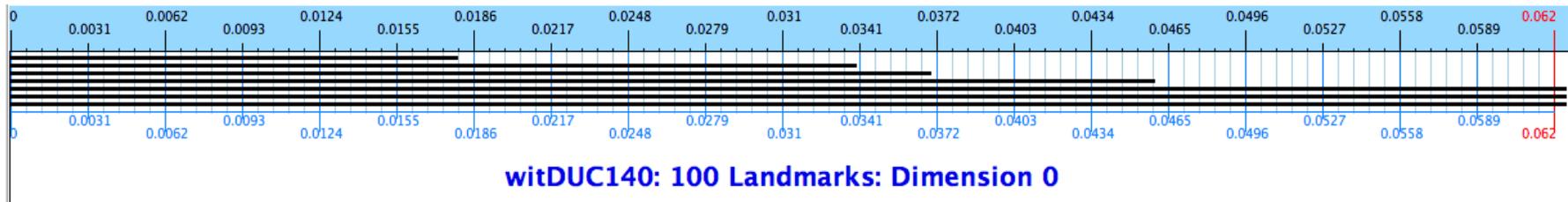


JPlex: Number of Simplices



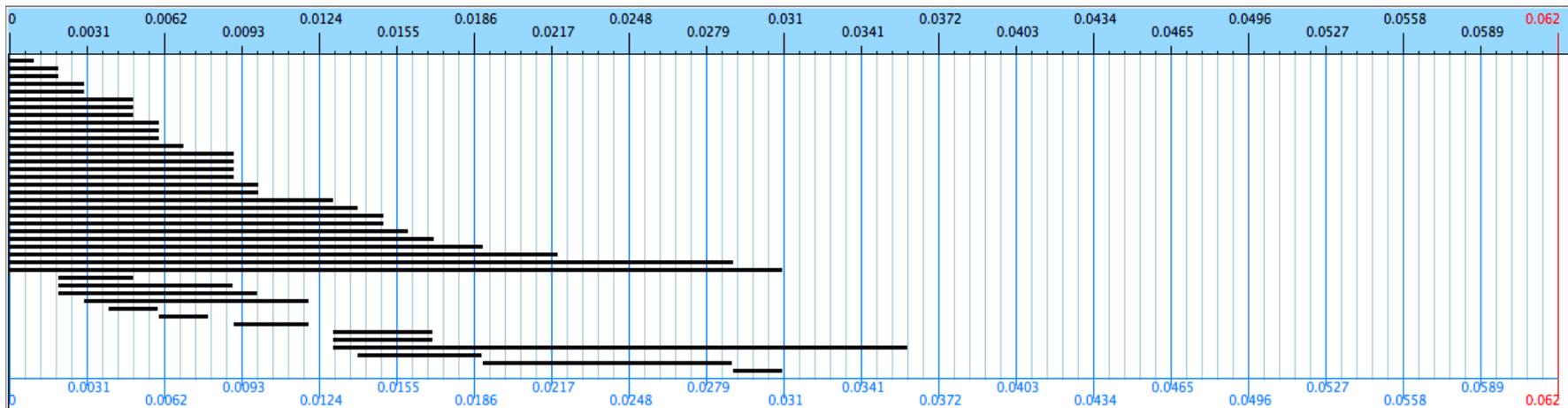


JPlex: Witness Stream

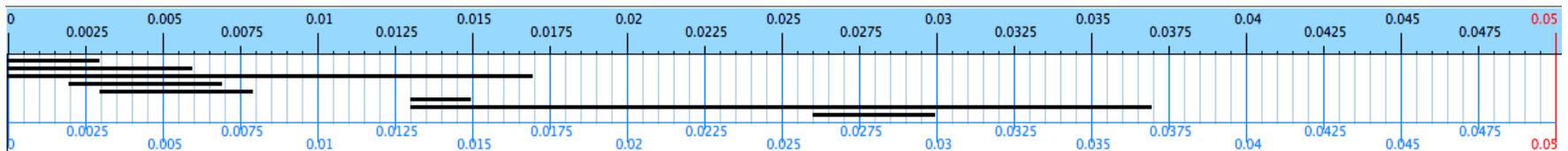




JPlex: Witness Stream



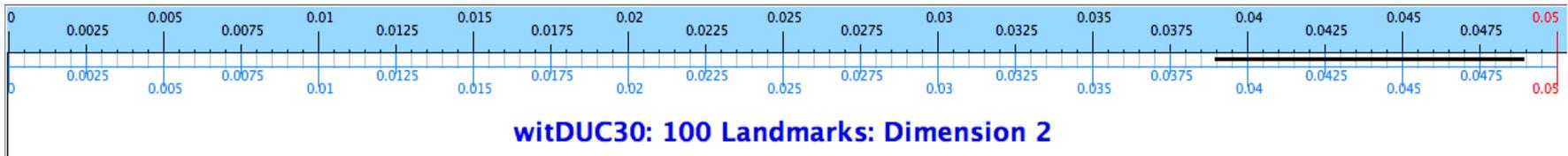
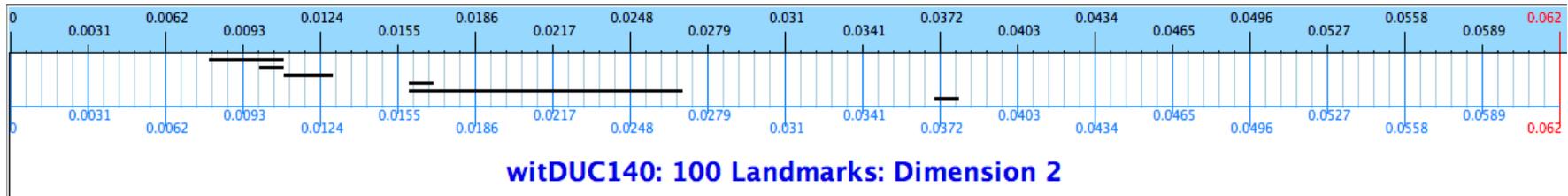
witDUC140: 100 Landmarks: Dimension 1



