Recursive Bipartite Spectral Clustering for Document Categorization

Jeff Solka[1,2], Avory Bryant[1],
and Edward J. Wegman[3]

1 – NSWCDD Code B10
2 – SCS GMU
3 – Department of Applied and Engineering Statistics GMU

Army Conference on Applied
Statistics, Atlanta, 2004

Agenda

- Our treatise.
- Our datasets.
- Our features.
- Mathematical background.
- Results on the Science News data.
- Results on the ONR ILIR data.

In a Nutshell?

- What are we trying to do?
  - Develop a semi-automated system to facilitate the text data mining
    - Discovery of articles from disparate corpora that may contain subtle relationships.
    - Discovery of interesting clusters of articles.
- What is our approach predicated on?
  - The synthesis of methodologies from statistics, mathematics and visualization.
  - Use of minimal spanning trees and spectral graph theory as technological enablers.
- What are the test cases?
  - Roughly 1200 Science News abstracts that have been precategorized into 8 categories.
  - Roughly 343 Office of Naval Research In-house Laboratory Independent Research documents.

The Science News Corpus

- 1117 documents from 1994–2002.
- Obtained from the SN website on December 2002 19,2002 using wget.
- Each article ranges from 1/2 a page to roughly a page in length.
- The corpus html/xml code was subsequently parsed into straight text.
- The corpus was read through and categorized into 8 categories.

The Science News Corpus Breakdown

- Anthropology and Archeology (48).
- Astronomy and Space Sciences (124).
- Behavior (88).
- Earth and Environmental Sciences (164).
- Life Sciences (174).
- Mathematics and Computers (65) .
- Medical Sciences (310) .
- Physical Sciences and Technology (144)

The Office of Naval Research (ONR) In-House Laboratory Independent Research (ILIR) Corpus

- 343 Documents
- Obtained from ONR
- Support on-line querying and mining of their ILIR database

ILIR Corpus Breakdown

- Advanced Naval Materials (82)
- Air Platforms and Systems (23)
- Electronics
- Expeditionary/USMC
- Human Performance /Factors (49)
- Information Technology and Operations (18)
- Manufacturing Technologies (21)
- Medical S&T (19)
- Naval & Joint Experimentation
- Naval Research Enterprise Programs
- Operational Environments (27)
- RF Sensing, Surveil, & Countermeasures (27)
- Sea Platform and Systems (38)
- Strike Weapons
- Undersea Weapons
- USW-ASW (5)
- USW-MIW (17)
- Visible and IR Sensing, Surveil & Countermeasures (17)

Denoising and Stemming

- These steps are performed prior to subsequent feature extraction steps.

- Various approaches to denoising were used
  - Simplest consists of removal of all words that appear on a stopper or noise word list.
  - the, a, an, …
  - More on this later

- Stemming transforms a given word into its base
  - walking → walk
  - walked → walk

- Denoising is implemented within the current system stemming is implemented in some versions but is not in others

Document Features

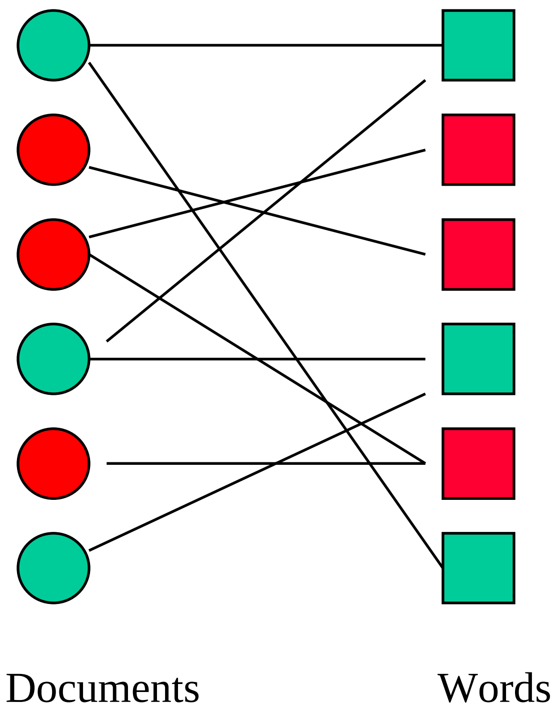
- Bigram Proximity Matrices ala Martinez 2002
  - Angel Martinez, "A Framework for the Representation of Semantics," *Ph.D Dissertation under the direction of Edward Wegman*, October 2002.

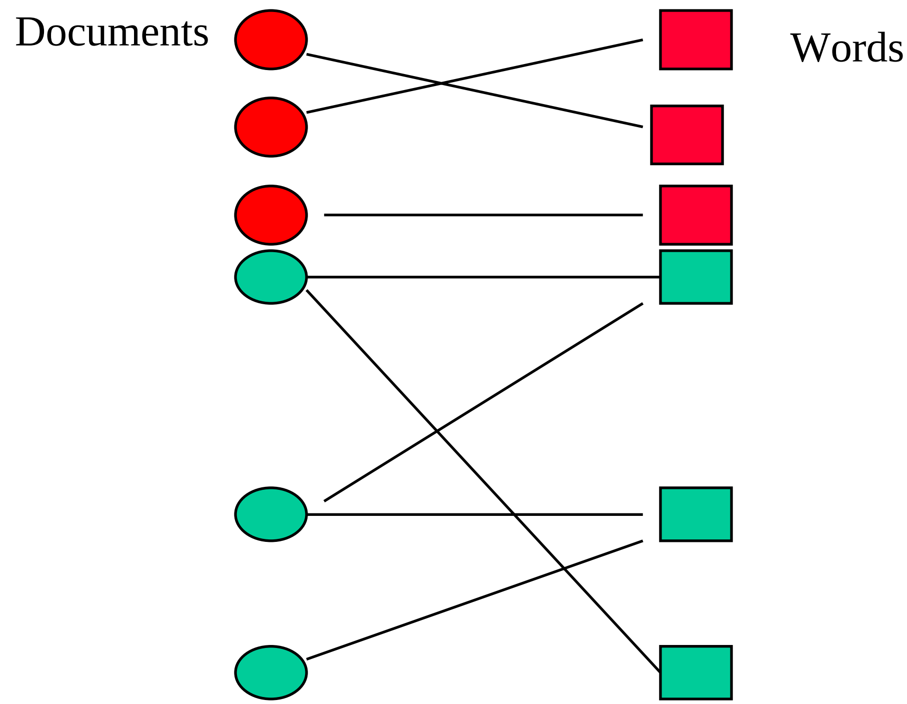
- Mutual Information Features ala Lin 2002
  - Patrick Pantel and Dekang Lin, "Discovery word senses from text," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pgs. 613-619, 2002.

- "Normalized" term document matrices ala Dhillon 2001
  - Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," UT CS Technical Report # TR 2001-05.

Bipartite Spectral Based Clustering

- Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," KDD 2001.





Cut measures the sum of the crossing between vertex set V_1 and vertex set V_2.

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}$$

The Graph Theoretic Formulation

Our Graph $G = (\mathcal{V}, E)$

Vertex Set $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$

Edge Set $\{i, j\}$

Edge Weights E_{ij}

Adjacency Matrix

$$M = \begin{cases} E_{ij}, & \text{if there is an edge } \{i, j\}, \\ 0, & \text{otherwise.} \end{cases}$$

The cut between two subsets of vertices.

$$\operatorname{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

The cut between k subsets of vertices.

$$\operatorname{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k) = \sum_{i<j} \operatorname{cut}(\mathcal{V}_i, \mathcal{V}_j)$$

The Document Word Bipartite Model

Our graph consisting of a vertex set consisting of documents and words along with associated edges.

$G = (\mathcal{D}, \mathcal{W}, E)$

The word vertices.

$\mathcal{W} = \{w_1, w_2, \ldots, w_m\}$

The document vertices.

$\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$

One strategy for setting the edge weights.

$$E_{ij} = t_{ij} \times \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}_i|}\right)$$

where t_{ij} is the number of times word w_i occurs in document d_j, $|\mathcal{D}| = n$ is the total number of documents and $|\mathcal{D}_i|$ is the number of documents that contain word w_i.

$$M = \begin{bmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{bmatrix}$$

Adjacency Matrix – $A_{ij} = E_{ij}$, 0's reflect no word to word or document to document connections

$$\mathrm{cut}(\mathcal{W}_1 \cup \mathcal{D}_1, \mathcal{W}_2 \cup \mathcal{D}_2, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k)$$

Our Clustering Criteria

Corpus Dependent Stop Word Removal

- Stop words are removed.
- Words occurring in less than 0.2% of the documents are removed.
- Words occurring in greater than 15% of the documents are removed.
- N. B.
  - The methodology has been shown successful even if stopper words are not removed.
  - 0.2% and 15% are user "tunable" parameters.

Graph Partitioning

Given a graph $G = (\mathcal{V}, E)$, the classical graph bipartitioning or bisection problem is to find nearly equally-sized vertex subsets $\mathcal{V}_1^*, \mathcal{V}_2^*$ of $\mathcal{V}$ such that

$$\mathrm{cut}(\mathcal{V}_1^*, \mathcal{V}_2^*) = \min_{\mathcal{V}_1, \mathcal{V}_2} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2).$$

The graph partitioning problem is known to be NP-complete.

We will follow Dhillon and use graph spectral methods to obtain an approximate solution based on a suitably formulated objective function.

Assuring An Equitable Partition – An Objective Function

$$W_{ij} = \begin{cases} \text{weight}(i), & i = j, \\ 0, & i \neq j. \end{cases}$$

The weight for a particular vertex.

$$\text{weight}(\mathcal{V}_l) = \sum_{i \in \mathcal{V}_l} \text{weight}(i) = \sum_{i \in \mathcal{V}_l} W_{ii}$$

The weight for a set of vertices.

A figure of merit function that helps assure near equal number of points in each cluster.

$$Q(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_1)} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\text{weight}(\mathcal{V}_2)}$$

One can think of this as being analogous to the ratio of between group and within group distances in our usual statistical clustering framework.

Choice of Vertex Weights

$$\text{weight}(i) = 1$$

$$\text{Ratio-cut}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_1|} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{|\mathcal{V}_2|}$$

$$\text{weight}(i) = \sum_k E_{ik}$$

Normalized cut.

$$\mathcal{N}(\mathcal{V}_1, \mathcal{V}_2) = \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_1} \sum_k E_{ik}} + \frac{\text{cut}(\mathcal{V}_1, \mathcal{V}_2)}{\sum_{i \in \mathcal{V}_2} \sum_k E_{ik}}$$

Algorithm Bipartition

$$D_1(i,i) = \sum_j A_{ij} \quad \text{(sum of edge-weights incident on word } i\text{)},$$
$$D_2(j,j) = \sum_i A_{ij} \quad \text{(sum of edge-weights incident on document } j\text{)}.$$

$$\boldsymbol{z}_2 = \begin{bmatrix} \boldsymbol{D_1}^{-1/2}\boldsymbol{u}_2 \\ \boldsymbol{D_2}^{-1/2}\boldsymbol{v}_2 \end{bmatrix} \quad (4.13)$$

Algorithm Bipartition

1. Given $\boldsymbol{A}$, form $\boldsymbol{A_n} = \boldsymbol{D_1}^{-1/2}\boldsymbol{A}\boldsymbol{D_2}^{-1/2}$.
2. Compute the second singular vectors of $\boldsymbol{A_n}$, $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ and form the vector $\boldsymbol{z}_2$ as in (4.13).
3. Run the k-means algorithm on the 1-dimensional data $\boldsymbol{z}_2$ to obtain the desired bipartitioning.

The singular vectors $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ of $\boldsymbol{A_n}$ give a real approximation to the discrete optimization problem of minimizing the normalized cut.

The Left and Right Singular Vectors

$$\boldsymbol{A}_n \boldsymbol{v}_2 = \sigma_2 \boldsymbol{u}_2, \qquad \boldsymbol{A}_n{}^T \boldsymbol{u}_2 = \sigma_2 \boldsymbol{v}_2, \tag{4.12}$$

$$\sigma_2 = 1 - \lambda_2$$

The right singular vector $\boldsymbol{v}_2$ will give us a bipartitioning of documents while the left singular vector $\boldsymbol{u}_2$ will give us a bipartitioning of the words. By examining the relations (4.12) it is clear that this solution agrees with our intuition that a partitioning of documents should induce a partitioning of words, while a partitioning of words should imply a partitioning of documents.

The curious fact is that the obtained transformation allows one to map the documents and words into the same one-dimensional space.

Algorithm Multipartition(k)

$$Z = \begin{bmatrix} {D_1}^{-1/2} U \\ {D_2}^{-1/2} V \end{bmatrix} \quad (4.14)$$

$$U = [u_2, u_3, \ldots, u_{\ell+1}], \text{ and } V = [v_2, v_3, \ldots, v_{\ell+1}], \quad \ell = \lceil \log_2 k \rceil$$

Algorithm Multipartition(k)

1. Given A, form $A_n = {D_1}^{-1/2} A {D_2}^{-1/2}$.
2. Compute $\ell = \lceil \log_2 k \rceil$ singular vectors of A_n, $u_2, u_3, \ldots u_{\ell+1}$ and $v_2, v_3, \ldots v_{\ell+1}$ and form the matrix Z as in (4.14).
3. Run the k-means algorithm on the ℓ-dimensional data Z to obtain the desired k-way multipartitioning.

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - I

- Confusion Matrix Performance Measures
  - Inderjit S. Dhillon, "Co-clustering documents and words using Bipartite Spectral Graph Partitioning," KDD 2001.
  - Inderjit S. Dhillon, " Co-clustering documents and words using Bipartite Spectral Graph Partitioning," Ut CS Technical Report # TR 2001-05.
  - These were obtained using "mixtures" of MEDLINE (medical database), CISI (Institute of Scientific Information database), and CRANFIELD (document searching database) document sets along with YAHOO_K5 (Reuter News Articles from Yahoo where words are stemmed and heavily pruned) and YAHOO_K1 (Reuters News Articles from Yahoo: words are stemmed and only stop words are pruned)

How Do We Know That the Dhillon 2001 Strategy is Worthwhile - II

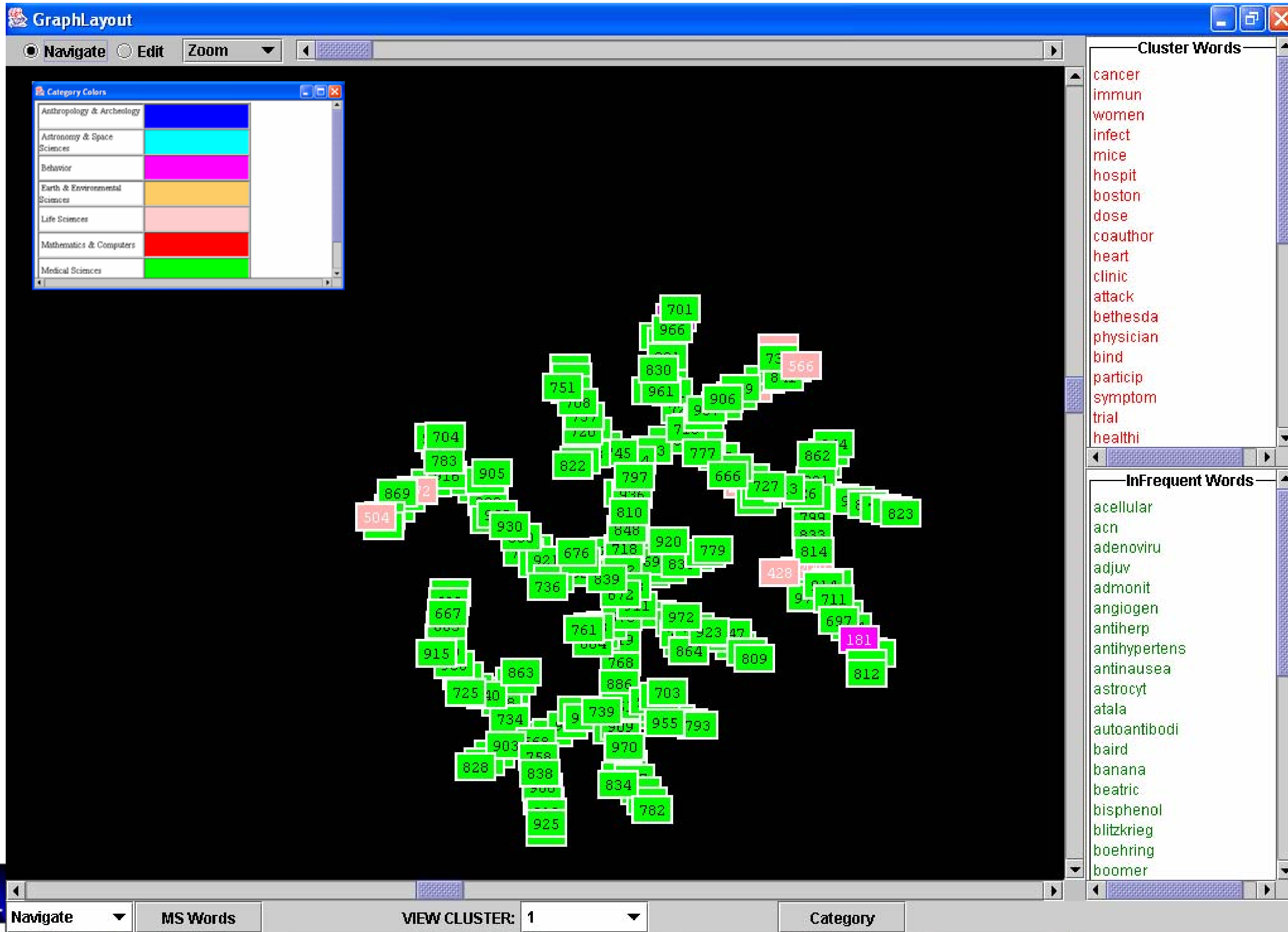
- Confusion matrix performance on the
  - Science News
  - ONR ILIR Data

- Theoretical results that insure us that the spectral based approach is a good approximation to solving the NP-compete problem.

