Visualização de Informação
Parte IV

Multi-dimensional Visualization: Detalhamento Ávores de Similaridade e Desenvolvimento Atual/Futuro

Rosane Minghim +
The team

Instituto de Ciências Matemáticas e de Computação
USP-São Carlos

Baseado em:
Visual Analysis of Metric and Multidimensional Data

M. Cristina F. Oliveira
Rosane Minghim
Agma Traina

Instituto de Ciências Matemáticas e de Computação
University of São Paulo, São Carlos

Julho 2010 IME – USP São Paulo
São Paulo Advanced School of Computing
Image Processing and Visualization

Part 1

Visual analysis of high-dimensional data

M. Cristina F. Oliveira
Rosane Minghim

LCAD - High Performance Computing Laboratory
VICG – Visualization, Imaging and Computer Graphics Group

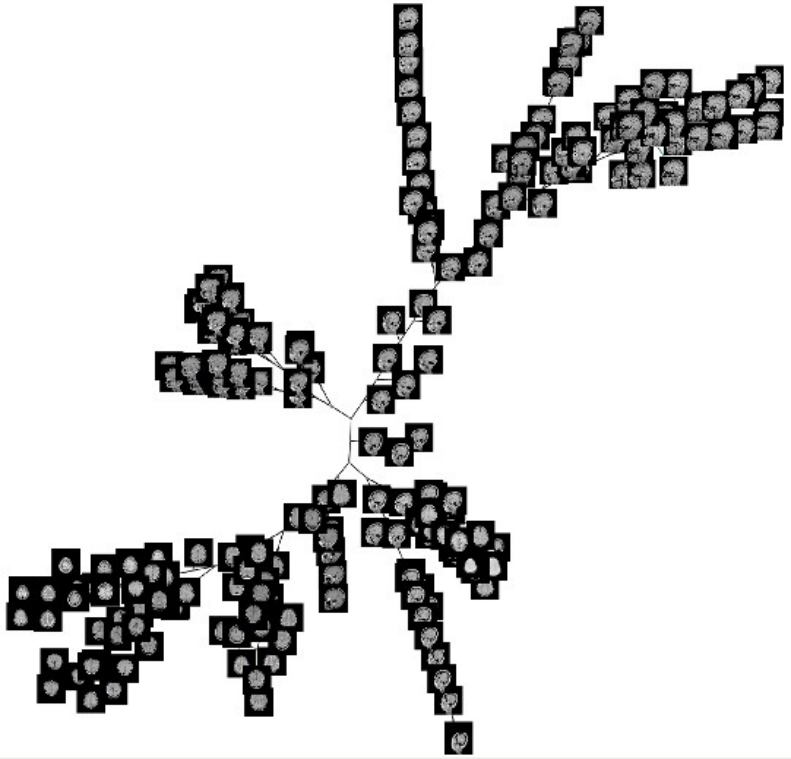
ICMC – University of São Paulo, São Carlos

Nomenclature and concepts review

- A data point, data item or sample:
  - An individual in a data set
- A data point is (usually) defined by m attributes (or dimensions)
- The goal is to build a mapping on 2D (or 3D) that reflects similarity amongst items
- Distance (or similarity) between data points is at the core of the visual mapping strategy

Similarity Trees

- From a distance relationship:
  1. Imposes a hierarchy creating a t structure
  2. Generates a tree layout