witDUC30: 100 Landmarks: Dimension 1

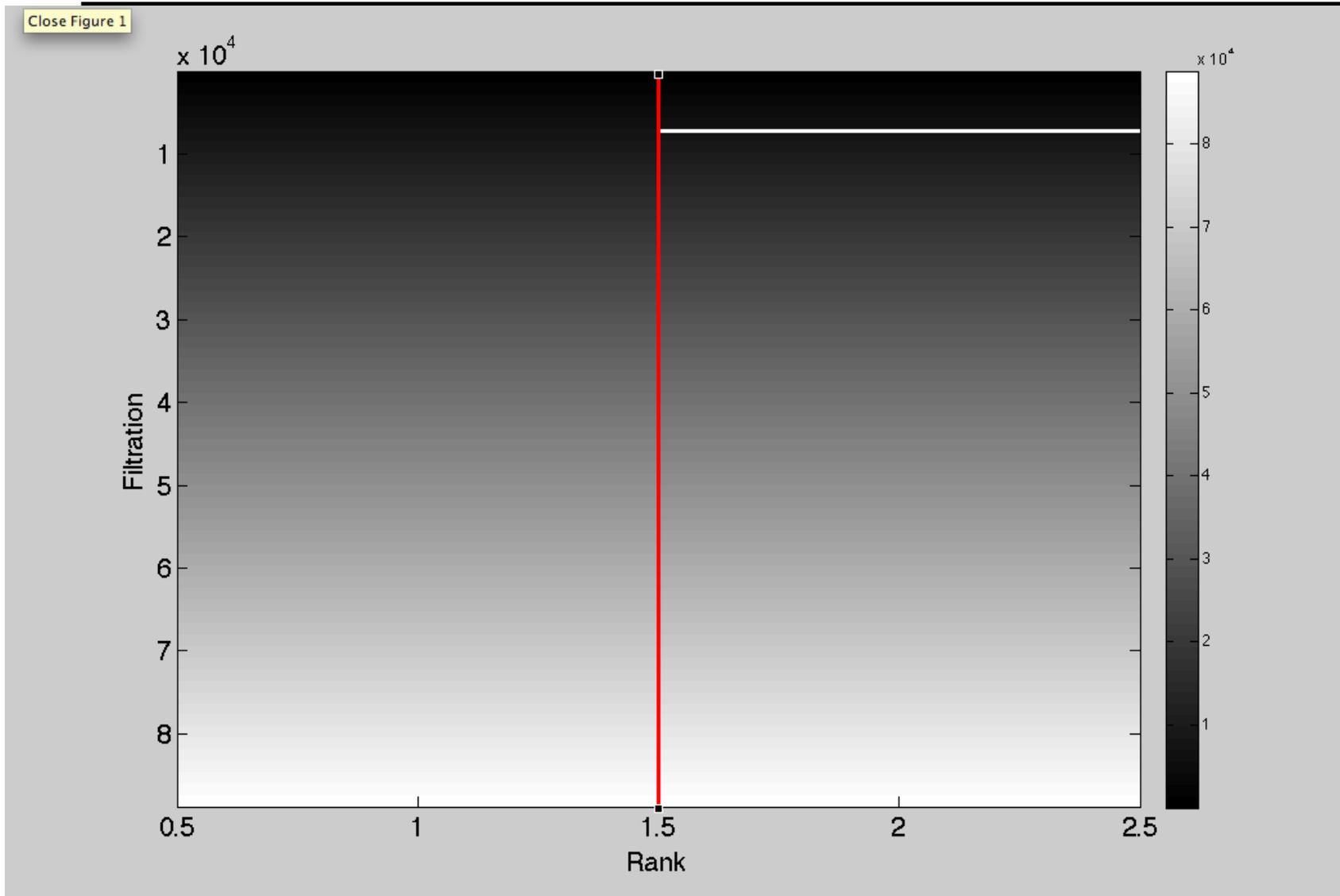


JPlex: Witness Stream



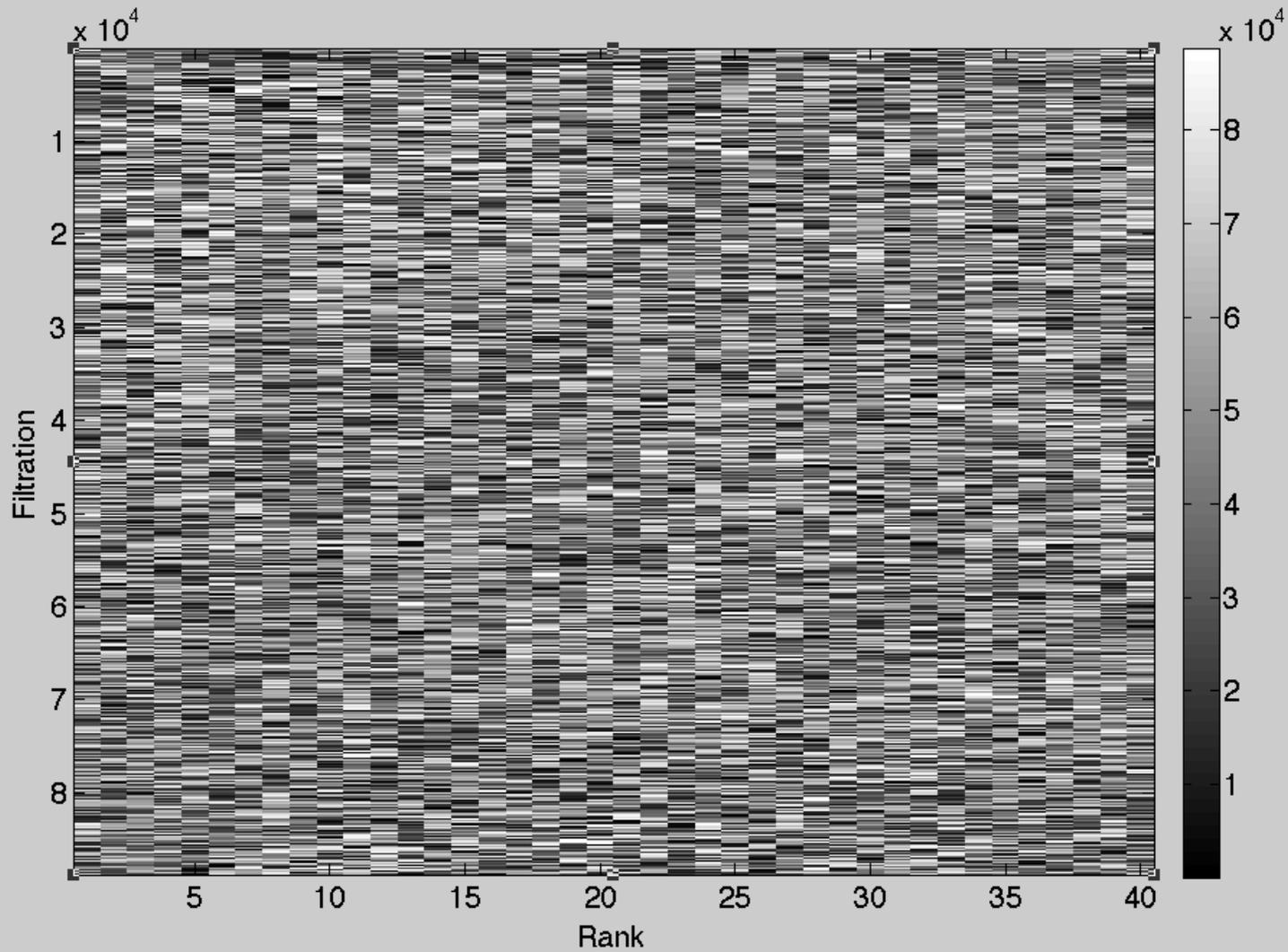


JPlex: Filtration versus Rank





JPlex: Filtration versus Rank

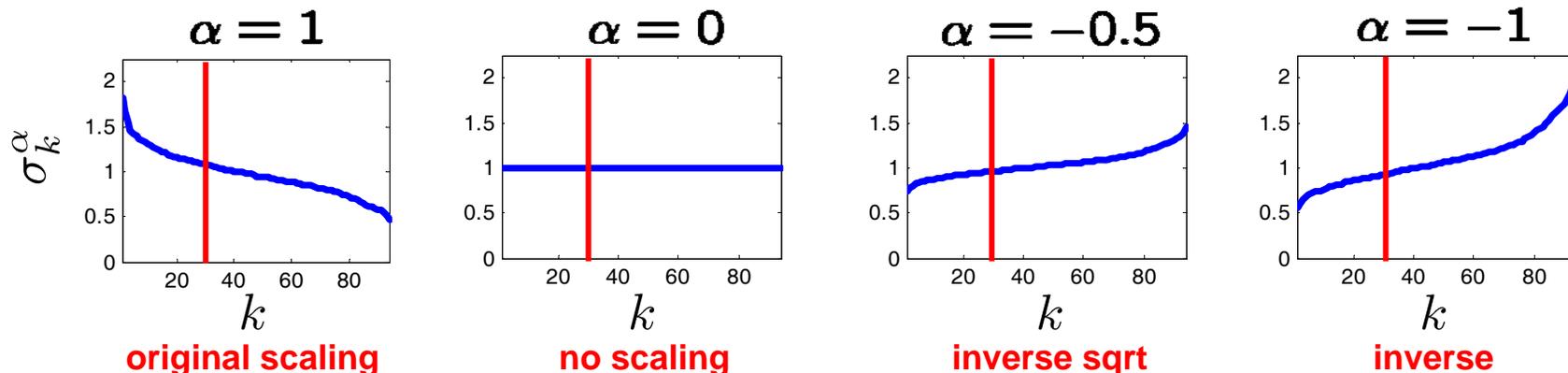