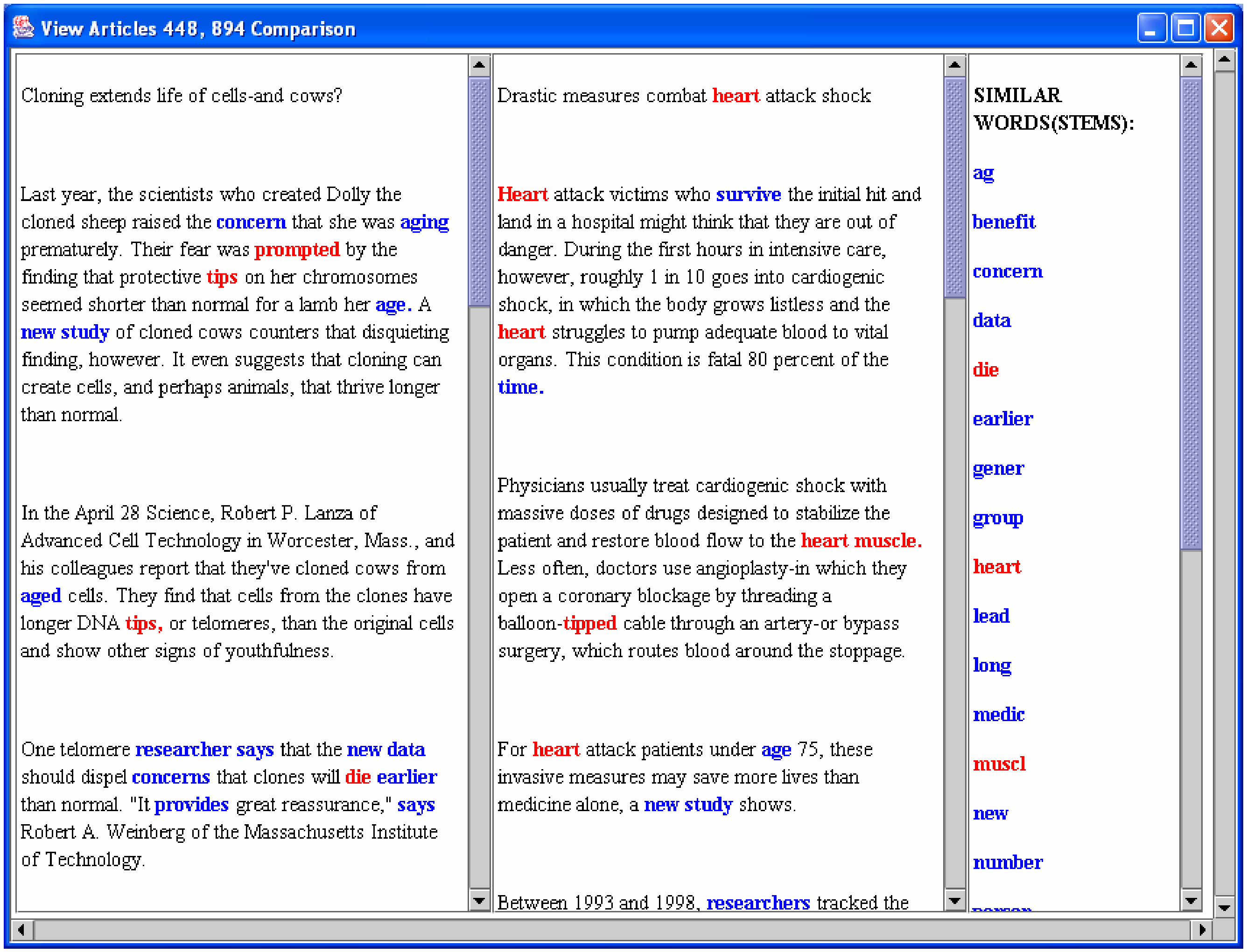
Recursive Bipartite Bipartition Methodology of Solka, Bryant and Wegman

- Alternative to the multipartition approach.
- Recursively use the bipartite bipartition methodology to obtain a multipartition of the data.
- Which cluster to split next is currently based on a simple mean distance of all observations to the centroid measure.
  - Certainly could be the subject of a more advanced statistical methodology.
- A visualization framework for exploration of the clusters (documents and words) and their associated concepts is provided.

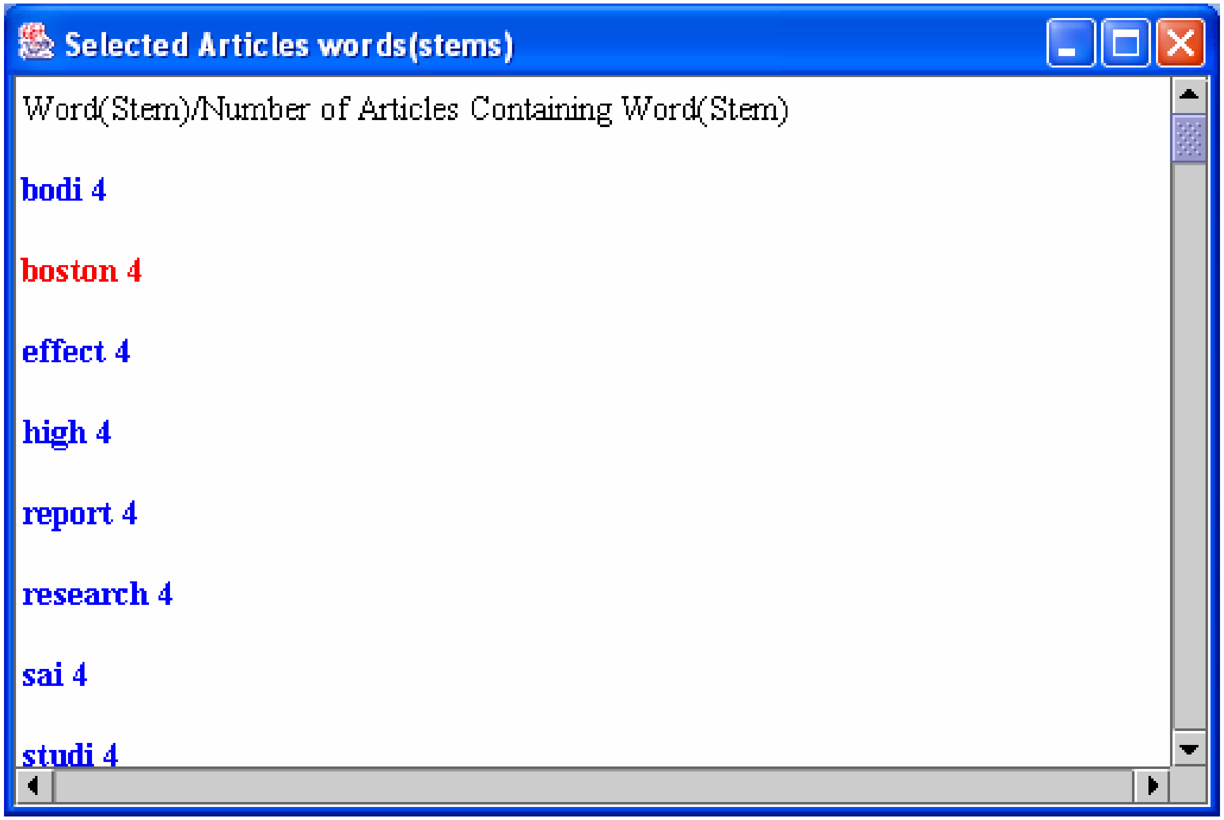
Visualization Framework - I



Visualization Framework – II (Comparison File for a Biology and Medical Sciences Article)



Visualization Framework - III (Multi-select Operation)



How Do We Measure the Quality of Our Clustering

- The clustering figure of merit is based on the ability of the methodology to match a set of user obtained categorizations.

- Deviations from these categorizations are measured via cluster:
  - Purity
  - Entropy

Purity

- A large value of purity indicates a good cluster.

$$P(D_j) = \frac{1}{n_j} \max_i (n_j^{(i)})$$

$n_j = |D_j|$ *and* n_j^i *is the number of documents in* D_j *that belong to class i*

Entropy

- A small value of entropy indicates a good cluster.

$$H(D_j) = -\frac{1}{\log(c)} \sum_{i=1}^{c} \frac{n_j^{(i)}}{n_j} \log\left(\frac{n_j^{(i)}}{n_j}\right)$$

Science News Spectral Clustering Results

Average Purity Per Observation

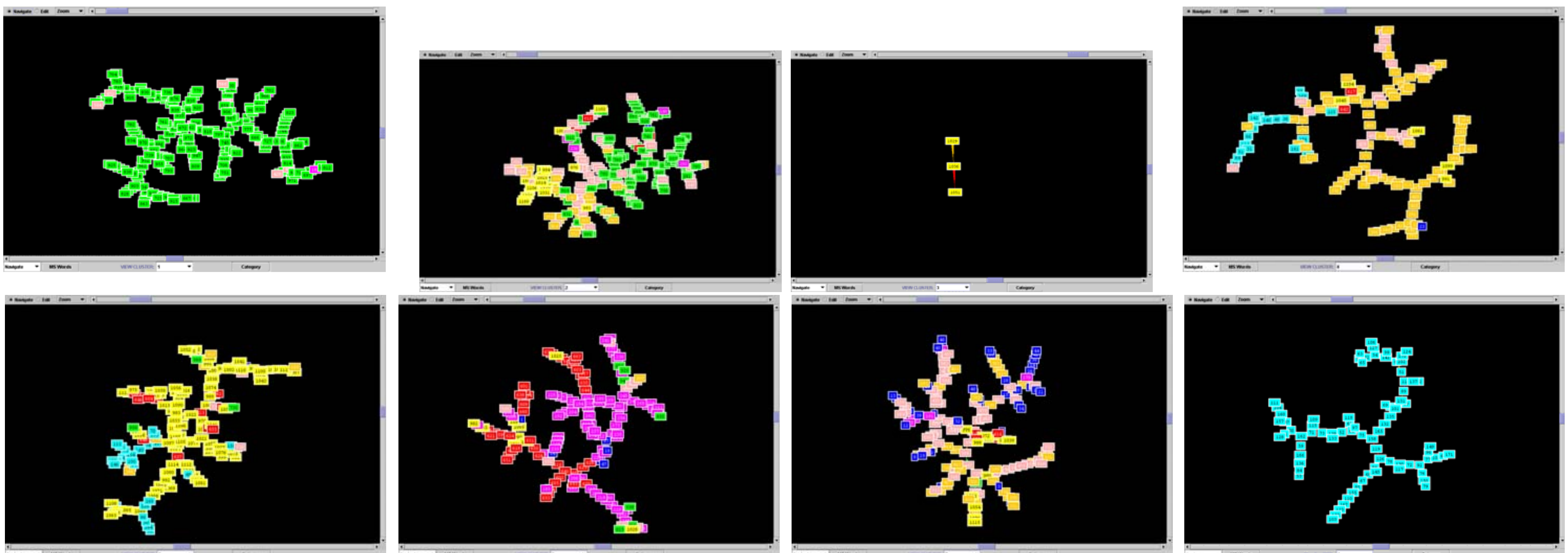
$$\hat{P} = \frac{\sum_{i=1}^{c} n_i P(D_i)}{n}$$

Average Purity Per Cluster

$$\overline{\overline{P}} = \frac{\sum_{i=1}^{c} n_i P(D_i)}{c}$$

Science News 8 Multi-partitioning

ANTHROPOLOGY & ARCHEOLOGY	ASTRONOMY & SPACE SCIENCES
BEHAVIOR	EARTH & ENVIRONMENTAL SCIENCES
LIFE SCIENCES	MATHEMATICS & COMPUTERS
MEDICAL SCIENCES	PHYSICAL SCIENCE & TECHNOLOGY



Science News 8 Multi-Partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	3	0	9	0	208	0
Cluster2	2	0	5	32	77	4	91	20
Cluster3	0	0	0	0	0	0	0	3
Cluster4	1	19	0	89	18	2	0	5
Cluster5	1	25	0	6	5	12	3	95
Cluster6	7	0	75	1	8	43	7	9
Cluster7	37	0	5	36	57	4	1	12
Cluster8	0	80	0	0	0	0	0	0

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

SciNews 8 Multi-Partitioning Purity & Entropy

PURITY

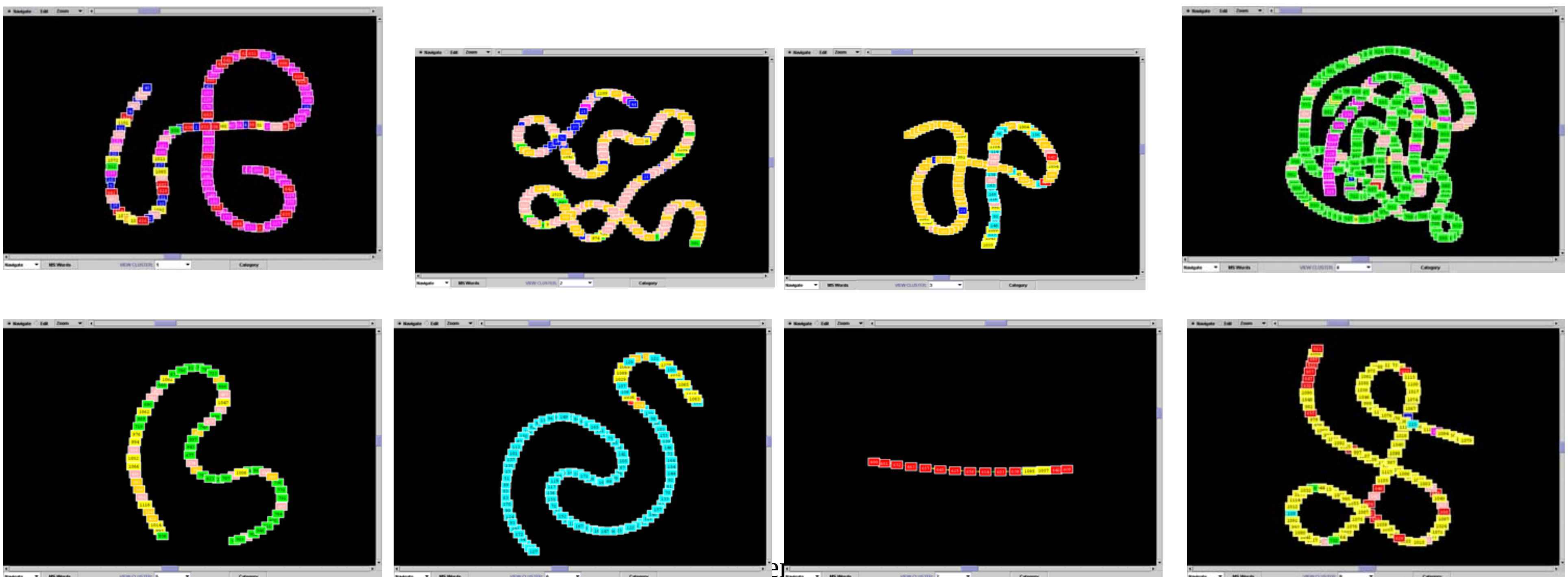
Cluster1	0.9454545454545454
Cluster2	0.3939393939393939
Cluster3	1.0
Cluster4	0.664179104477612
Cluster5	0.6462585034013606
Cluster6	0.5
Cluster7	0.375
Cluster8	1.0
Avg Purity	0.690603943409114
Avg Purity Per Observation	0.6248880931065354
Avg Aggregate Purity Per Cluster	87.25

ENTROPY

Cluster1	0.1165507016468692
Cluster2	0.6795872298749444
Cluster3	0.0
Cluster4	0.500340206242551
Cluster5	0.5515312792869759
Cluster6	0.6488882742515858
Cluster7	0.7186710223086614
Cluster8	0.0
Avg Entropy:	0.4019460892014485
Avg Entropy Per Observation	0.4810364607732902
Avg Aggregate Entropy Per Cluster	67.16471583547064

Science News 8 Recursive Bi-partitioning

ANTHROPOLOGY & ARCHEOLOGY
BEHAVIOR
LIFE SCIENCES
MEDICAL SCIENCES
ASTRONOMY & SPACE SCIENCES
EARTH & ENVIRONMENTAL SCIENCES
MATHEMATICS & COMPUTERS
PHYSICAL SCIENCE & TECHNOLOGY



Science News 8 Recursive-Bipartitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	19	0	55	1	10	29	2	9
Cluster2	25	0	3	69	79	0	10	9
Cluster3	2	20	0	75	10	2	0	13
Cluster4	1	0	29	8	57	3	263	4
Cluster5	0	0	0	8	13	0	33	11
Cluster6	0	102	0	3	0	1	0	9
Cluster7	0	0	0	0	0	13	0	2
Cluster8	1	2	1	0	5	17	2	87

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

.

SciNews 8 Recursive Bi-Partitioning Purity & Entropy

PURITY

Cluster1	0.44
Cluster2	0.405128205128205 13
Cluster3	0.6147540983606558
Cluster4	0.7205479452054795
Cluster5	0.5076923076923077
Cluster6	0.8869565217391304
Cluster7	0.8666666666666667
Cluster8	0.7565217391304347
Avg Purity	0.649783435490 36
Avg Purity Per Observation	0.6329453894359892
Avg Aggregate Purity Per Cluster	88.375

ENTROPY

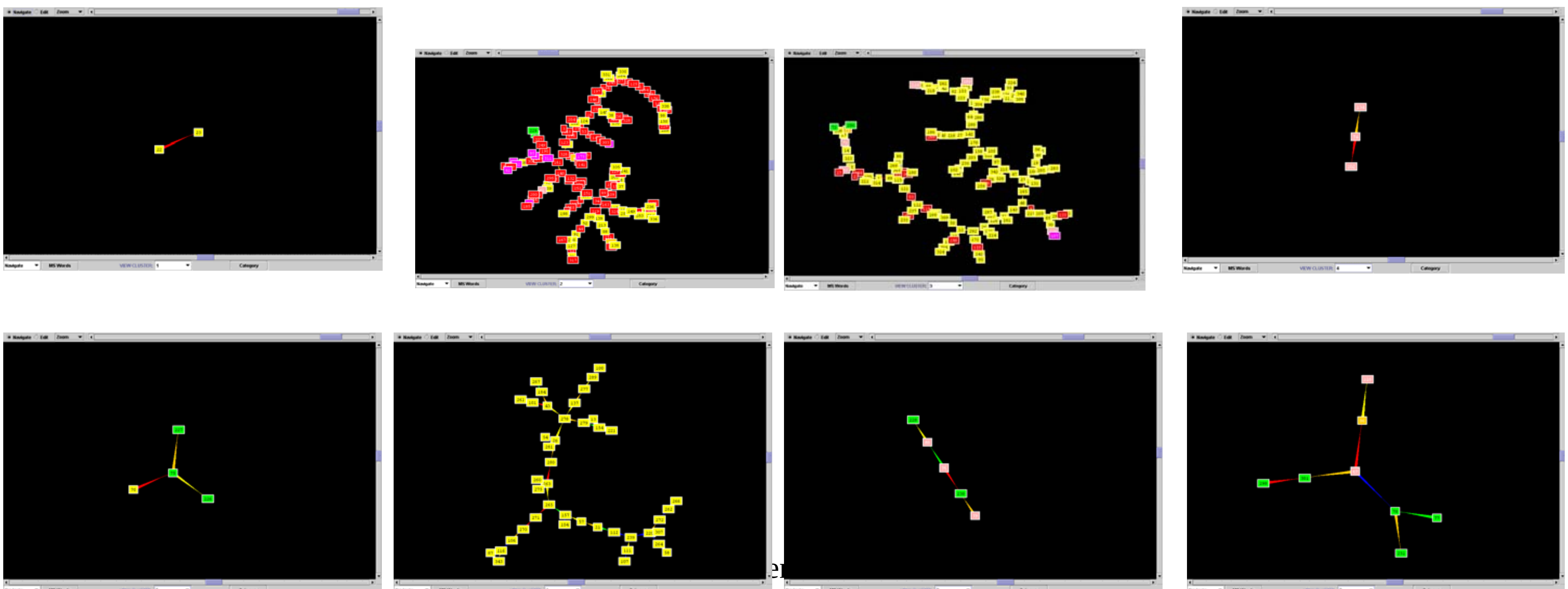
Cluster1	0.7130868633677933
Cluster2	0.6518677715369288
Cluster3	0.5645491645164644
Cluster4	0.44058717559174865
Cluster5	0.5888689116265443
Cluster6	0.21263698105033718
Cluster7	0.18883650218430179
Cluster8	0.41043119693933683
Avg Entropy	0.4713580708516819
Avg Entropy Per Observation	0.5001801770724928
Avg Aggregate Entropy Per Cluster	69.83765722374682

ONR ILIR Spectral Clustering Results

(Using Science News Categories)

ILIR 8 Multi-partitioning

ANTHROPOLOGY & ARCHEOLOGY	ASTRONOMY & SPACE SCIENCES
BEHAVIOR	EARTH & ENVIRONMENTAL SCIENCES
LIFE SCIENCES	MATHEMATICS & COMPUTERS
MEDICAL SCIENCES	PHYSICAL SCIENCE & TECHNOLOGY



ILIR 8 Multi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	0	0	0	0	0	2
Cluster2	0	0	9	2	1	83	2	46
Cluster3	0	0	1	1	5	15	2	111
Cluster4	0	0	0	0	3	0	0	0
Cluster5	0	0	0	0	0	0	3	1
Cluster6	0	0	0	0	0	0	0	43
Cluster7	0	0	0	0	3	0	2	0
Cluster8	0	0	0	1	2	0	5	0

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

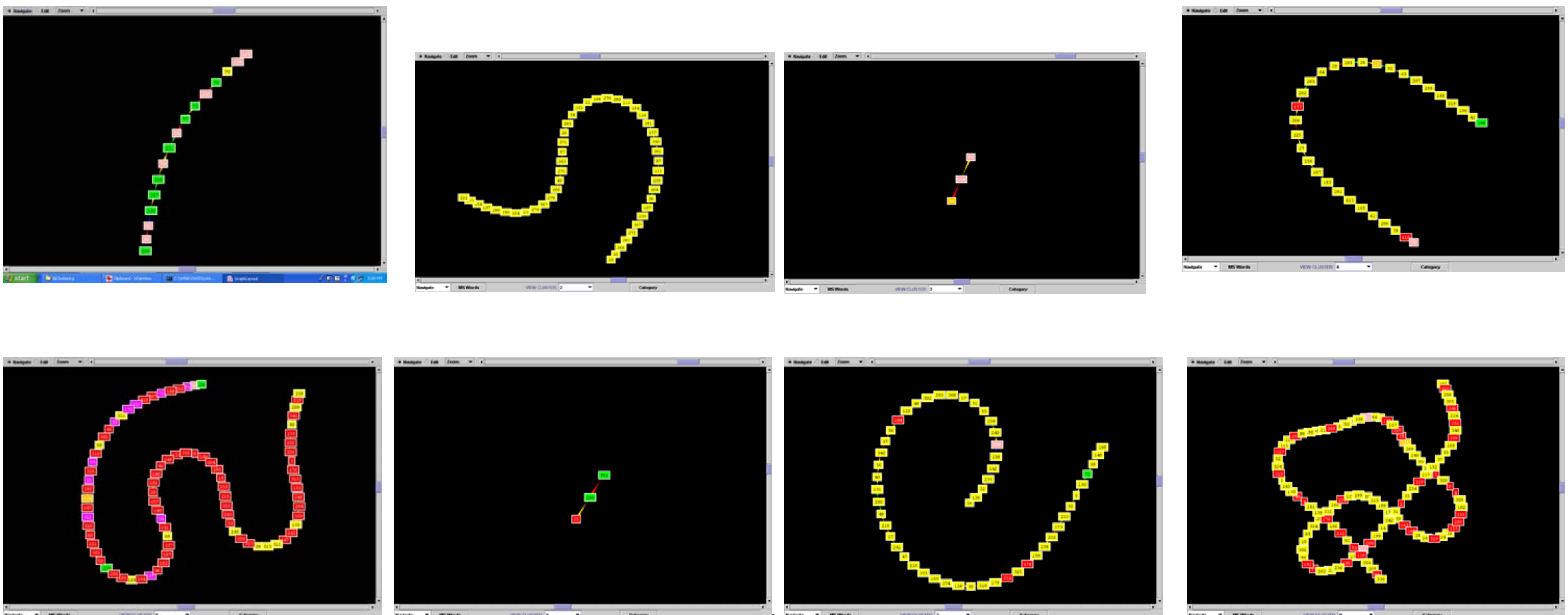
.