LSA: Singular Value (Re)Scaling

- **Document similarities:** $\mathbf{A}_k^\top \mathbf{A}_k = \mathbf{D}_k \Sigma_k^2 \mathbf{D}_k^\top$
- **Inner product view:** $(\mathbf{D}_k \Sigma_k) (\mathbf{D}_k \Sigma_k)^\top$
- **Scaled inner product view:** $(\mathbf{D}_k \Sigma_k^\alpha) (\mathbf{D}_k \Sigma_k^\alpha)^\top$

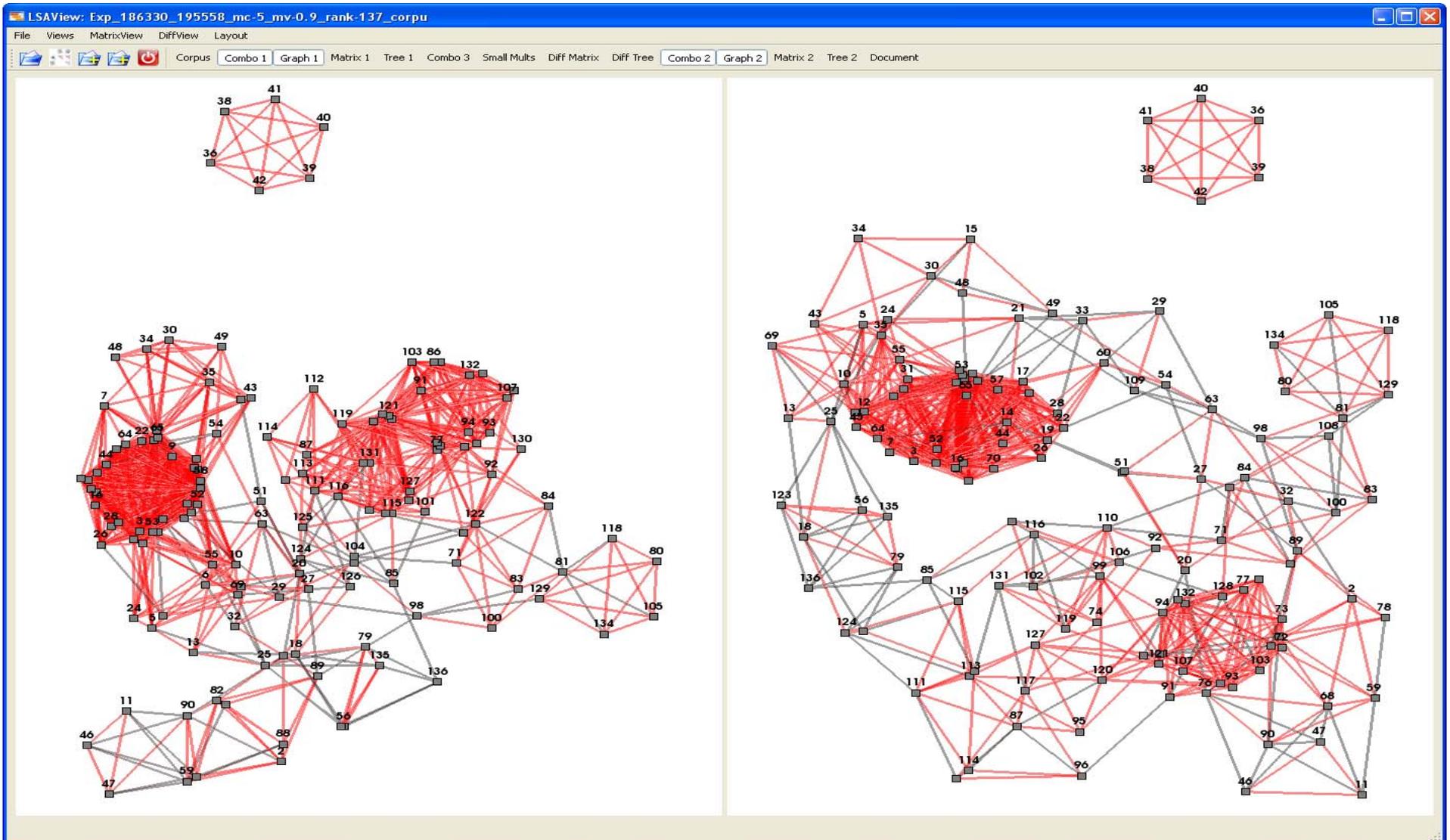
What is the best scaling for document similarity graph generation? [Data: 97 documents, 335 terms]