ILIR 8 Multi-Partitioning Purity & Entropy

PURITY		ENTROPY	
Cluster1	1.0	Cluster1	0.0
Cluster2	0.5804195804195804	Cluster2	0.48512856262450244
Cluster3	0.8222222222222222	Cluster3	0.31846159266280455
Cluster4	1.0	Cluster4	0.0
Cluster5	0.75	Cluster5	0.2704260414863776
Cluster6	1.0	Cluster6	0.0
Cluster7	0.6	Cluster7	0.32365019815155627
Cluster8	0.625	Cluster8	0.43293164689846625
Avg Purity	0.7972052253302253	Avg Entropy	0.22882475522796342
Avg Purity Per Observation	0.7376093294460642	Avg Entropy Per Observation	0.3455659119436545
Avg Aggregate Purity Per Cluster	31.625	Avg Aggregate Entropy Per Cluster	14.816138474584186

ILIR 8 Recursive-Bipartitioning

ANTHROPOLOGY & ARCHEOLOGY
BEHAVIOR
LIFE SCIENCES
MEDICAL SCIENCES
ASTRONOMY & SPACE SCIENCES
EARTH & ENVIRONMENTAL SCIENCES
MATHEMATICS & COMPUTERS
PHYSICAL SCIENCE & TECHNOLOGY



ILIR 8 Recursive-Bipartitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8
Cluster1	0	0	0	0	7	0	8	1
Cluster2	0	0	0	0	0	0	0	45
Cluster3	0	0	0	1	2	0	0	0
Cluster4	0	0	0	1	1	2	1	26
Cluster5	0	0	10	1	1	58	2	12
Cluster6	0	0	0	0	0	1	2	0
Cluster7	0	0	0	0	1	3	1	48
Cluster8	0	0	0	1	2	34	0	71

Class 1 is anthropology and archaeology, class 2 astronomy and space sciences, class 3 is behavior, class 4 is earth and environmental sciences, class 5 is life sciences, class 6 is mathematics and computers, class 7 is medical sciences, and class 8 is physical sciences and technology.

.

ILIR 8 Recursive Bi-Partitioning Purity & Entropy

PURITY

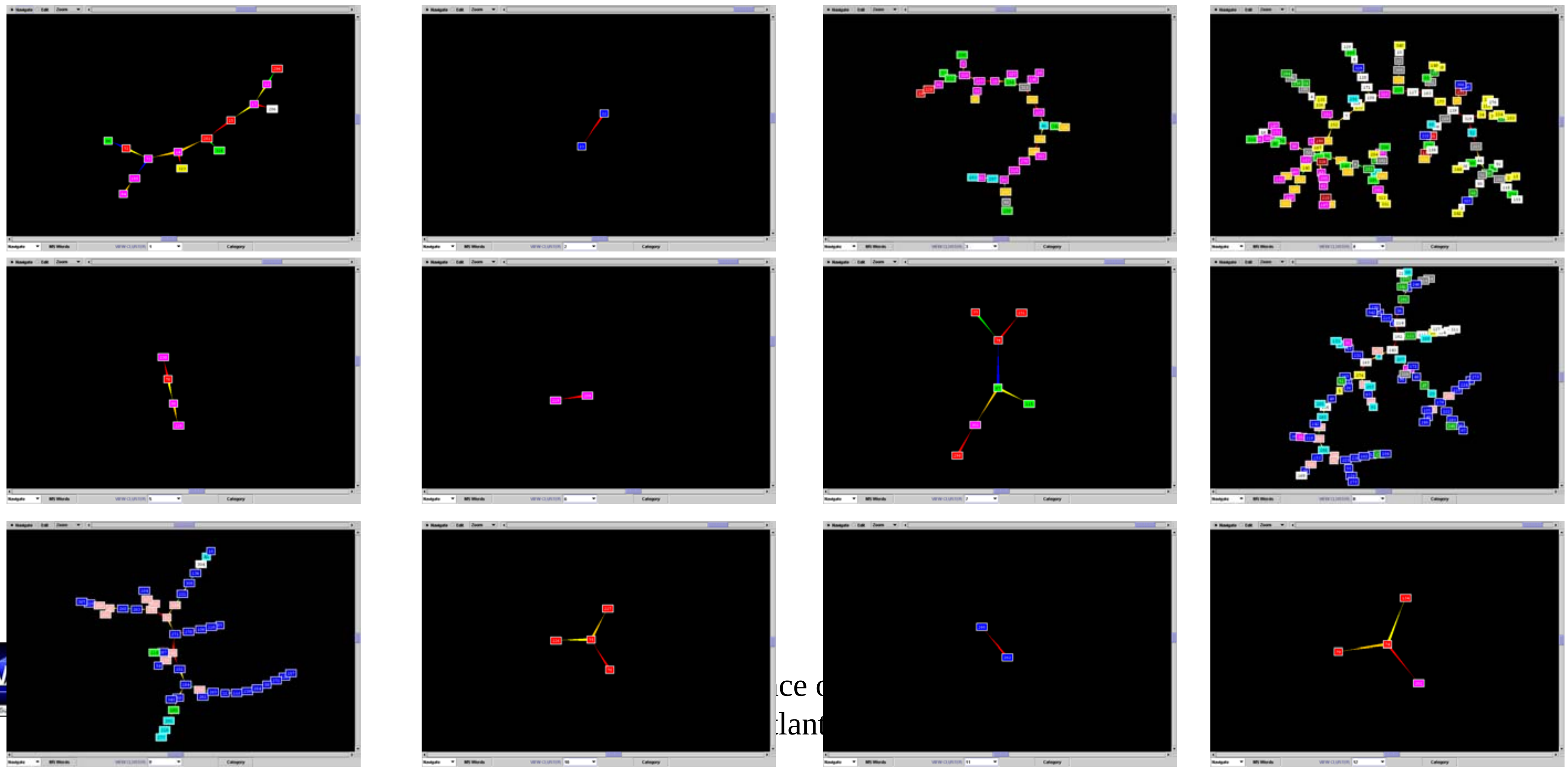
Cluster1	0.5
Cluster2	1.0
Cluster3	0.6666666666666666
Cluster4	0.8387096774193549
Cluster5	0.6904761904761905
Cluster6	0.6666666666666666
Cluster7	0.9056603773584906
Cluster8	0.6574074074074074
Avg Purity	0.7406983732493471
Avg Purity Per Observation	0.7580174927113703
Avg Aggregate Purity Per Cluster	32.5

ENTROPY

Cluster1	0.42392740719993277
Cluster2	0.0
Cluster3	0.3060986113514965
Cluster4	0.3157919672077929
Cluster5	0.4720354557082904
Cluster6	0.3060986113514965
Cluster7	0.1933754500318905
Cluster8	0.36395700045157614
Avg Entropy	0.29766056291280946
Avg Entropy Per Observation	0.3137498960545373
Avg Aggregate Entropy Per Cluster	13.452026793338288

ILIR 12 Multi-partitioning

Advanced Naval Materials	Air Platforms and Systems
Human Performance /Factors	Information Technology and Operations
Manufacturing Technologies	Medical S&T
Operational Environments	RF Sensing, Surveil, & Countermeasures
Sea Platform and Systems	USW-ASW
USW-MIW	Visible and IR Sensing, Surveil & Countermeasures



Advanced Naval Materials(1) Air Platforms and Systems(2)
Human Performance /Factors(3) Information Technology and Operations(4)
Manufacturing Technologies(5) Medical S&T(6)
Operational Environments(7) RF Sensing, Surveil, & Countermeasures(8)
Sea Platform and Systems(9) USW-ASW(10)
USW-MIW(11) Visible and IR Sensing, Surveil & Countermeasures(12)

ILIR 12 Multi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class 10	Class 11	Class 12
Cluster1	0	0	6	0	0	4	2	1	1	0	0	0
Cluster2	2	0	0	0	0	0	0	0	0	0	0	0
Cluster3	0	3	15	6	0	2	6	0	0	0	2	0
Cluster4	5	4	18	12	0	1	15	23	24	5	12	9
Cluster5	0	0	3	0	0	1	0	0	0	0	0	0
Cluster6	0	0	2	0	0	0	0	0	0	0	0	0
Cluster7	0	0	1	0	0	4	2	0	0	0	0	0
Cluster8	43	12	3	0	10	0	0	3	12	0	3	8
Cluster9	30	4	0	0	11	0	2	0	1	0	0	0
Cluster10	0	0	0	0	0	4	0	0	0	0	0	0
Cluster11	2	0	0	0	0	0	0	0	0	0	0	0
Cluster12	0	0	1	0	0	3	0	0	0	0	0	0

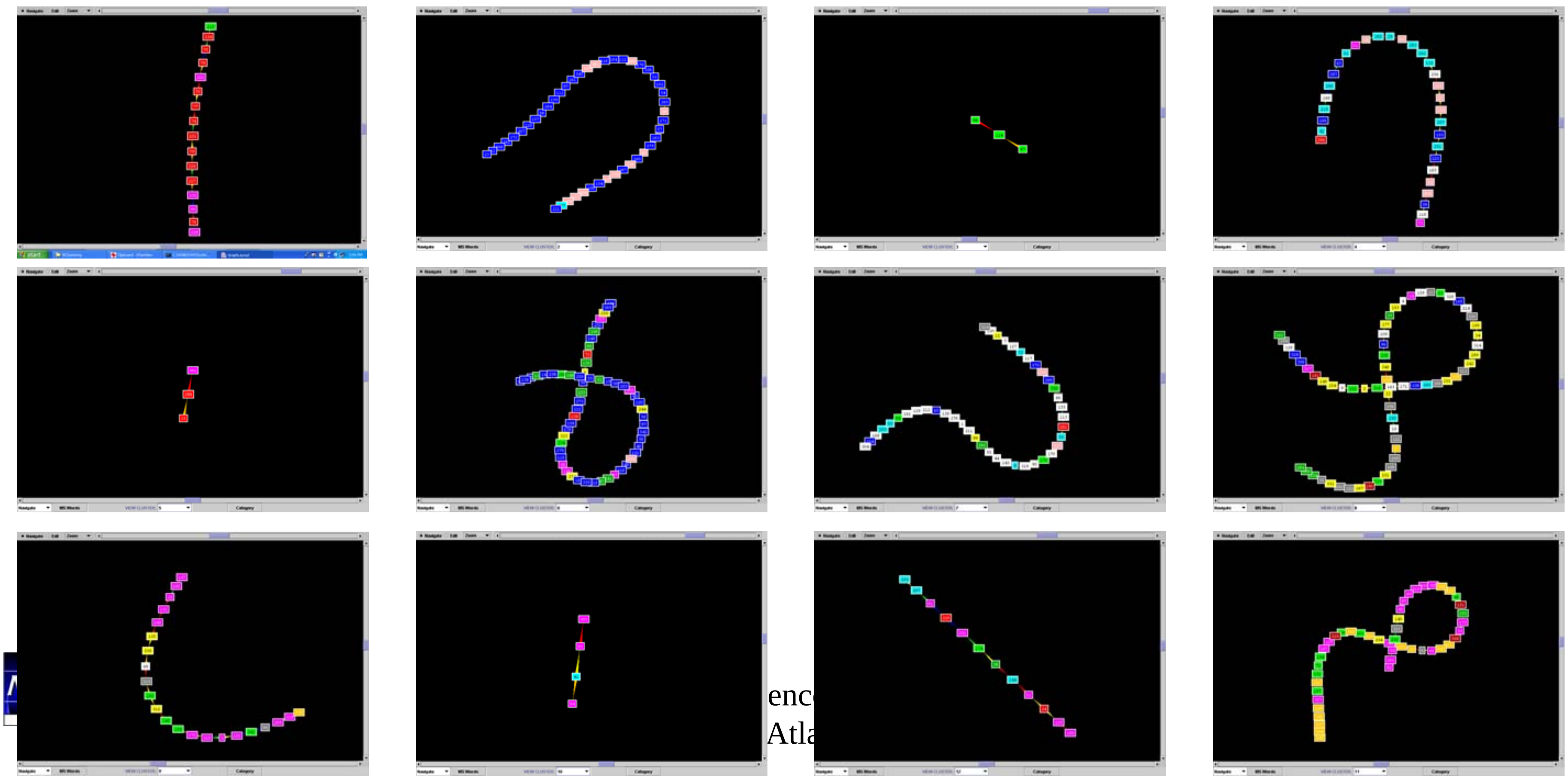
ILIR 12 Multi-Partitioning Purity & Entropy

PURITY	
Cluster1	0.42857142857142855
Cluster2	1.0
Cluster3	0.4411764705882353
Cluster4	0.1875
Cluster5	0.75
Cluster6	1.0
Cluster7	0.5714285714285714
Cluster8	0.4574468085106383
Cluster9	0.625
Cluster10	1.0
Cluster11	1.0
Cluster12	0.75
Avg Purity	0.6842602732582396
Avg Purity Per Observation	0.40233236151603496
Avg Aggregate Purity Per Cluster	11.5

ENTROPY	
Cluster1	0.5537653840548961
Cluster2	0.0
Cluster3	0.612000331799029
Cluster4	0.877077819793199
Cluster5	0.22630030977895443
Cluster6	0.0
Cluster7	0.3846019290881892
Cluster8	0.6685102839439432
Cluster9	0.4231665274665911
Cluster10	0.0
Cluster11	0.0
Cluster12	0.22630030977895443
Avg Entropy	0.33097690797531293
Avg Entropy Per Observation	0.666126132893414
Avg Aggregate Entropy Per Cluster	19.04010529853675

ILIR 12 Recursive Bi-partitioning

Advanced Naval Materials	Air Platforms and Systems
Human Performance /Factors	Information Technology and Operations
Manufacturing Technologies	Medical S&T
Operational Environments	RF Sensing, Surveil, & Countermeasures
Sea Platform and Systems	USW-ASW
USW-MIW	Visible and IR Sensing, Surveil & Countermeasures



Advanced Naval Materials(1) Air Platforms and Systems(2)
Human Performance /Factors(3) Information Technology and Operations(4)
Manufacturing Technologies(5) Medical S&T(6)
Operational Environments(7) RF Sensing, Surveil, & Countermeasures(8)
Sea Platform and Systems(9) USW-ASW(10)
USW-MIW(11) Visible and IR Sensing, Surveil & Countermeasures(12)

ILIR 12 Recursive-Bi-partitioning Confusion Matrix

	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class 10	Class 11	Class 12
Cluster1	0	0	4	0	0	11	1	0	0	0	0	0
Cluster2	33	1	0	0	11	0	0	0	0	0	0	0
Cluster3	0	0	0	0	0	0	3	0	0	0	0	0
Cluster4	6	11	2	0	7	1	0	0	4	0	0	0
Cluster5	0	0	1	0	0	2	0	0	0	0	0	0
Cluster6	34	0	5	0	1	2	2	5	0	0	0	9
Cluster7	4	5	0	0	2	1	3	2	22	0	1	1
Cluster8	5	2	2	3	0	0	5	15	11	2	12	5
Cluster9	0	0	11	1	0	0	4	3	1	0	2	0
Cluster10	0	1	3	0	0	0	0	0	0	0	0	0
Cluster11	0	0	16	14	0	0	8	2	0	3	2	1
Cluster12	0	3	5	0	0	2	1	0	0	0	0	1

ILIR 12 Recursive Bi-Partitioning Purity & Entropy

PURITY

Cluster1	0.6875
Cluster2	0.7333333333333333
Cluster3	1.0
Cluster4	0.3548387096774194
Cluster5	0.6666666666666666
Cluster6	0.5862068965517241
Cluster7	0.5365853658536586
Cluster8	0.24193548387096775
Cluster9	0.5
Cluster10	0.75
Cluster11	0.34782608695652173
Cluster12	0.4166666666666667
Avg Purity	0.5684632674647465
Avg Purity Per Observation	0.4839650145772595
Avg Aggregate Purity Per Cluster	13.833333333333334

ENTROPY

Cluster1	0.31287377816678064
Cluster2	0.2641567106578208
Cluster3	0.0
Cluster4	0.6331562009104378
Cluster5	0.2561521449303204
Cluster6	0.5340341300193764
Cluster7	0.6340006351665233
Cluster8	0.8273794447785848
Cluster9	0.5743544330688691
Cluster10	0.22630030977895443
Cluster11	0.6308112406531853
Cluster12	0.5731120392262111
Avg Entropy	0.45552758894642l9
Avg Entropy Per Observation	0.5684854885128293
Avg Aggregate Entropy Per Cluster	16.24921021332504

Future

- Development of visualization frameworks that allow for simultaneous display of words and documents.

- Tree-based displays for the recursive bipartitioning tree.

- Higher dimensional visualization in the case of the multipartition algorithm.

Backup Slides

Methodology

- ILIR1:
- 12 classification categories
- Heirarchical Clustering
- Method: Average
- Tree Cut: 24

An Alternate Approach

Multi-Discipline Document Set → Exemplar Term Production Via Synonym Analysis → Feature Extraction (BPM, TPM) → Dimensionality Reduction ISOMAP/LLE → Model-Based Clustering With Adaptive Mixtures Initialization → Serendipity Identification and Visualization

A Paradigm

"you don't reach Serendip by plotting a course for it. You have to set out in good faith for elsewhere and lose your bearings … serendipitously."

-- John Barth, *The Last Voyage of Somebody the Sailor*

Acknowledgements

- Jim Gentle (Opportunity to speak)
- Algotek (Funding and Program Management)
  - Anna Tsao
- Algotek Team (Helpful discussions and encouragement)
  - Carey Priebe
  - David Marchette

The Porter Stemming Algorithm

- "The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems. " ('official' home page for distribution of the Porter Stemming Algorithm http://www.tartarus.org/~martin/PorterStemmer)

Our Approach to be Discussed Today

Multi-Discipline Document Set

Feature Extraction (Denoising, stemming, BPM, TPM)

Interpoint Distance Calculation

Minimal Spanning Tree (MST) Calculation

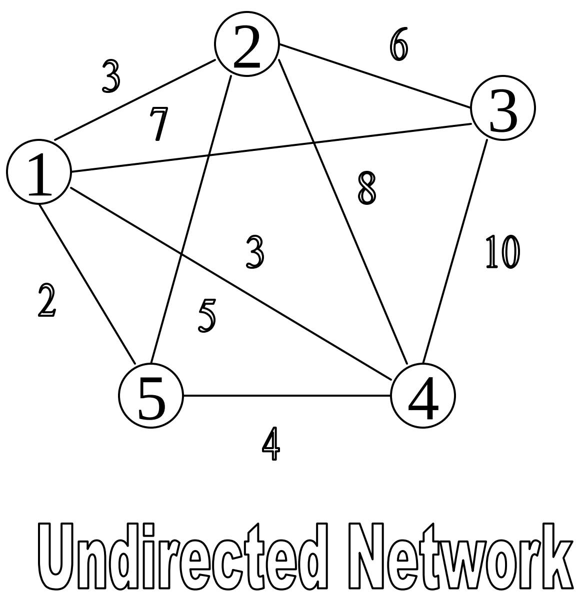
MST Layout Via Spring Based Models

Text Data Mining Via MST Exploration

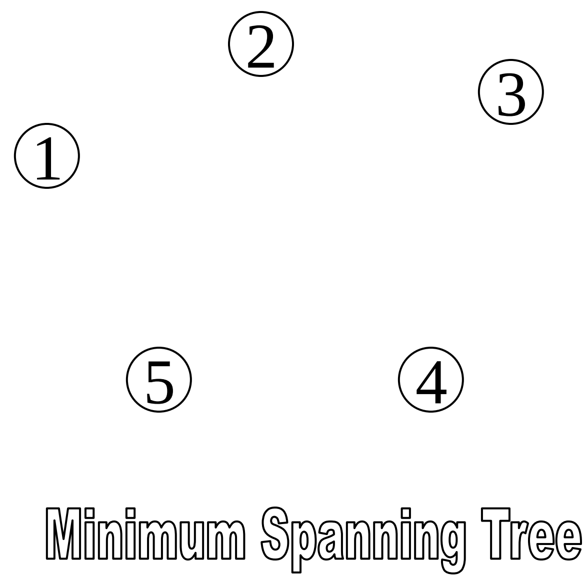
Cross Corpora Associations

Cluster Determination and Exploration

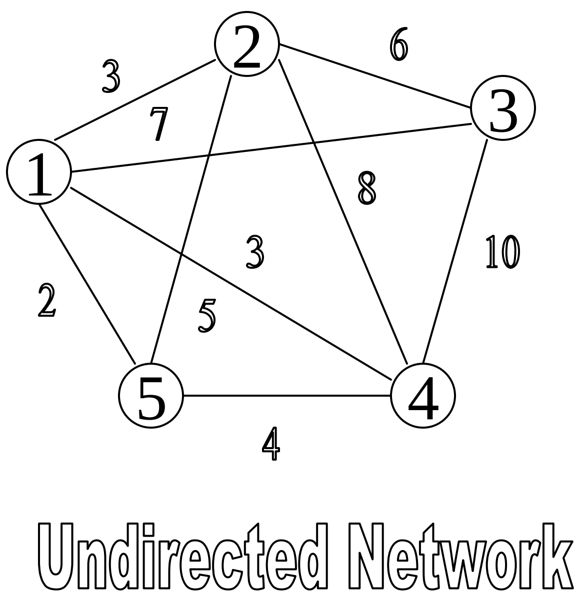
Calculation of the MST : Kruskal's Algorithm



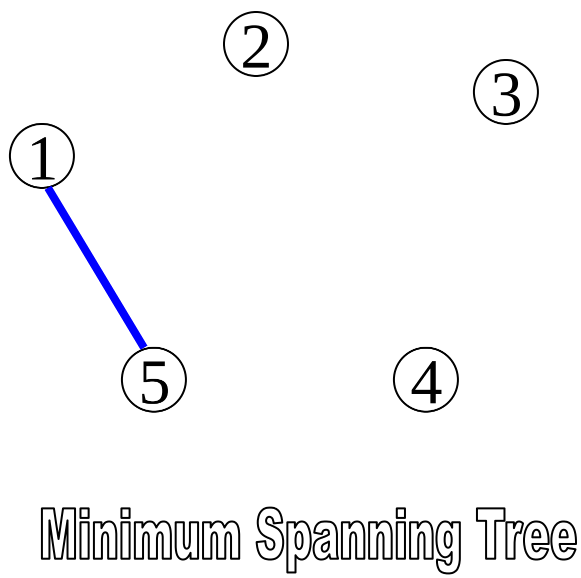
2
3
3
4
5
6
7
8
10



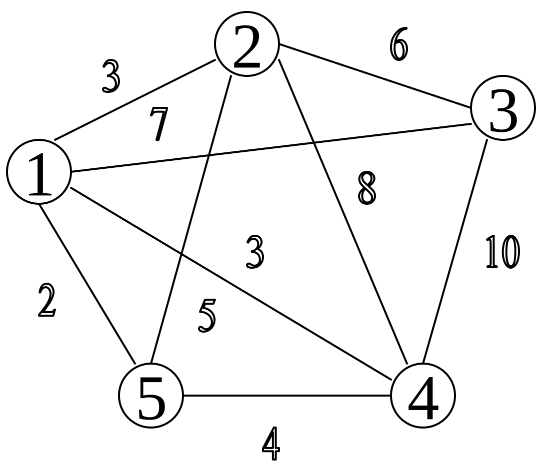
Calculation of the MST : Kruskal's Algorithm



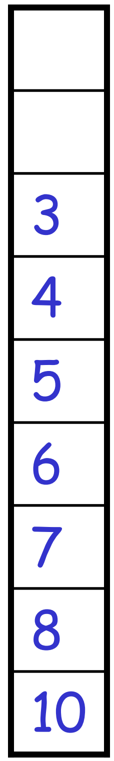
3
3
4
5
6
7
8
10

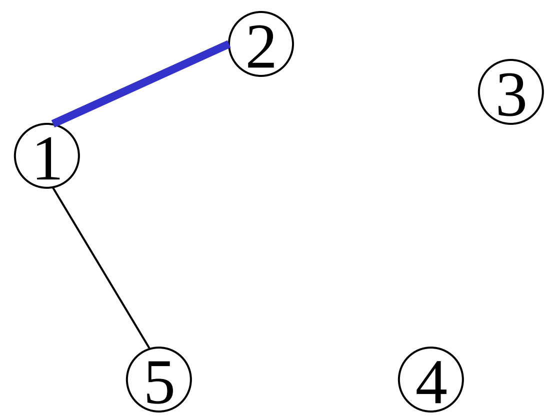


Calculation of the MST : Kruskal's Algorithm



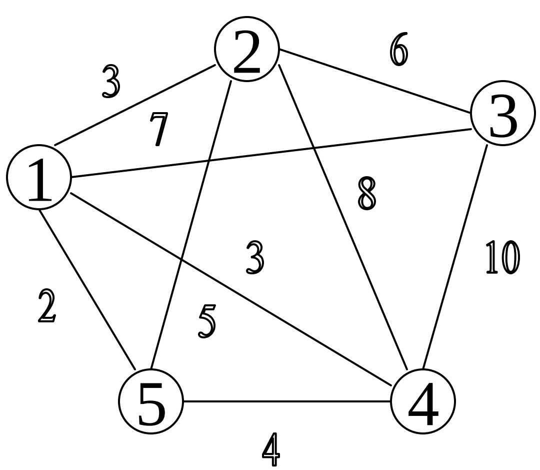
Undirected Network



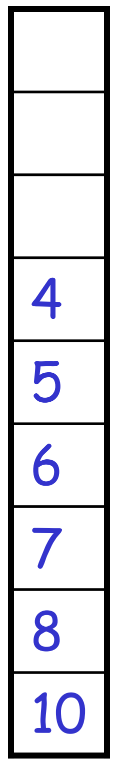


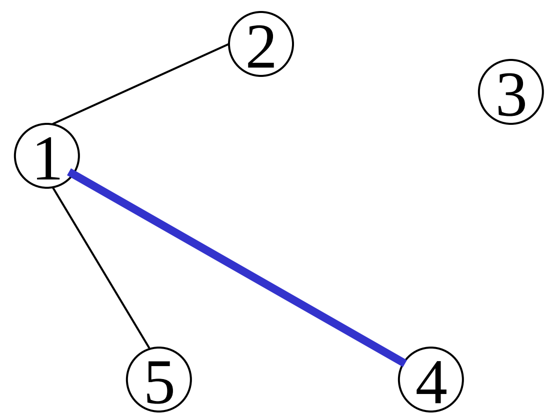
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



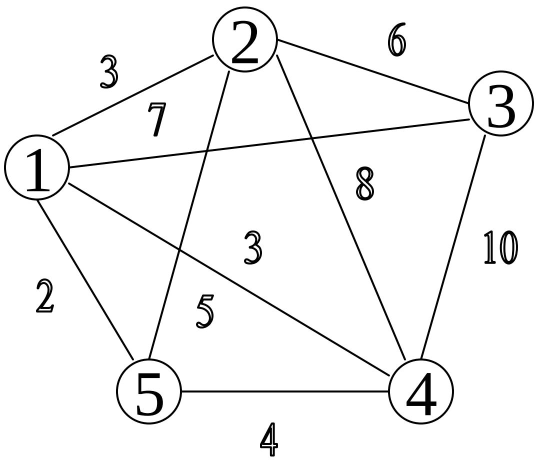
Undirected Network



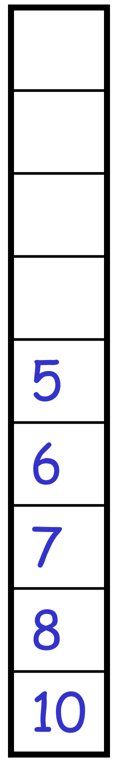


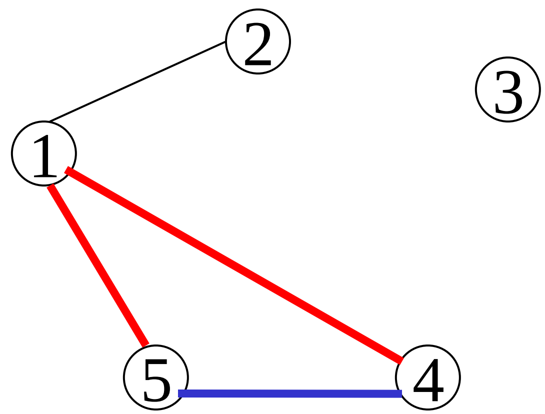
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



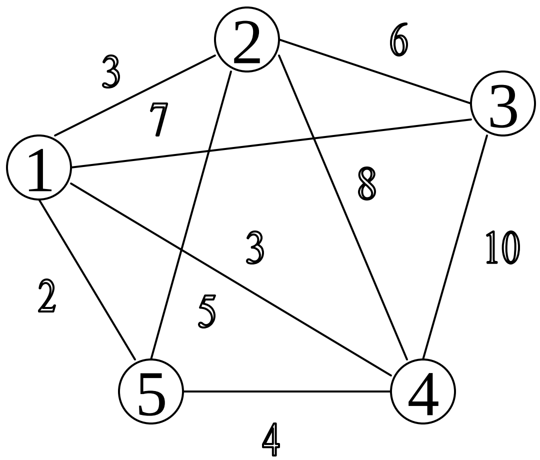
Undirected Network



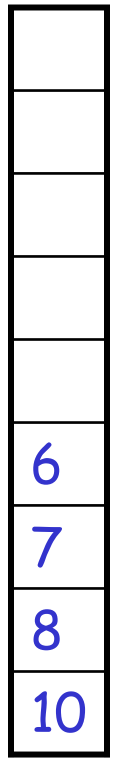


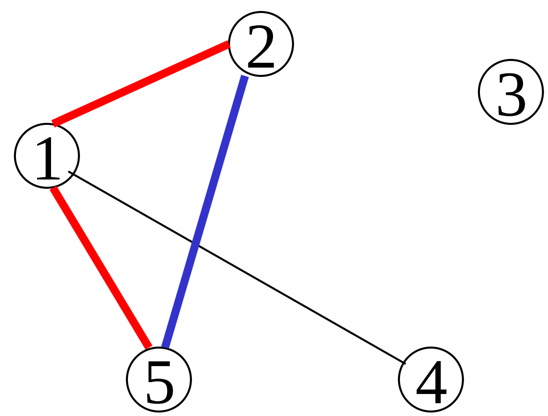
Minimum Spanning Tree

Calculation of the MST : Kruskal's Algorithm



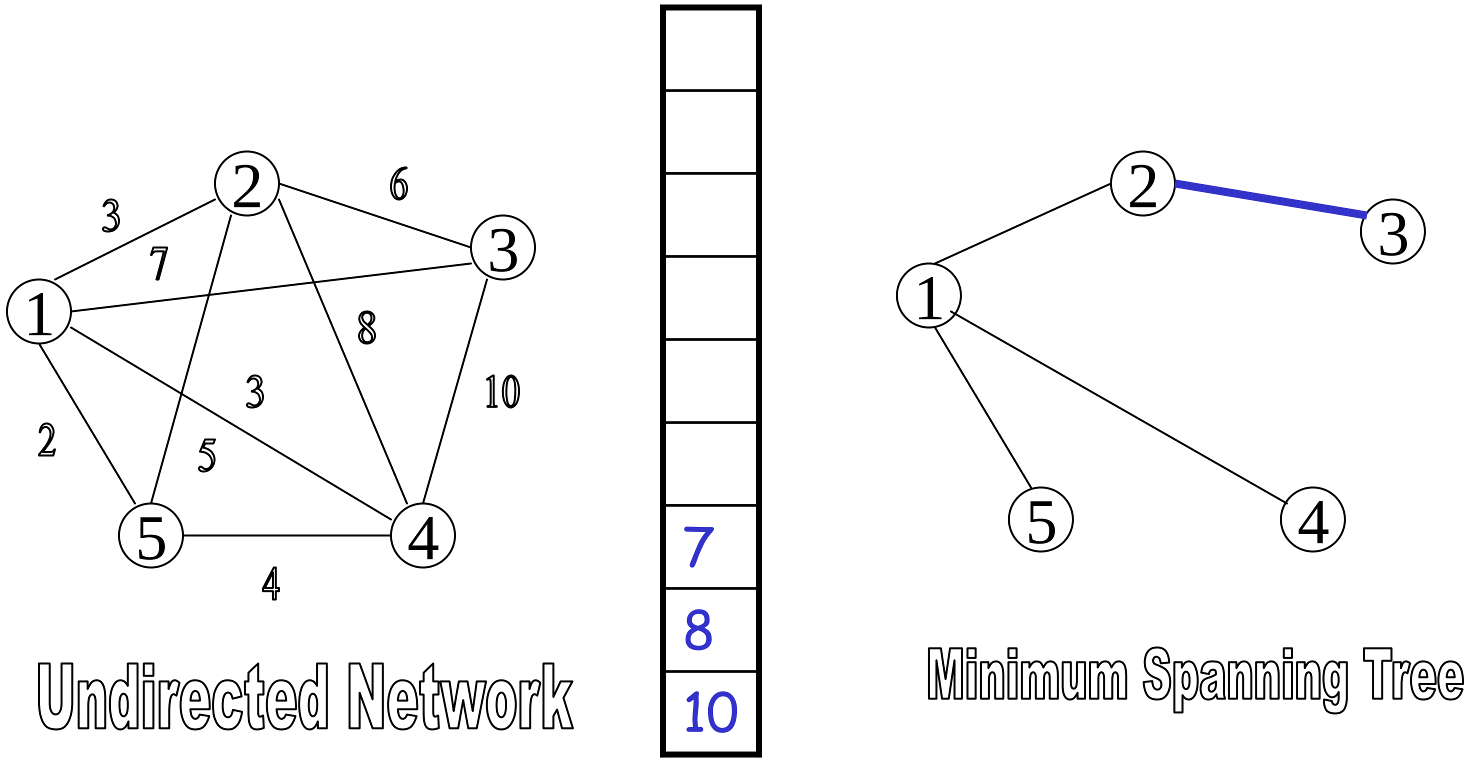
Undirected Network



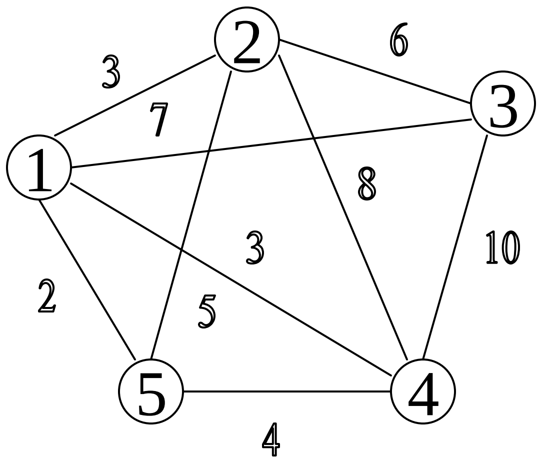


Minimum Spanning Tree

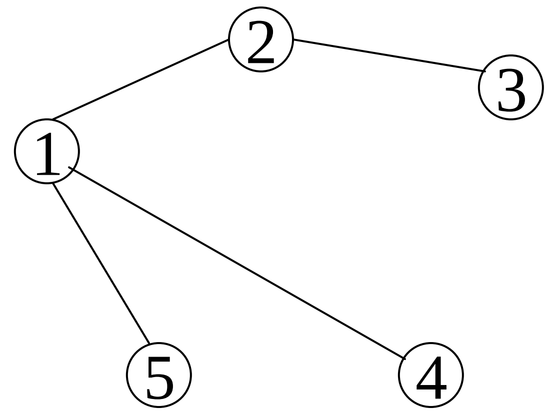
Calculation of the MST : Kruskal's Algorithm



Calculation of the MST : Kruskal's Algorithm



Undirected Network



Minimum Spanning Tree

Implementation Issues (The Devil in the Details)

- BPM extraction and interpoint distance calculation:
  - Implemented in C#.
- BPM similarity and distance calculation:
  - Implemented in C#.
- MST calculation:
  - Implemented using Kruskal's algorithm in JAVA.
- Cluster calculations are performed using JAVA
- Visualization environment:
  - Implemented in JAVA.
  - Graph layout facilitated using TouchGraph.

TouchGraph

- TouchGraph is a general public license JAVA-based library for the visualization of graphs. (www.touchgraph.com)
- Graph layout in TouchGraph:
  - When a graph is first loaded, nodes start out at the center with slightly random positions, and then spread out because of node-node repulsions.
- Graph manipulation tools provided by TouchGraph.
  - Zooming.
  - Rotation.
  - Hyperbolic manipulation.
  - Graph dragging.