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

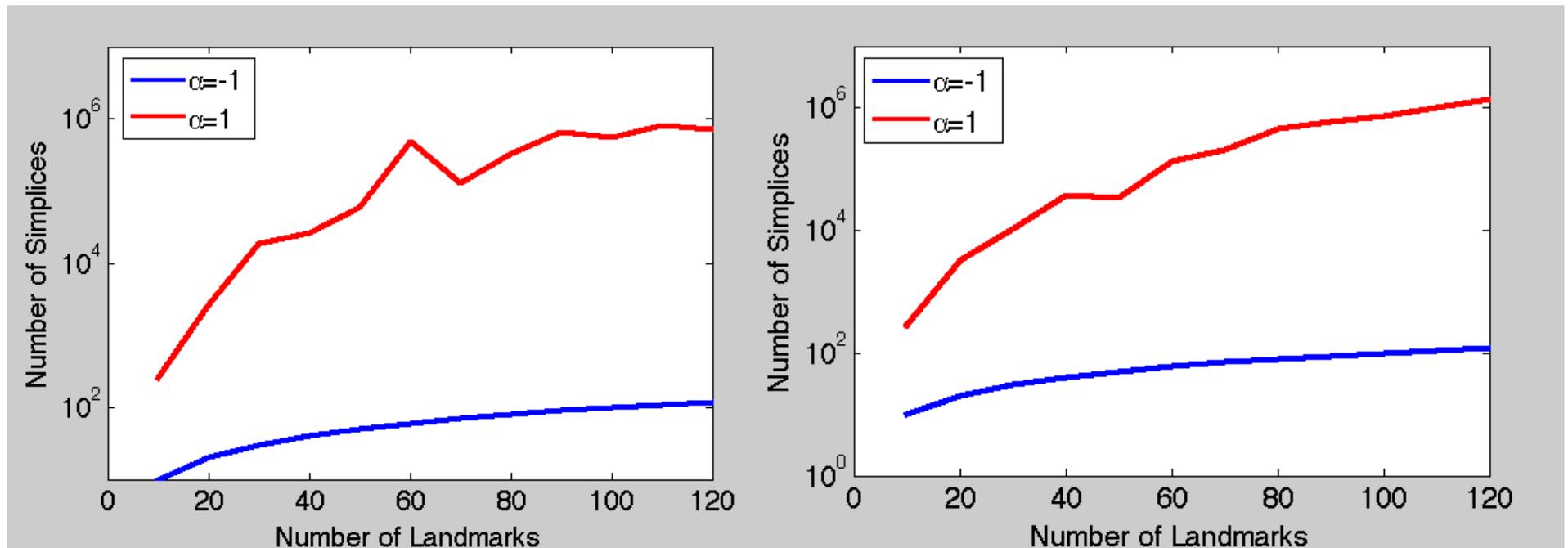


LSAView: Singular Value Scaling



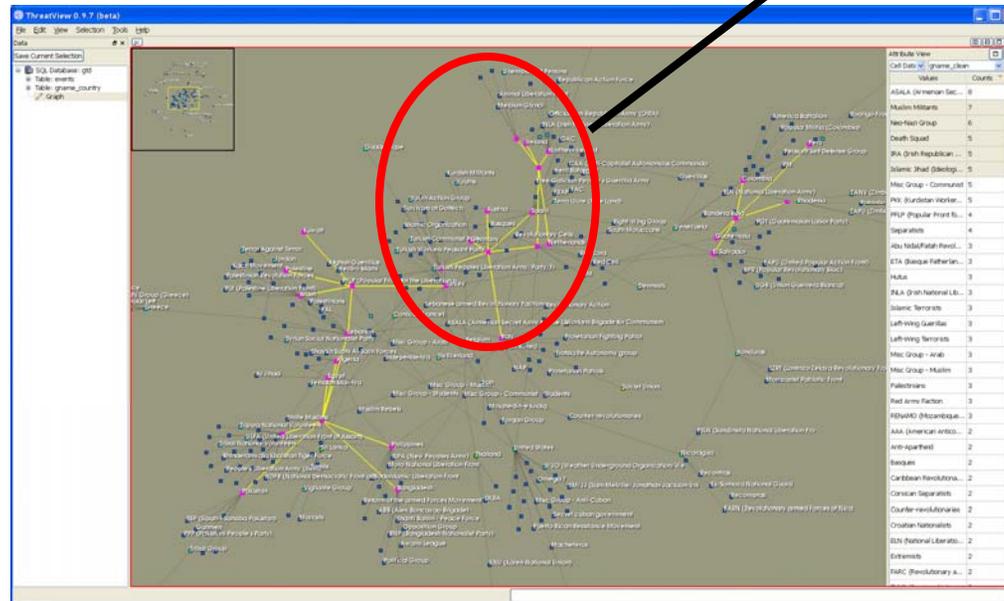
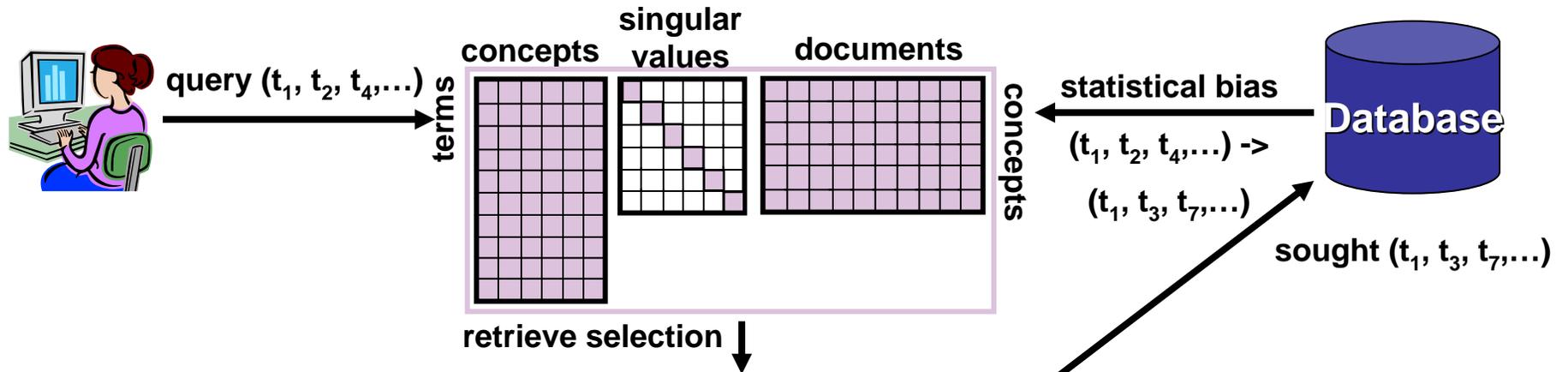


JPlex: Number of Simplices





LSA: Relevance Feedback





Alternative Approaches for LSA

- **SVD alternatives**

- Semi-discrete decomposition (SDD)
- Non-negative matrix factorization (NMF)