Equations - I

$$M = \begin{cases} E_{ij}, & \text{if there is an edge } \{i, j\}, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{cut}(\mathcal{V}_1^*, \mathcal{V}_2^*) = \min_{\mathcal{V}_1, \mathcal{V}_2} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2).$$

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2) = \sum_{i \in \mathcal{V}_1, j \in \mathcal{V}_2} M_{ij}.$$

$$\mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k) = \sum_{i<j} \mathrm{cut}(\mathcal{V}_i, \mathcal{V}_j)$$

$$E_{ij} = t_{ij} \times \log\left(\frac{|\mathcal{D}|}{|\mathcal{D}_i|}\right)$$

$$M = \begin{bmatrix} \mathbf{0} & A \\ A^T & \mathbf{0} \end{bmatrix}$$

$$\mathrm{cut}(\mathcal{W}_1 \cup \mathcal{D}_1, \mathcal{W}_2 \cup \mathcal{D}_2, \ldots, \mathcal{W}_k \cup \mathcal{D}_k) = \min_{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k} \mathrm{cut}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k)$$

Wrap-up

- Demonstrated a new method for cross corpora document discovery

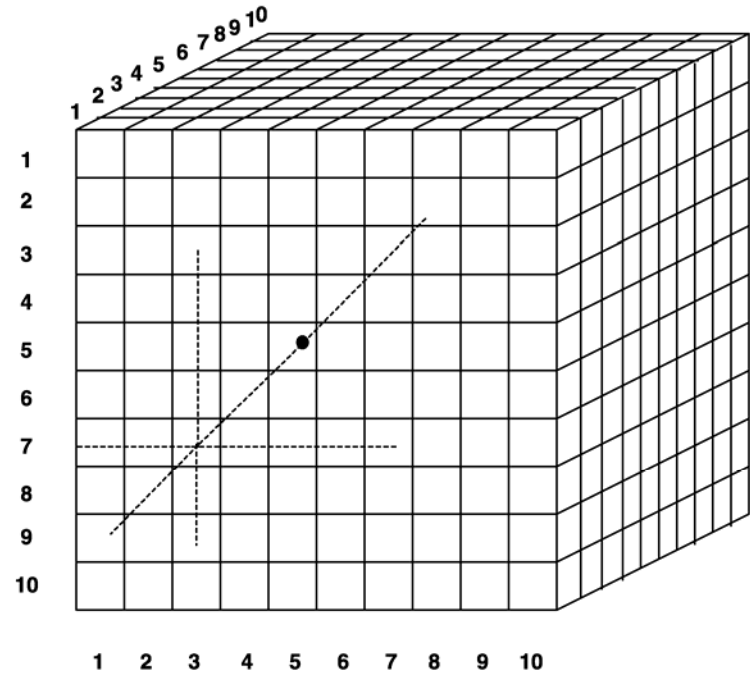
- Method predicated on the use of BPM and the MST as a convenient foil for the exploration of the cross corpora relationships.

- This work represents the tip of the iceberg of a new area that is not only of strategic importance to the United States but also is highly relevant to all who are currently conducting research in any discipline.

Feature Extraction
(Bigram Proximity Matrix (BPM) & Trigram Proximity Matrix (TPM))

Table 2.2 Example of Bigram Proximity Matrix

	.	crowd	his	in	father	man	sought	the	wise	young
.										
crowd	1									
his					1					
in								1		
father				1						
man							1			
sought			1							
the		1							1	
wise										1
young						1				



1 . (period)
2 crowd
3 his
4 in
5 father
6 man
7 sought
8 the
9 wise
10 young

"The wise young man <u>sought his father</u> in the crowd."

Evidence That BPM and TPM Capture Semantic Content

- Angel Martinez, "A Framework for the Representation of Semantics," *Ph.D Dissertation under the direction of Edward Wegman*, October 2002.
  - Supervised Learning.
  - Hypothesis Tests (3 sets of tests).
  - Unsupervised Learning.
  - Supervised Learning in a Reduced Dimension Space.

Similarity Measures and Pseudometrics on the BPM

- Following Martinez (2002) we propose the use of the Ochiai measure in the case of the BPM:

$$S(X,Y) = \frac{|X \text{ and } Y|}{\sqrt{(|X||Y|)}}$$

- This is converted to a distance via:

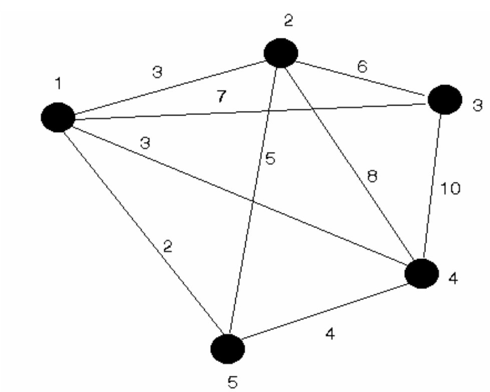
$$d(X,Y) = \sqrt{(2 - 2S(X,Y))}$$

How Do We Exploit This Interpoint Distance Matrix for Clustering?

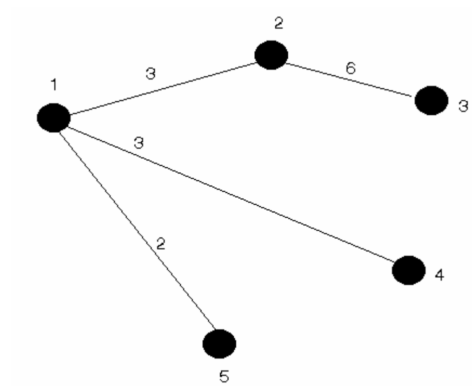
- First order exploration
  - Visualization of cluster structures
- Second order exploration
  - Exploration of cluster structures to ascertain interesting cross (within) corpora relationships

The Minimal Spanning Tree (MST): A Strategy for Effective Exploration of the Interpoint Distance Matrix and Cluster Computation

- Definition (Minimal Spanning Tree (MST)) – The collection of edges that join all of the points in a set together, with the minimum possible sum of edge values. The edge values that will be used here is the distance measures stored in our interpoint distance matrix.

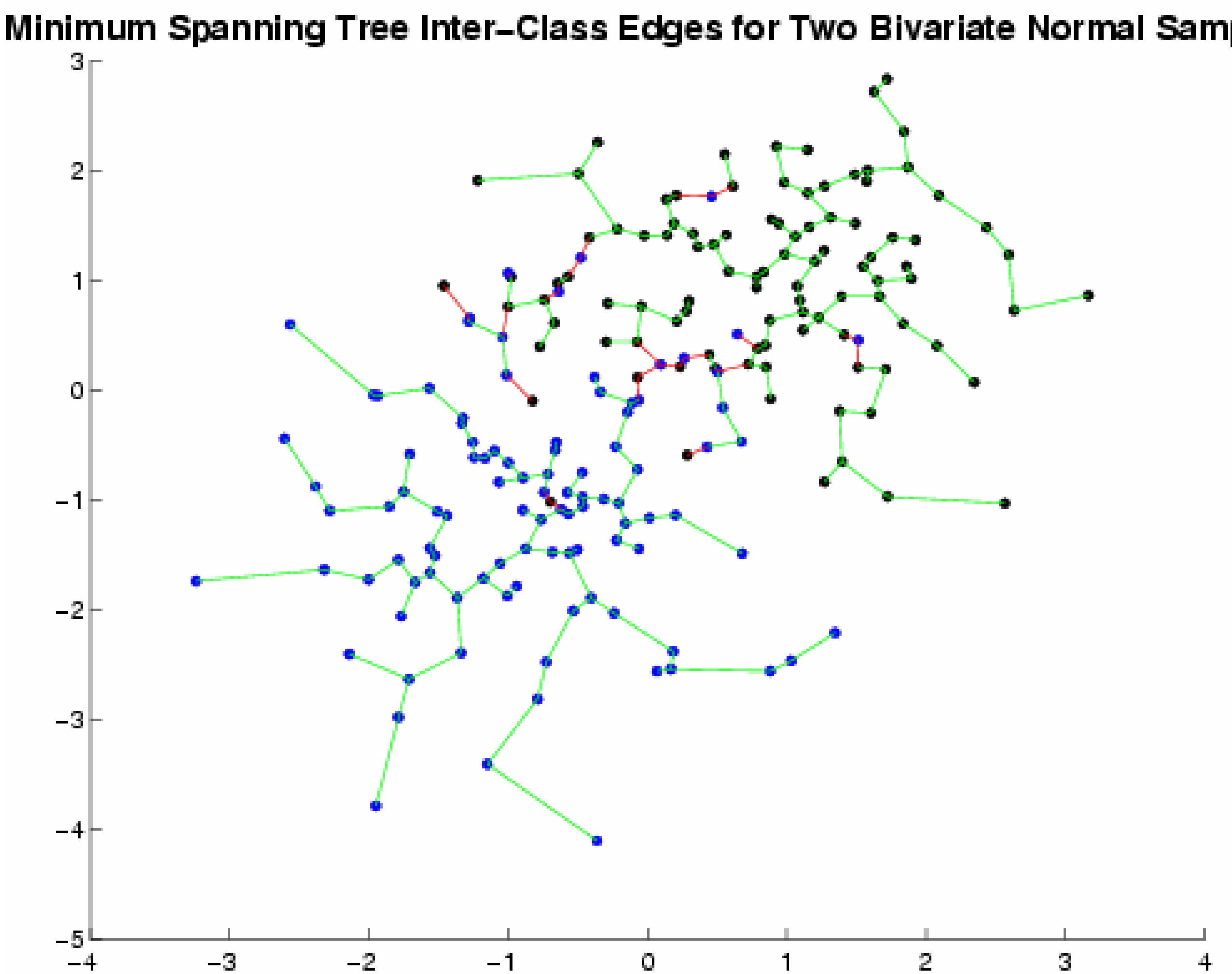


A complete graph.



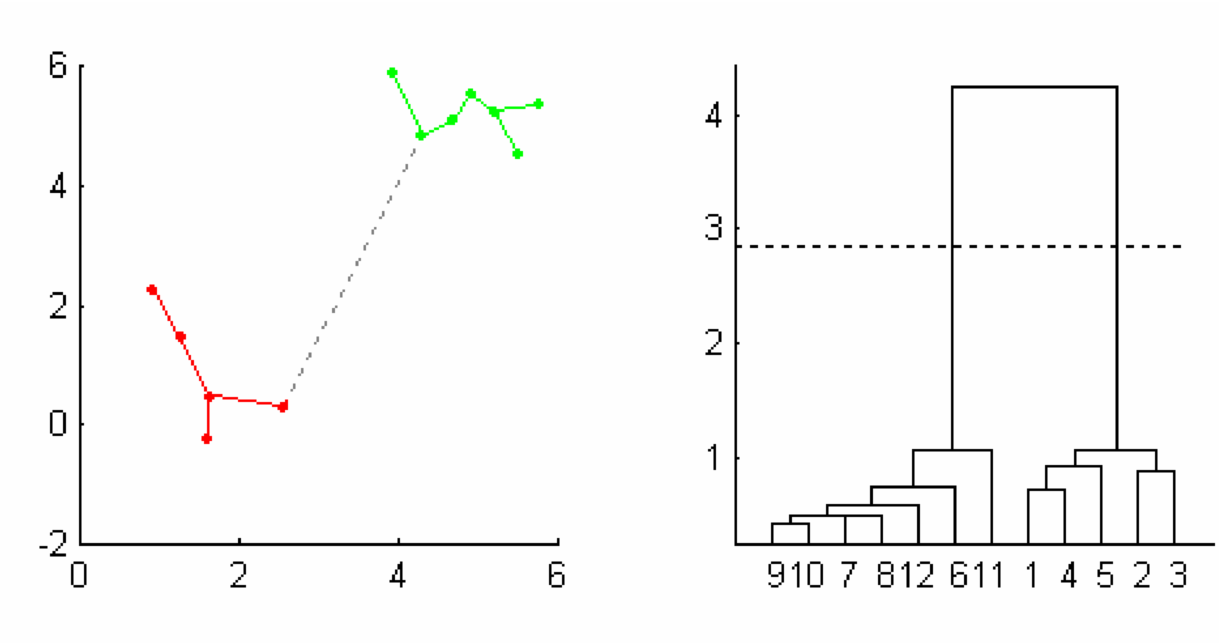
Associated MST.

MST Classifier Complexity Characterization



Previous work had suggested that the number of cross class edges can be used as a surrogate for classification complexity. <u>These cross class (corpora) edges will be used in our scheme to facilitate the cross-corpora discovery process.</u>

MST-based Clustering



All the single-linkage clusters could be obtained by deleting the edges of the MST, starting from the largest one.

Adapted from - Course: Cluster Analysis and Other Unsupervised Learning Methods (Stat 593 E)

Speakers: Rebecca Nugent[1], Larissa Stanberry[2]
Department of [1] Statistics, [2] Radiology,
University of Washington

Applications of MST-based Clustering and Data Mining to Geospatial Data

- Diansheng Guo, Donna Peuquet, Mark Gahegan, "Opening the Black Box: Interactive Hierarchical Clustering for Multivariate Spatial Patterns," *GIS'02*, November 8-9, 2002, McLean, Virginia, USA.

- Guo, D., D. Peuquet and M. Gahegan, "ICEAGE: Interactive Clustering and Exploration of Large and High-Dimensional Geodata" *GeoInformatica*, **7**(3): 229-253, 2003.

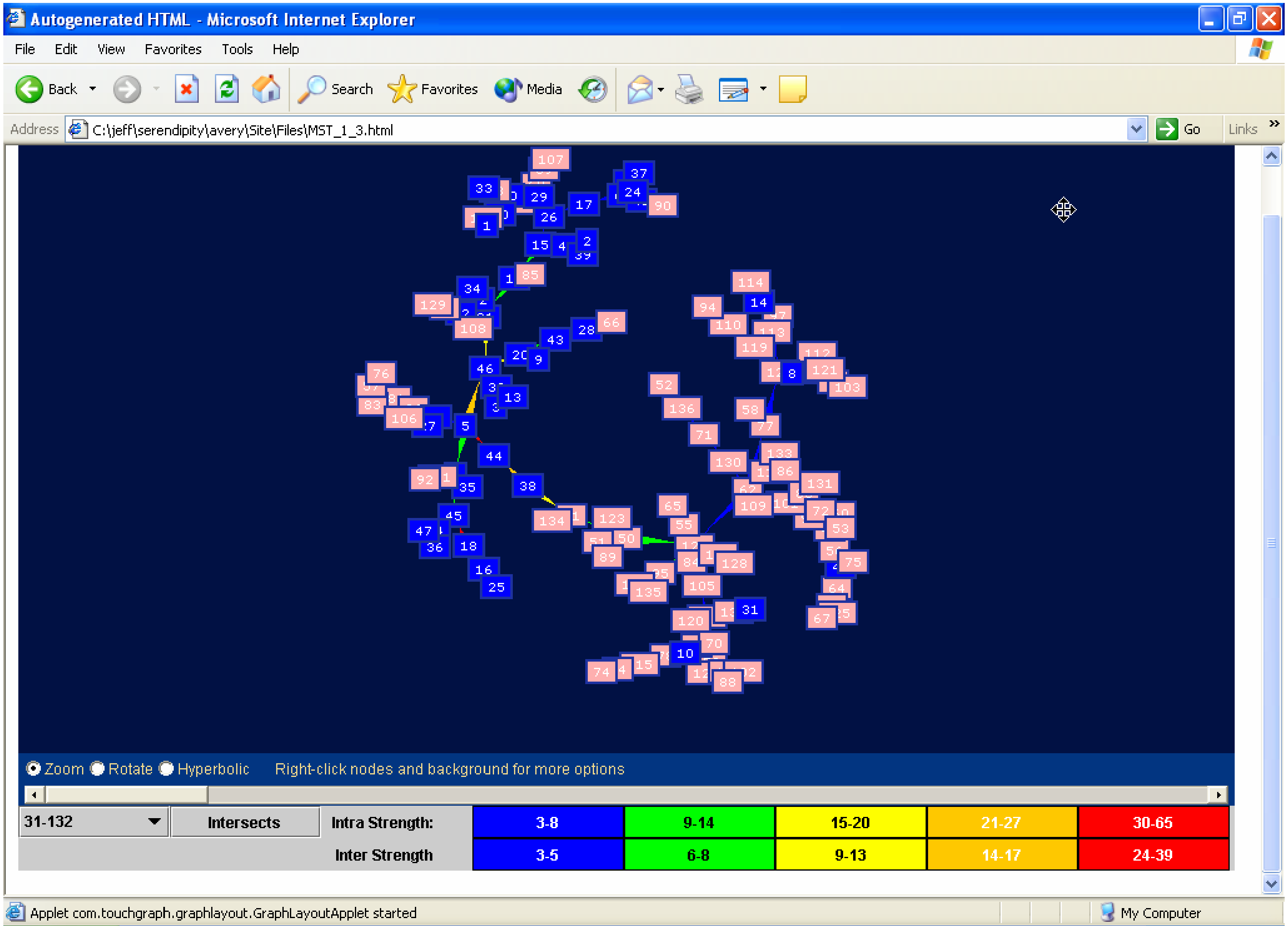
Applications of MST-based Clustering to Gene Expression Data

- Ying Xu, Victor Olman, and Dong Xu, "Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees," *Bioinformatics* Vol. 18 no. 4, pp. 536-545, 2002.

- Ying Xu, Victor Olman, and Dong Xu, "Minimum spanning trees for gene expression clustering," *Genome Informatics, 12: 24–33, 2001.*

The Environment (Opening Screen)

The Environment (MST)



Blue is anthropology and archaeology. Pink is behavior.

The Environment (The Comparison File)

Anthropology & Archeology/Behavior

Slumber's Unexplored Landscape People in traditional societies sleep in eye-opening ways

A Gebusi woman in New Guinea, decked out in her dance costume, catches a few winks on a woodpile during a male initiation ceremony. (Eileen Knauft) Ah, the sweet simplicity of sleep. You tramp into your bedroom with sagging eyelids and stifle a yawn. After disrobing, you douse the lights and climb into bed. Maybe a little reading or television massages the nerves, loosening them up for slumber's velvet fingers. In a while, you nod off. Suddenly, an alarm clock's shrill blast breaks up the dozefest as the sun pokes over the horizon. You feel a bit drowsy but shake it off and face the new day. Images of a dream dissolve like sugar in the morning's first cup of coffee.

There's a surprising twist, however, at the heart of this familiar ritual. It simply doesn't apply to people currently living outside of the modern Western world-or even to inhabitants of Western Europe as recently as 200 years ago.

Brains in Dreamland

Scientists hope to raise the neural curtain on sleep's virtual theater

After his father's death in 1896, Viennese neurologist Sigmund Freud made a momentous career change. He decided to study the mind instead of the brain. Freud began by probing his own mind. Intrigued by his conflicted feelings toward his late father, the scientist analyzed his own dreams, slips of the tongue, childhood memories, and episodes of forgetfulness.

Freud's efforts culminated in the 1900 publication of The Interpretation of Dreams. In that book, he depicted dreams as symbolic stories in which sleepers' unconscious sexual and aggressive desires play out in disguised forms.