- Matrix subset selection (e.g., CUR)

- **Probabilistic modeling**

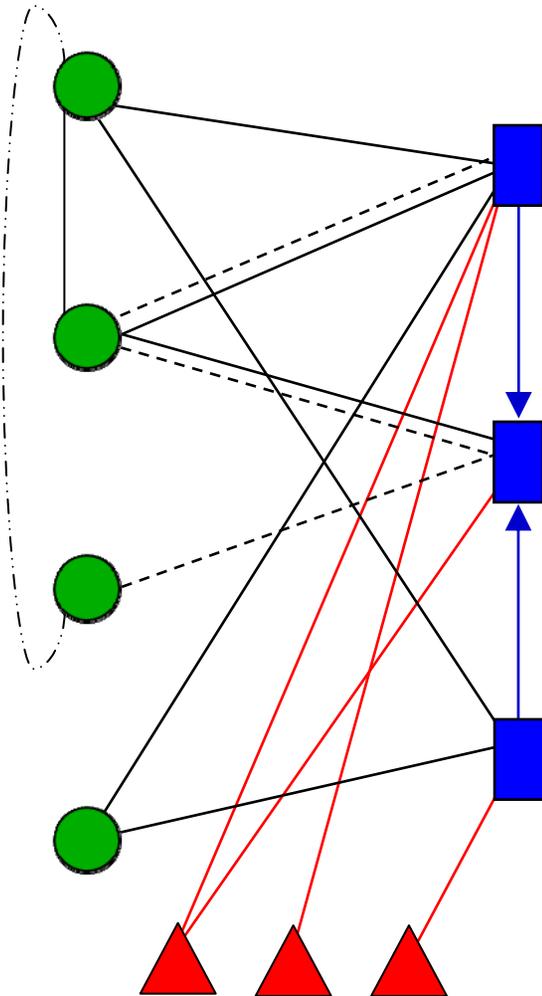
- Probabilistic LSA
- Latent Dirichlet allocation (LDA)

- **Multway modeling, semantic graphs**

- Examples: term-document-author, term-document-time
- Data is modeled as a multidimensional array (tensor)
- Tensor decompositions
 - PARAFAC, Tucker, DEDICOM, ...