Later in his life, Freud acknowledged that dreams don't always gratify wishes. For instance, he noted that some

SIMILAR WORD PAIRS:

shortly, midnight

rem, sleep

virginia, polytechnic

brain, body

falling, asleep

sleep, contrast

slept, nightly

people, slept

historian, roger

remember, dreams

looks, like

brain, activity

school, medicine

studies, indicate

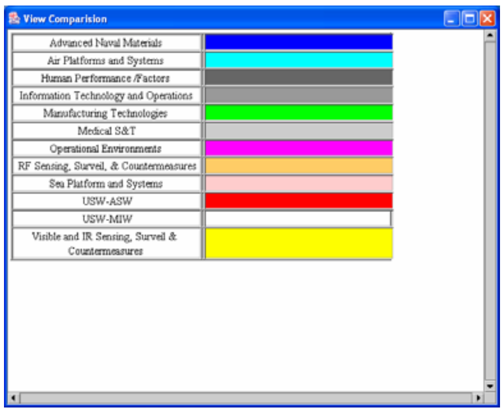
Java Applet Window

MST-Based Divisive Clustering Results on the ONR ILIR Data

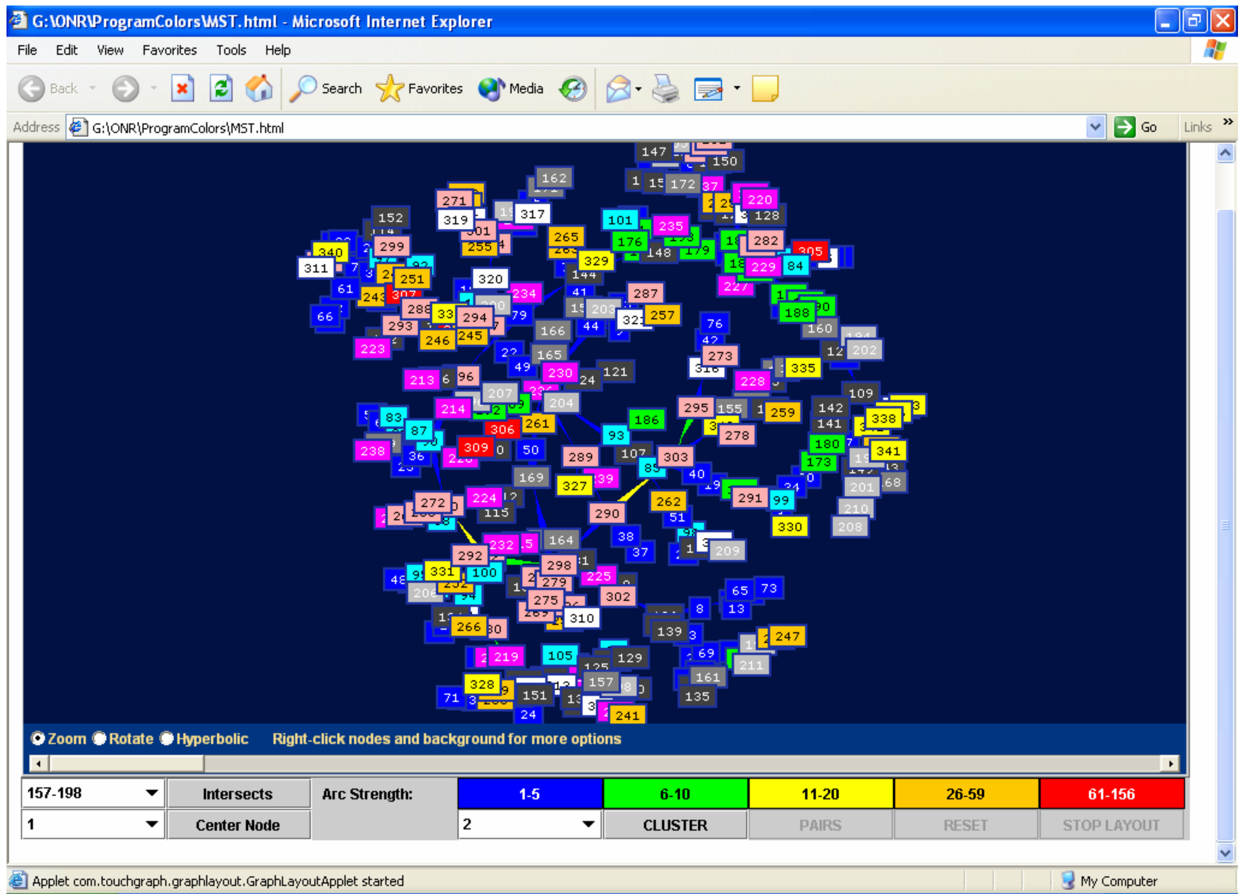
Options on the Clustering Program

- Decision to cut at an edge is determine by the the edge strength/(mean of associated edges of path length k). Choose the largest value
- Nuisance parameters
  - Maximum number of clusters
  - Minimum of points per cluster
  - k

Opening Screen for ILIR Cluster Program

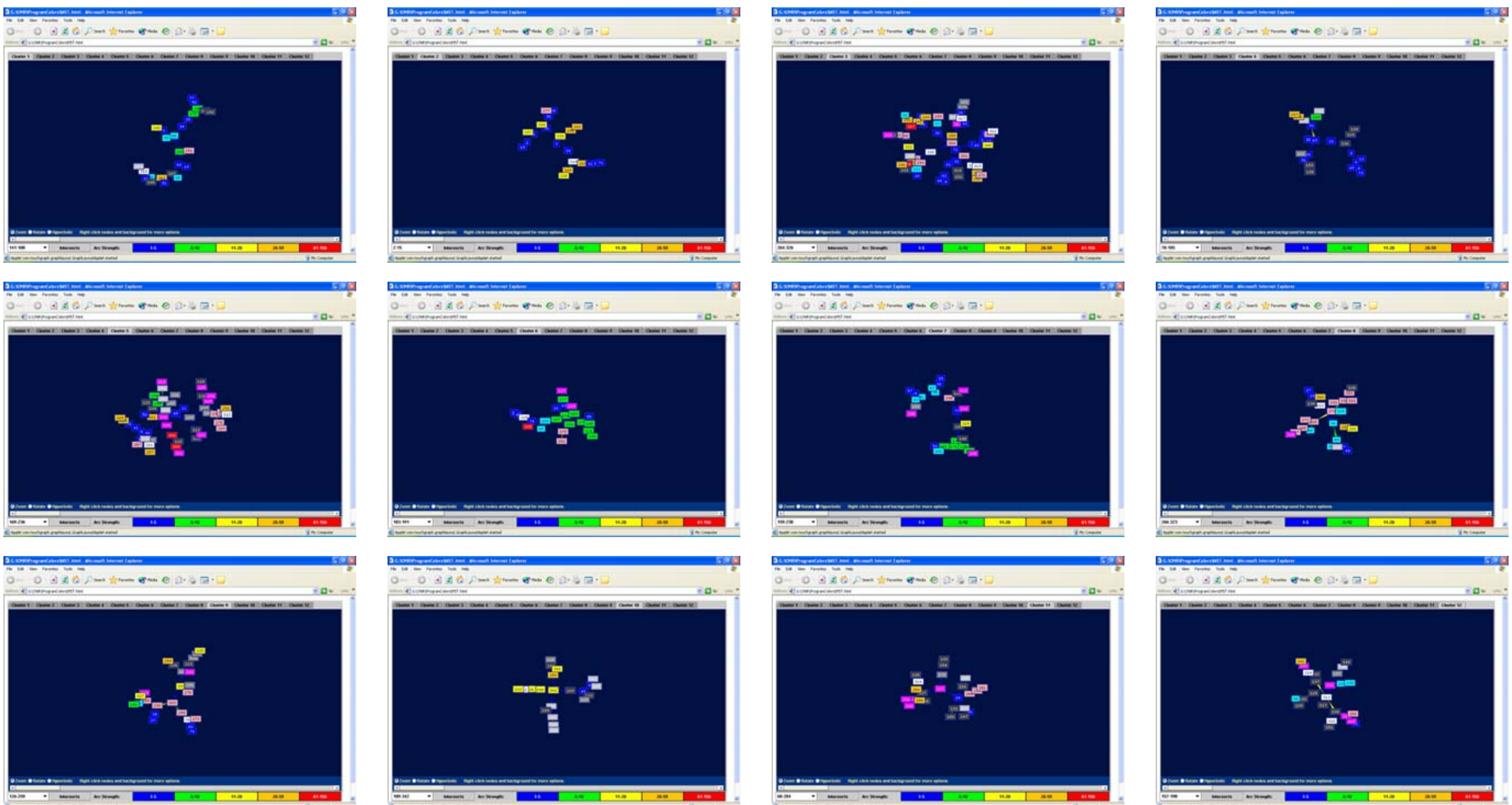


Associated
Color
Key

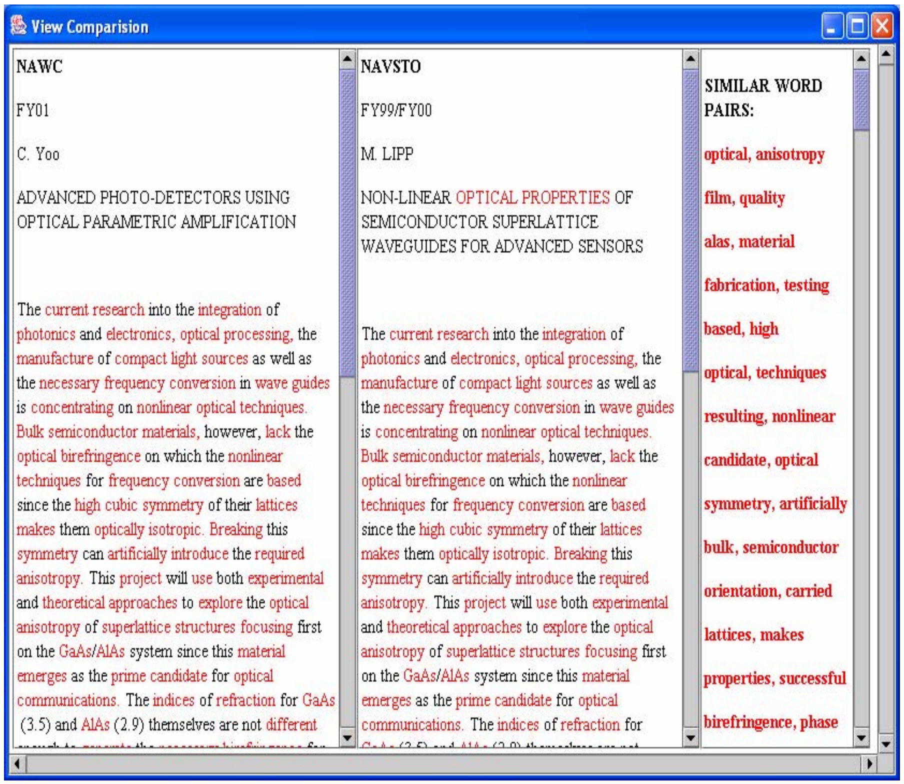


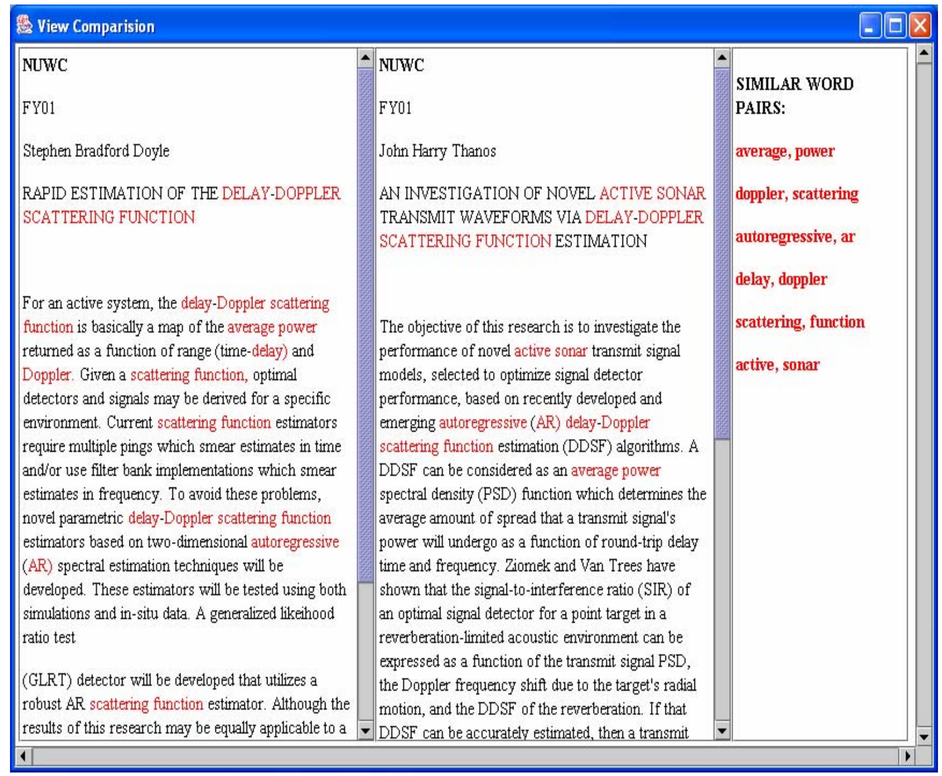
Overview of the ILIR MST Cluster Structure

Advanced Naval Materials	
Air Platforms and Systems	
Human Performance /Factors	
Information Technology and Operations	
Manufacturing Technologies	
Medical S&T	
Operational Environments	
RF Sensing, Surveil, & Countermeasures	
Sea Platform and Systems	
USW-ASW	
USW-MIW	
Visible and IR Sensing, Surveil & Countermeasures	



Some Additional Interesting Anomalies/Discoveries in the ONR ILIR Data Made Apparent Via User Exploration of the Clusters - I

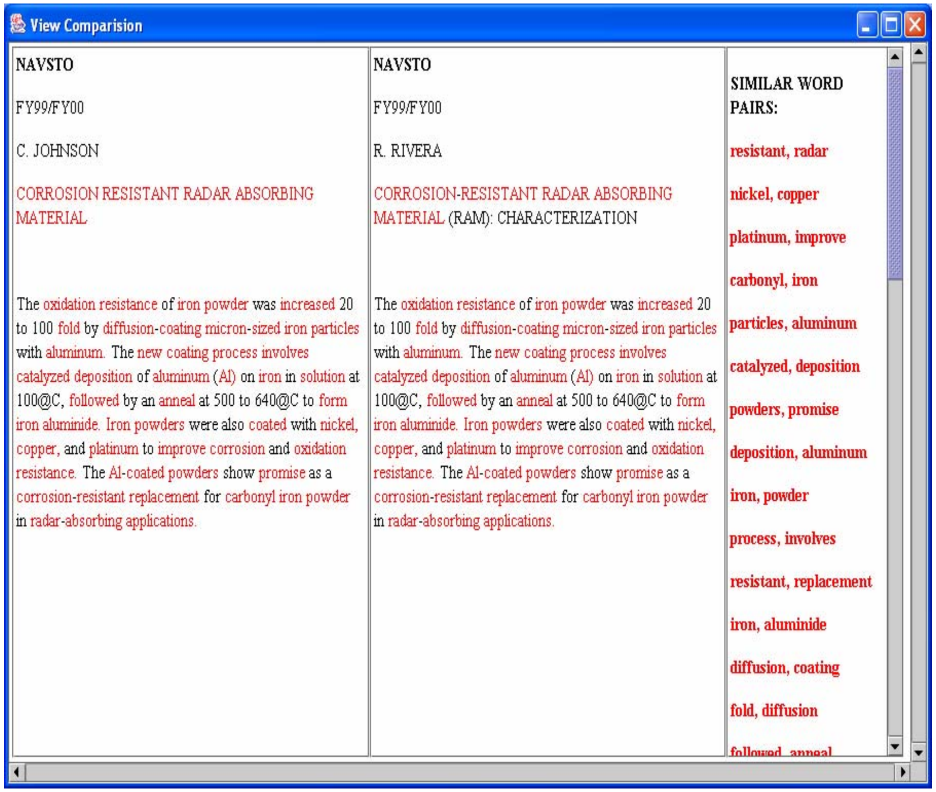


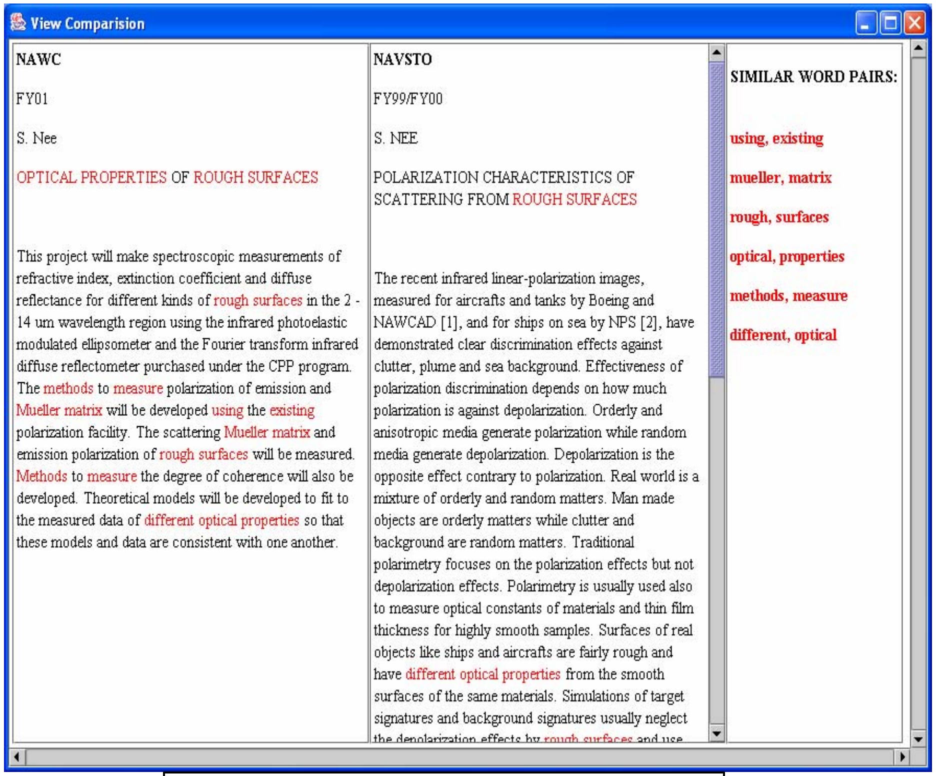


Identical Abstracts Different Author and Titles Same Year

Interesting Association Between the USW-MIW and RF Sensing, Surveillance, & Countermeasures Classes

Some Additional Interesting Anomalies/Discoveries in the ONR ILIR Data Made Apparent Via User Exploration of the Clusters - II



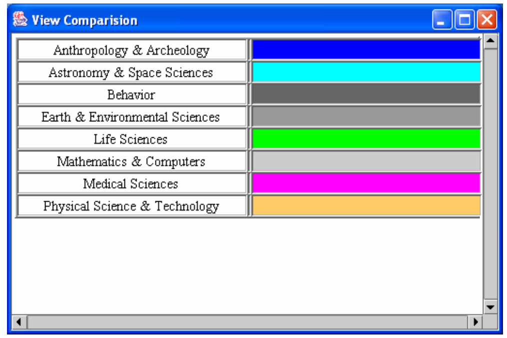


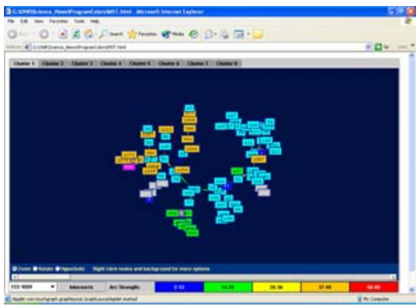
Identical Abstracts Different Author and Titles Same Year

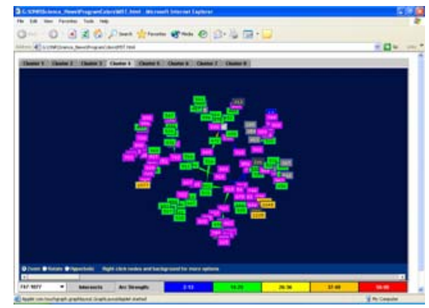
Interesting Association Between the Advanced Naval Materials and Visible and IR Sensing, Surveillance & Countermeasures Categories

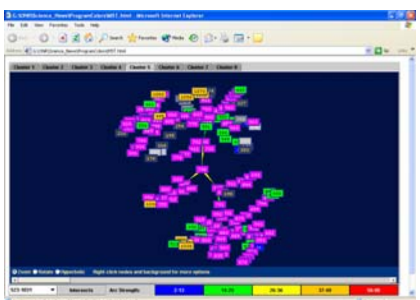
MST-Based Divisive Clustering Results on the Science News Data