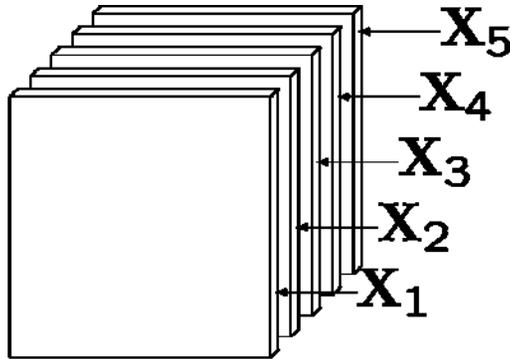
Semantic Graphs



- **Nodes (one type) connected by multiple types of links**
 - Node x Node x Connection
- **Two types of nodes connected by multiple types of links**
 - Node A x Node B x Connection
- **Multiple types of nodes connected by multiple types of links**
 - Node A x Node B x Node C x Connection
 - Directed and undirected links



Semantic Graph for Bibliometric Analysis



Slice (k)	Description	Nonzeros	$\sum_i \sum_j x_{ijk}$
1	Abstract Similarity	28476	7695.28
2	Title Similarity	120236	33285.79
3	Keyword Similarity	115412	16201.85
4	Author Similarity	16460	8027.46
5	Citation	2659	5318.00

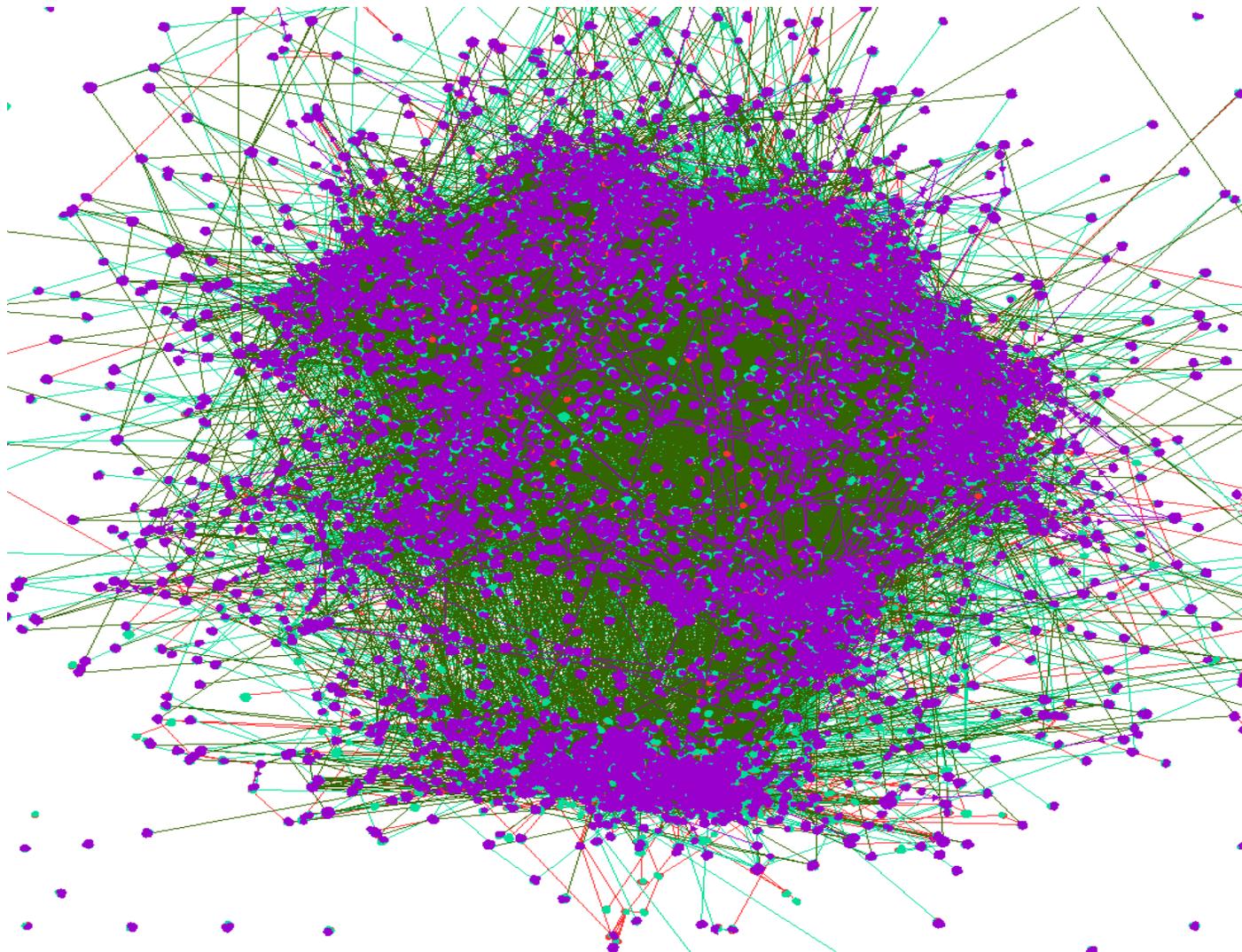
Frontal Slices \mathbf{X}_k

- $\mathbf{X}_1 = \mathbf{T}^T \mathbf{T}$ where $t_{ij} = f_{ij} \log_2(N/N_i)$ for terms in the **abstracts**
 - f_{ij} is the frequency of term i in document j
 - N_i is the number of documents that term i appears in
- $\mathbf{X}_2 = \mathbf{T}^T \mathbf{T}$ for terms in the **titles**
- $\mathbf{X}_3 = \mathbf{T}^T \mathbf{T}$ for terms in the author-supplied **keywords**
- $\mathbf{X}_4 = \mathbf{W}^T \mathbf{W}$ where $w_{ij} = \begin{cases} 1/\sqrt{M_j} & \text{if author } i \text{ wrote document } j \\ 0 & \text{otherwise,} \end{cases}$
- $x_{ij5} = \begin{cases} 2 & \text{if document } i \text{ cites document } j \\ 0 & \text{otherwise.} \end{cases}$