Overview of the Science News MST Cluster Structure



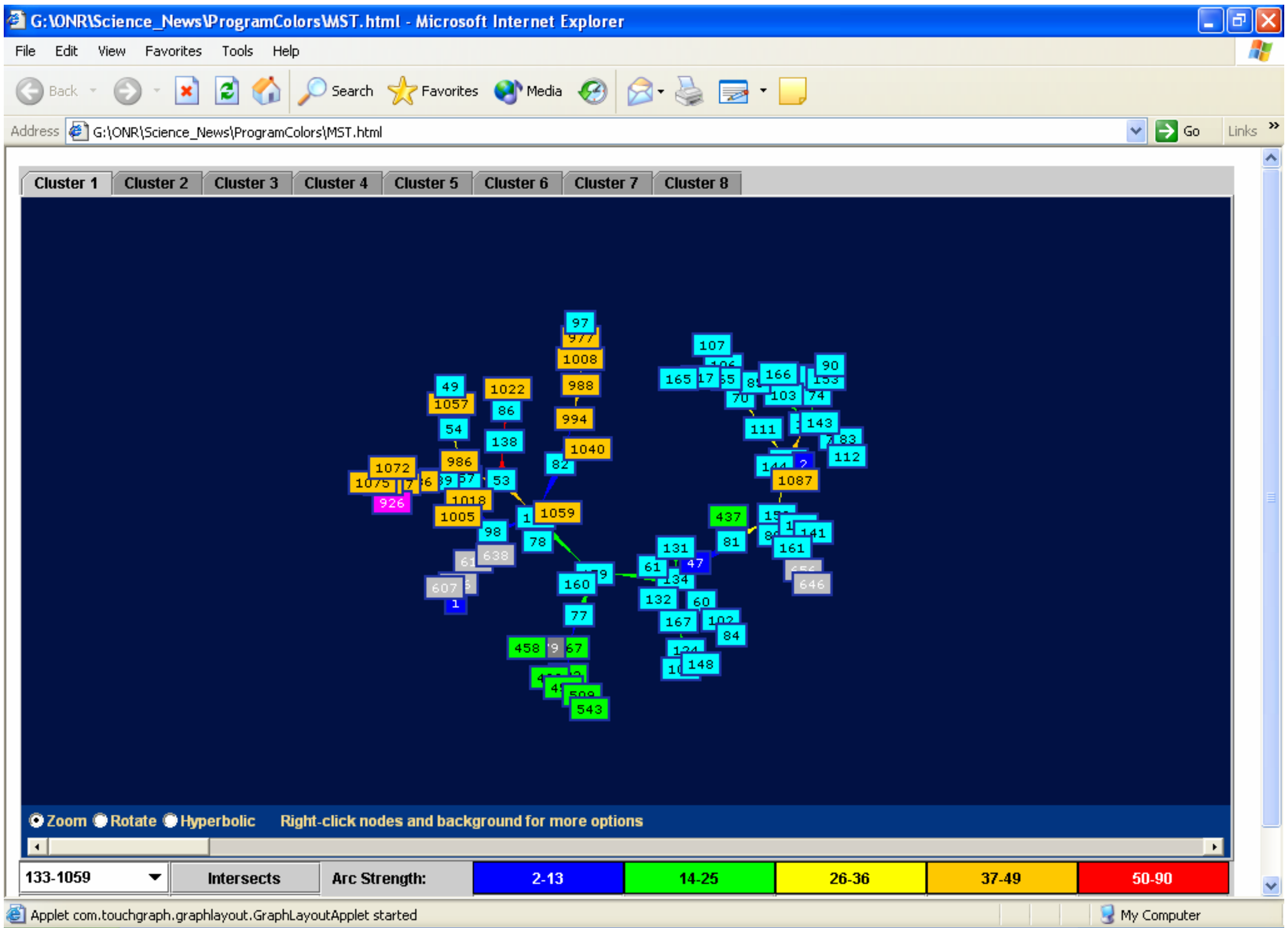






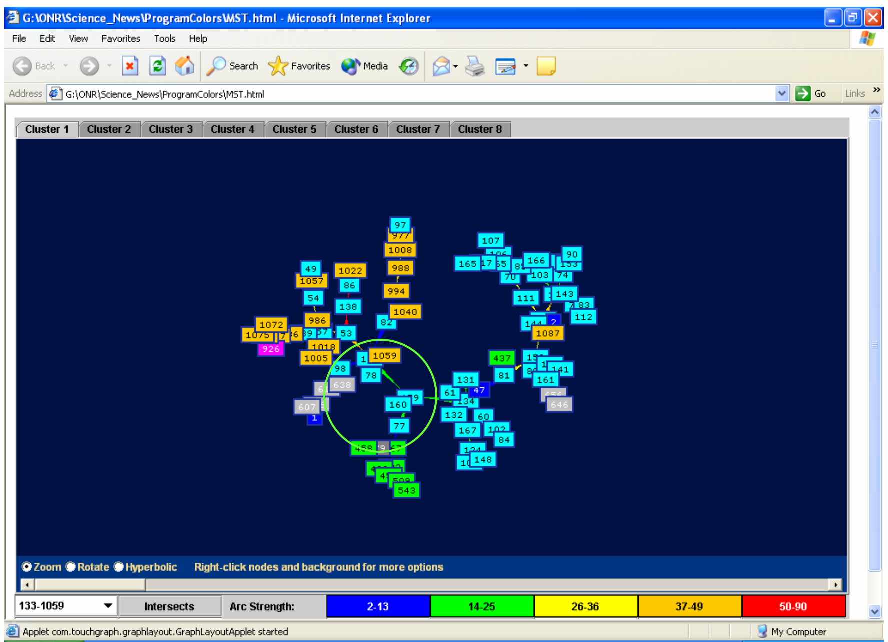
Exploration of Science News MST Cluster 1

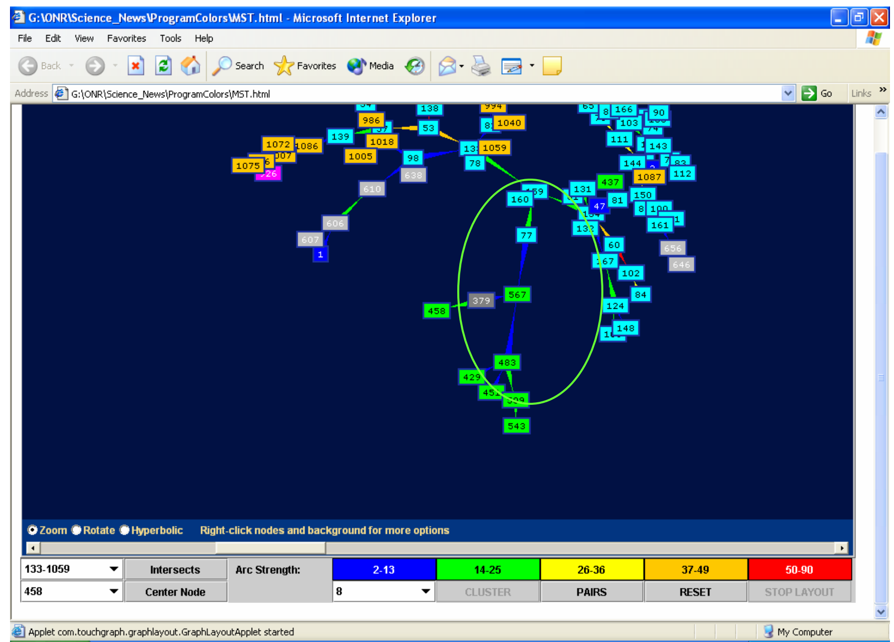
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	



Science News MST Cluster 1 Subcluster – Animal Behavior and Sexuality

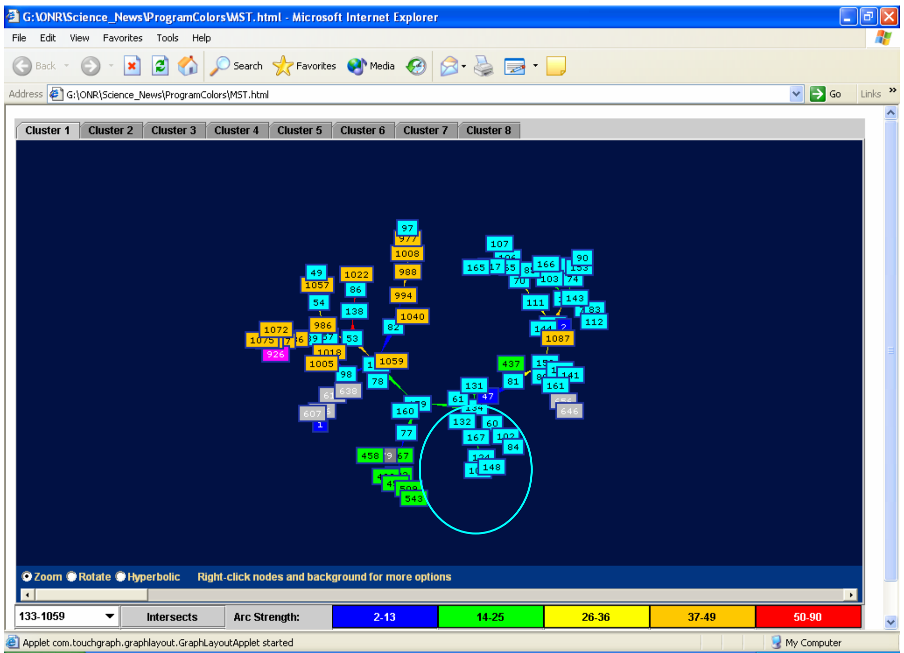
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

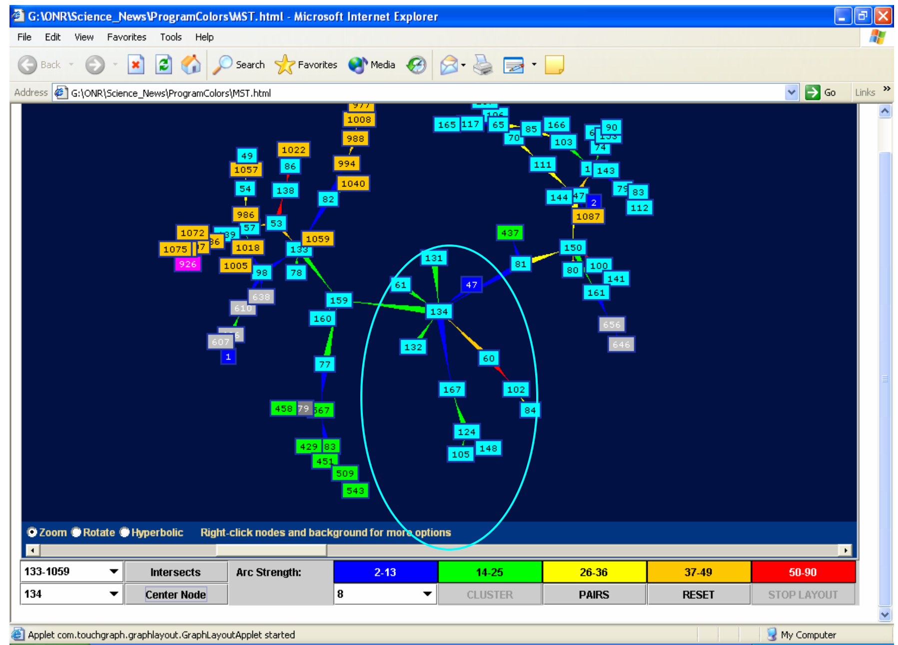




Science News MST Cluster 1 Subcluster – Infrared Camera and Its Applications to Cosmology

Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

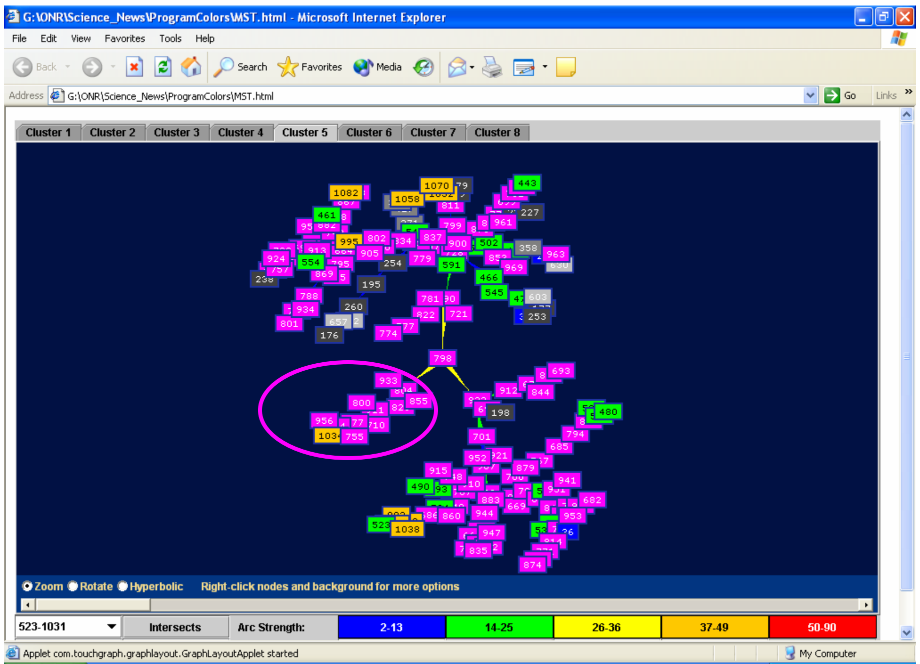


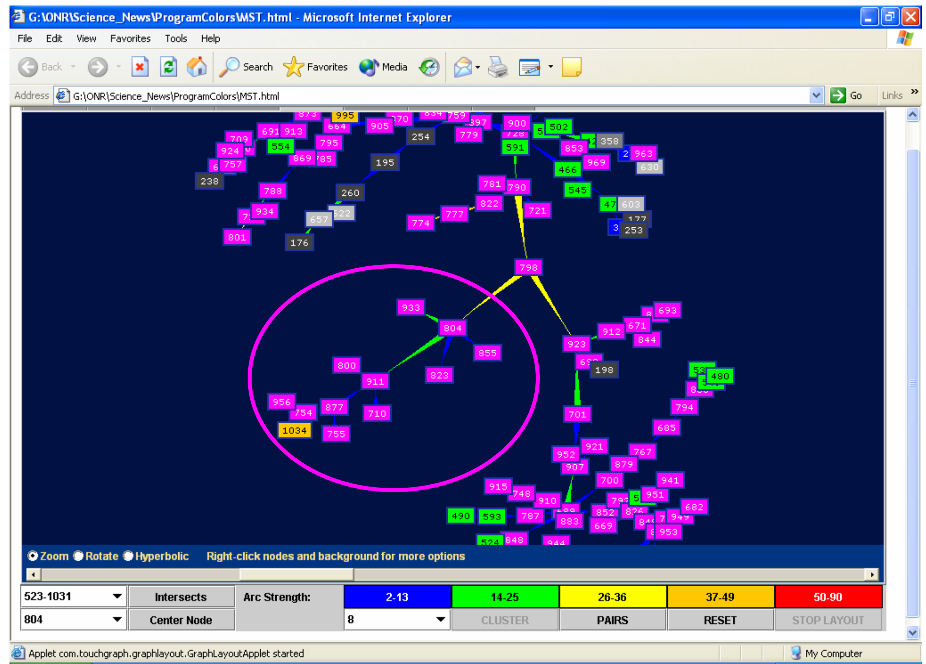


Article 134 Discusses an Enabling Technology "Infrared Camera Goes the Distance"

Science News MST Cluster 5 Subcluster - Aids

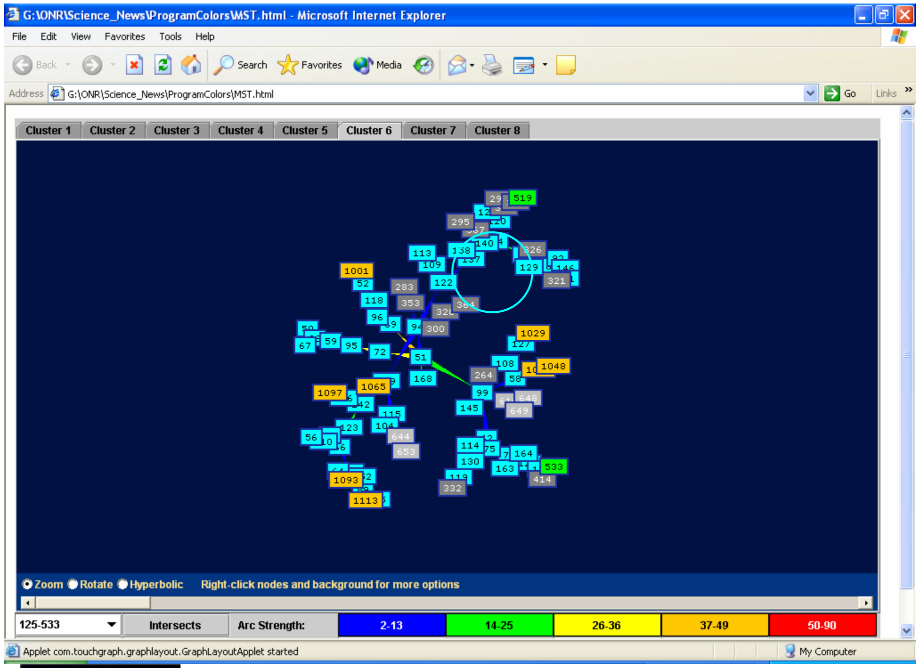
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	

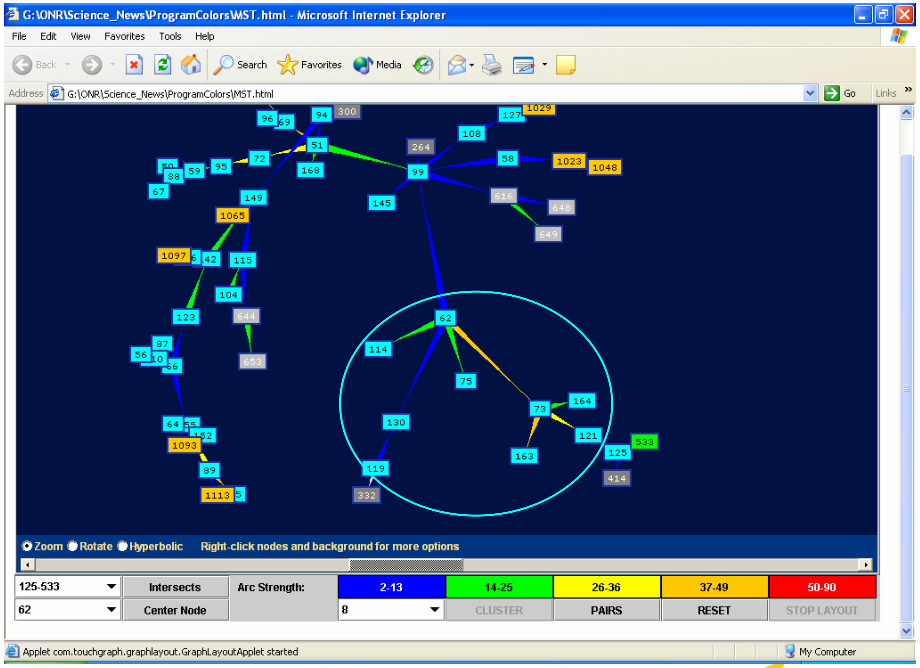




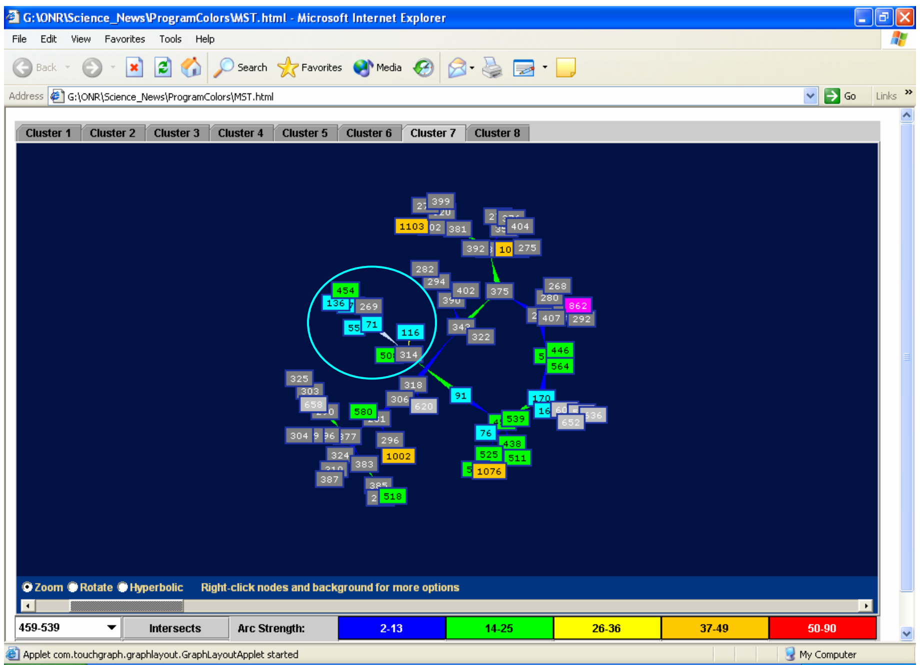
Science News MST Cluster 6 Subcluster – Solar Activity

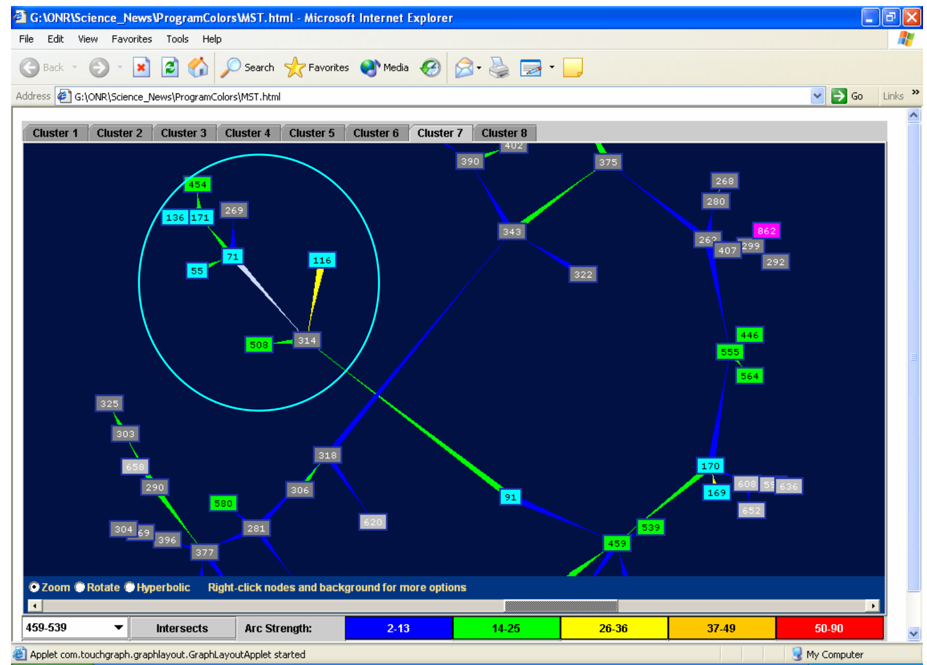
Anthropology & Archeology	
Astronomy & Space Sciences	
Behavior	
Earth & Environmental Sciences	
Life Sciences	
Mathematics & Computers	
Medical Sciences	
Physical Science & Technology	





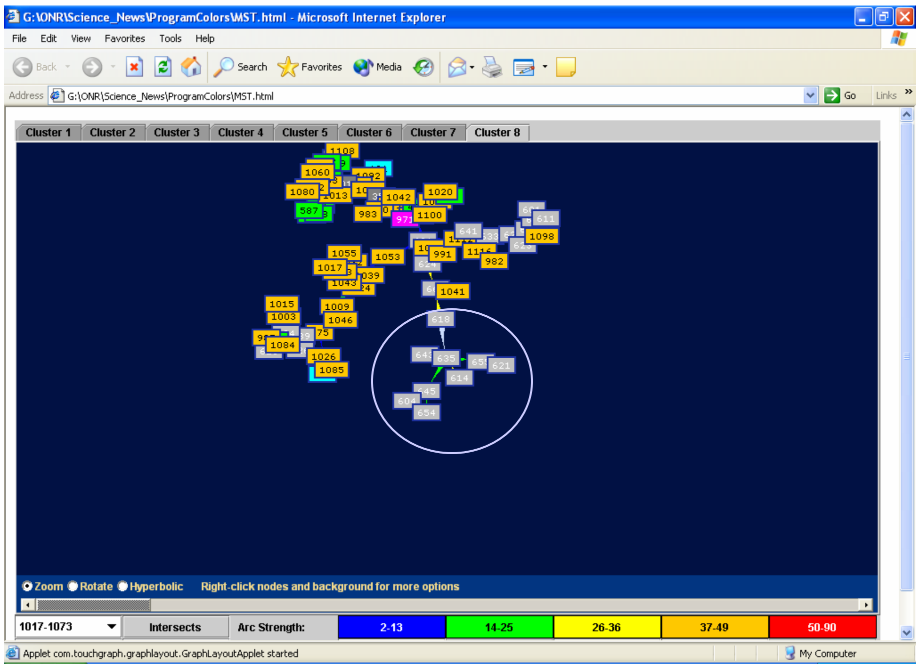
Science News MST Cluster 7 Subcluster – Evolution and the Origins of Life

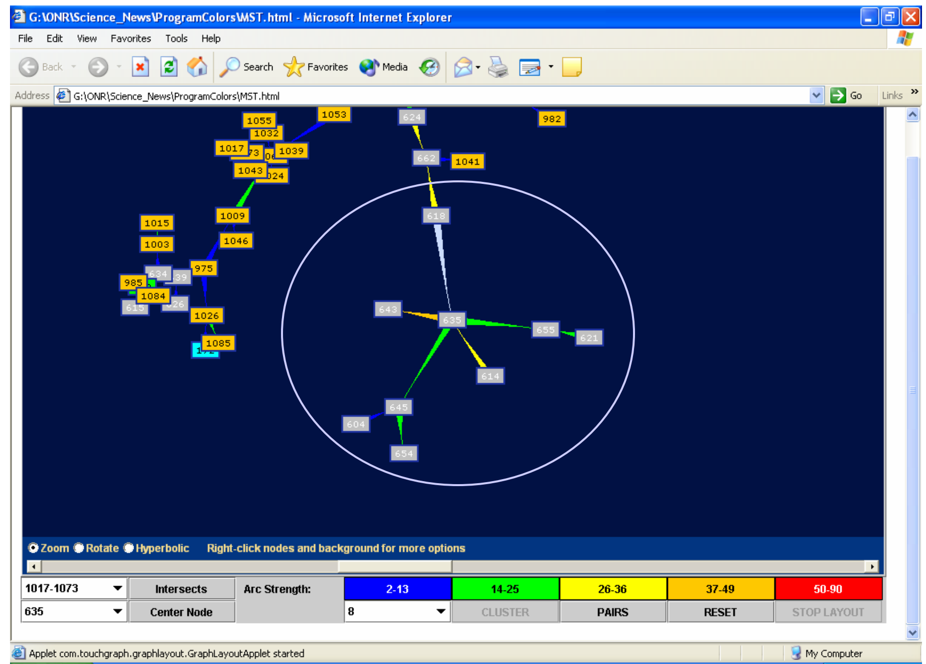




Note that cluster 7 has been rendered using a slightly different solution to the spring equations than was originally presented.

Science News MST Cluster 8 Subcluster – Artificial Intelligence





Note that cluster 8 has been rendered using a slightly different solution to the spring equations then was originally presented.

Agglomerative Clustering Results on the Science News Data

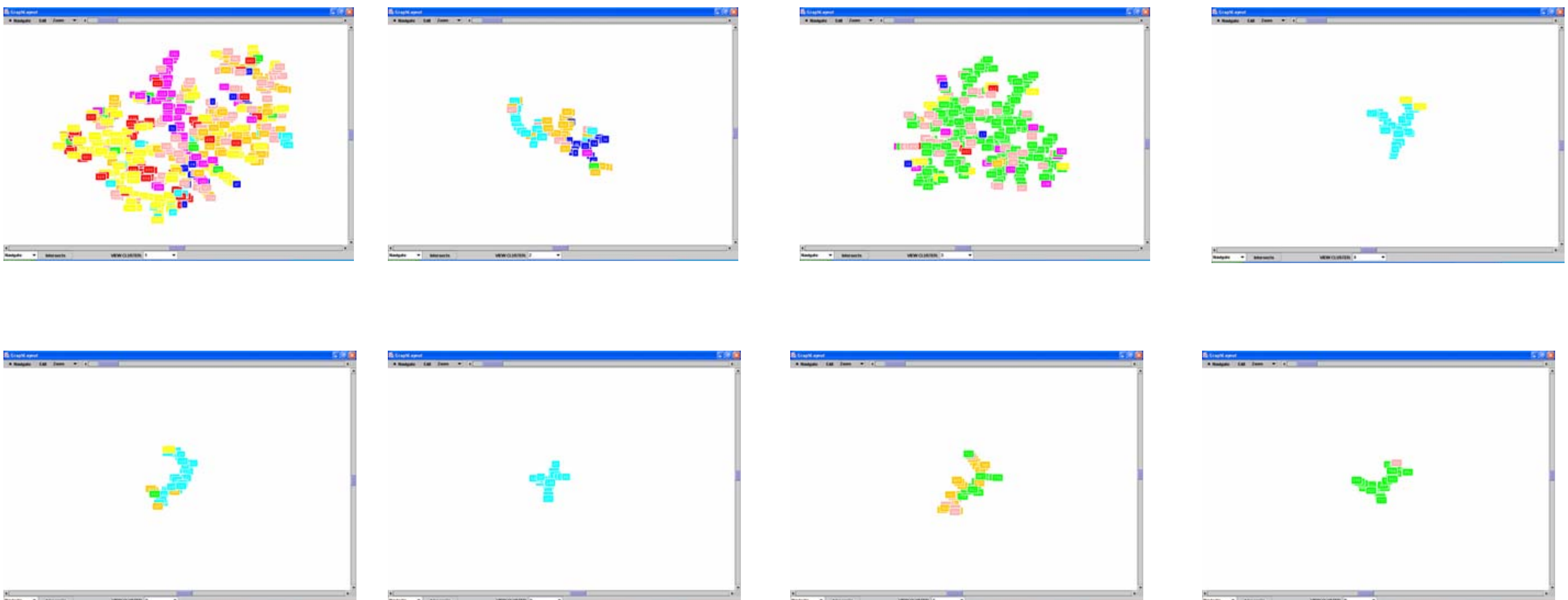
Methodology

- ScienceNews1:
  - 8 classification categories
  - Hierarchical Clustering
  - Method: Ward
    - Merge two clusters that produce the smallest variance in resultant cluster.
  - Tree Cut: 8

Science News 8 Agglomerative Clusters

Anthropology & Archeology
Behavior
Life Sciences
Medical Sciences

Astronomy & Space Sciences
Earth & Environmental Sciences
Mathematics & Computers
Physical Science & Technology



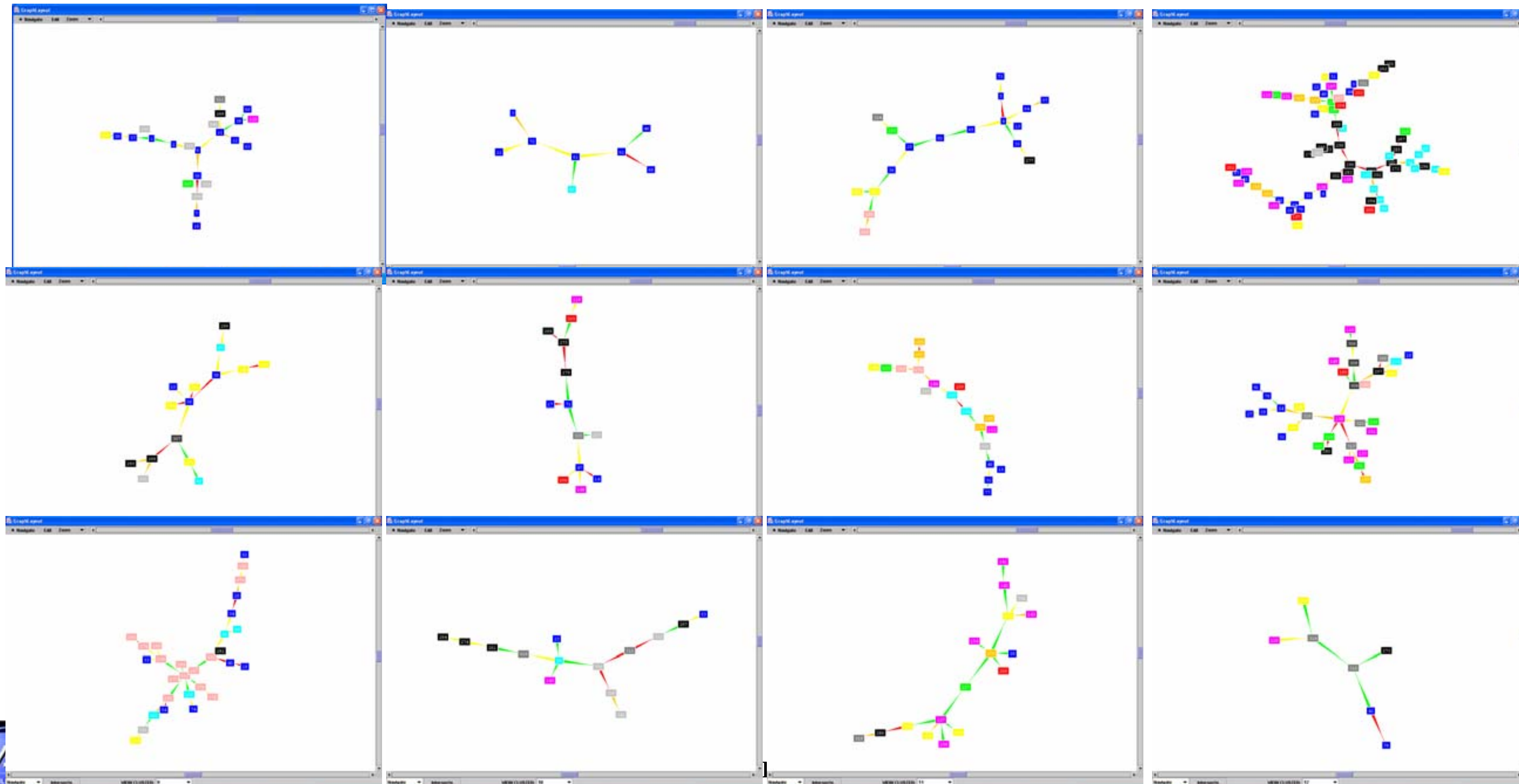
Agglomerative Clustering Results on the ONR ILIR Data

Methodology

- ILIR1:
  - 12 classification categories
  - Hierarchical Clustering
  - Method: Average
    - Merge clusters with smallest average distance.
  - Tree Cut: 24

ILIR 24 Agglomerative Clusters

Advanced Naval Materials
Information Technology and Operations
Operational Environments
USW-ASW
Air Platforms and Systems
Manufacturing Technologies
RF Sensing, Surveil, & Countermeasures
USW-MIW
Human Performance /Factors
Medical S&T
Sea Platform and Systems
Visible and IR Sensing, Surveil & Countermeasures

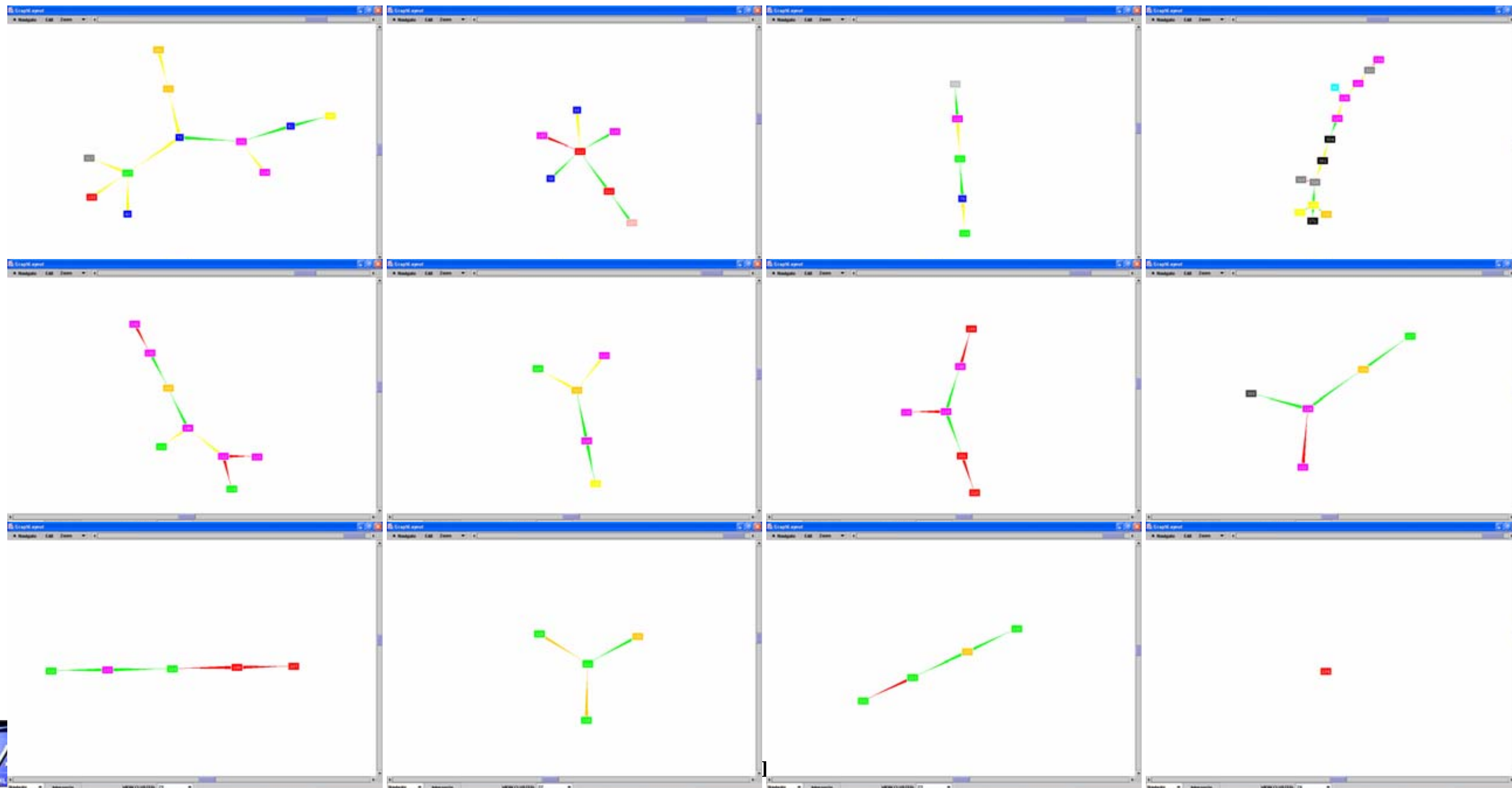


ILIR 24 Agglomerative Clusters

Advanced Naval Materials
Information Technology and Operations
Operational Environments
USW-ASW

Air Platforms and Systems
Manufacturing Technologies
RF Sensing, Surveil, & Countermeasures
USW-MIW

Human Performance /Factors
Medical S&T
Sea Platform and Systems
Visible and IR Sensing, Surveil & Countermeasures



Bipartite Spectral Based Results