not symmetric

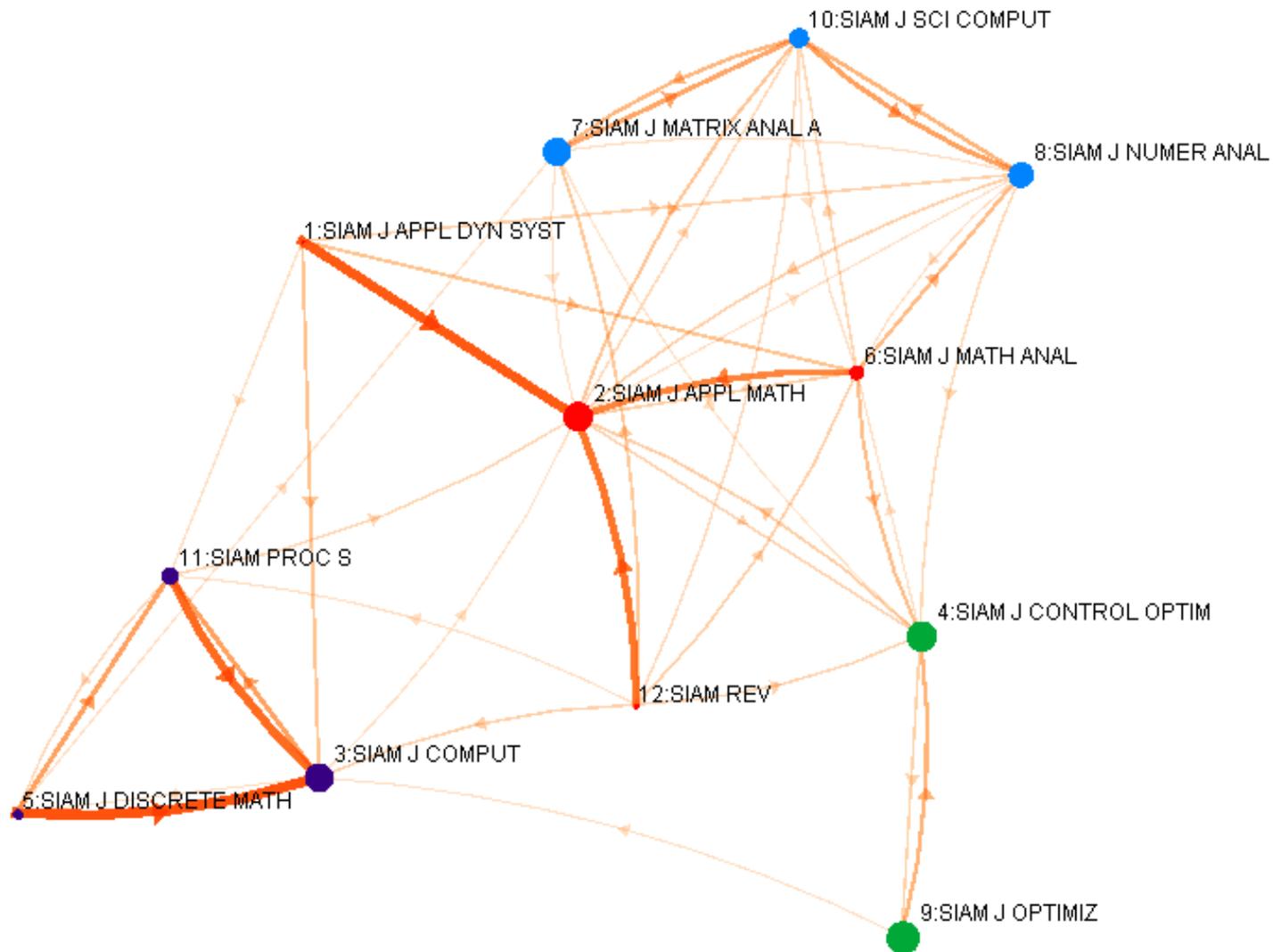


LSA Applied to Weighted Average of Slices





Journal Relationships using PARAFAC Tensor Decompositions





Research questions

- **Is there useful structure in the graphs generated using LSA?**
- **Can we associate topology and analyst knowledge?**
- **Can understanding of persistent homology lead to improved algorithms for knowledge discovery?**
- **How sensitive are topological structures with respect to ...**
 - LSA parameters? Data outliers? Data noise? Changes over time?
- **How do we communicate topological sensitivities to analysts?**
- **Can we compute persistent homology for dynamic data?**
- **Can we compute persistent homology for streaming data?**
- **What does structural persistence mean for semantic graphs and how can it be computed?**
- **Can we handle large-scale data sets?**



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>