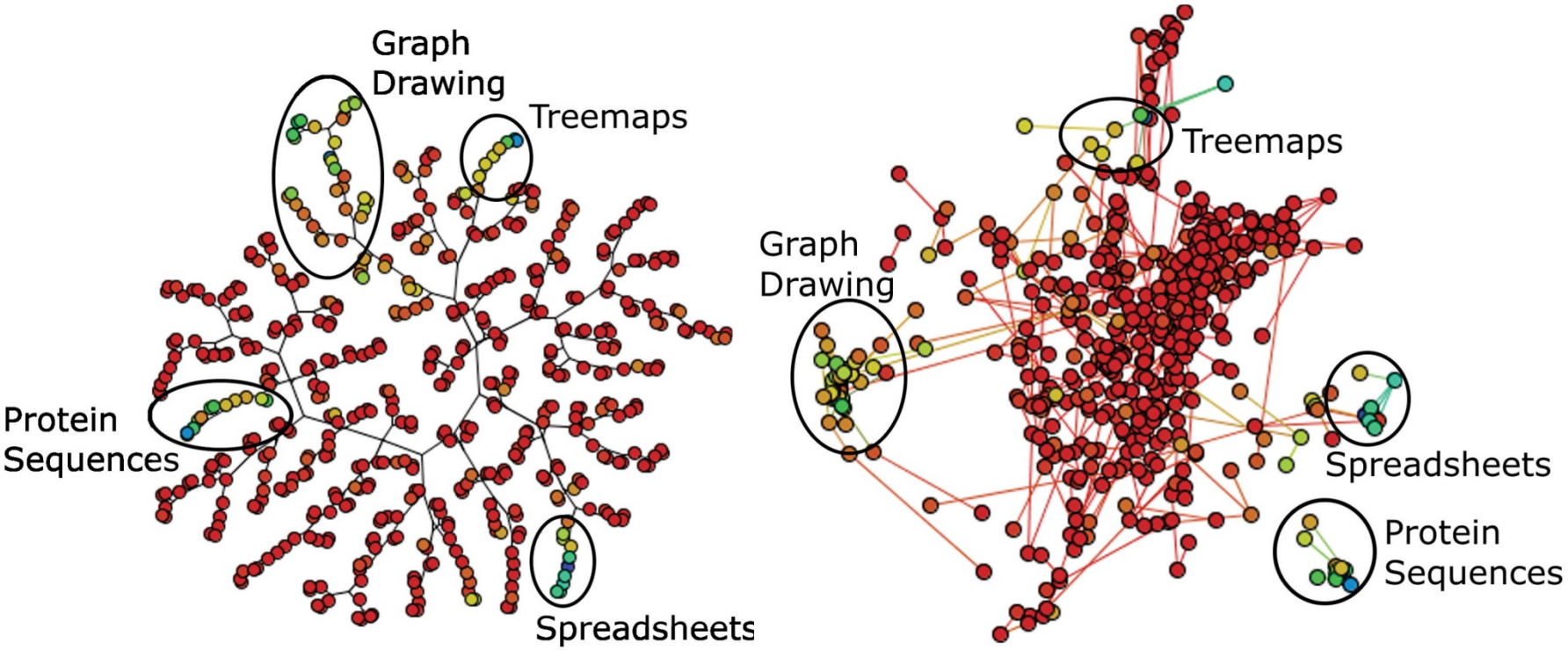
- Interpretation is subject to branch organization

- Offers multi-level views of data

Example

- **Mapping data sets for NJ and projection techniques**

2004 IEEE Infovis Contest, 515 files, Scientific papers on InfoVis

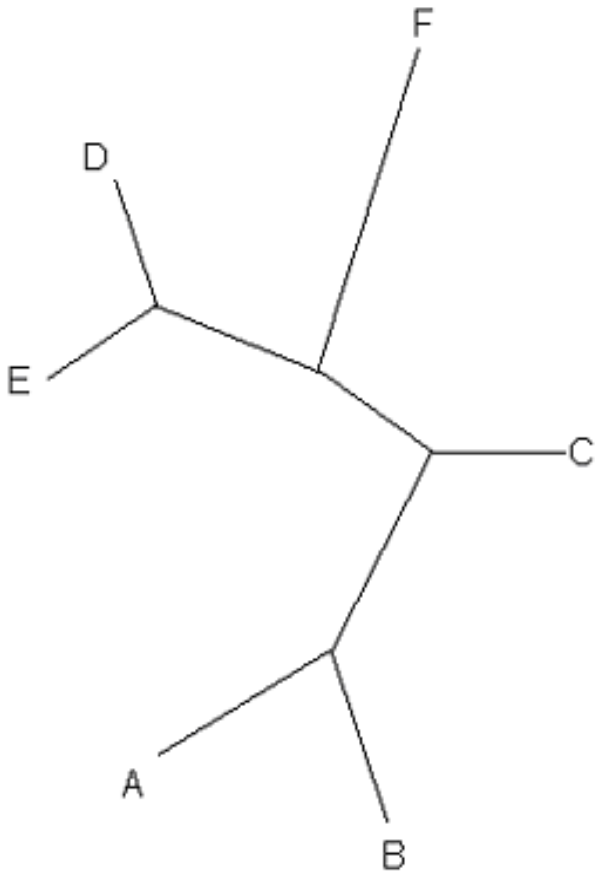


NJ **LSP**

1 – Hierarchy by Neighbor-Joining (NJ)

[10] Cuadros, Paulovich, Minghim, Telles, IEEE VAST 2007

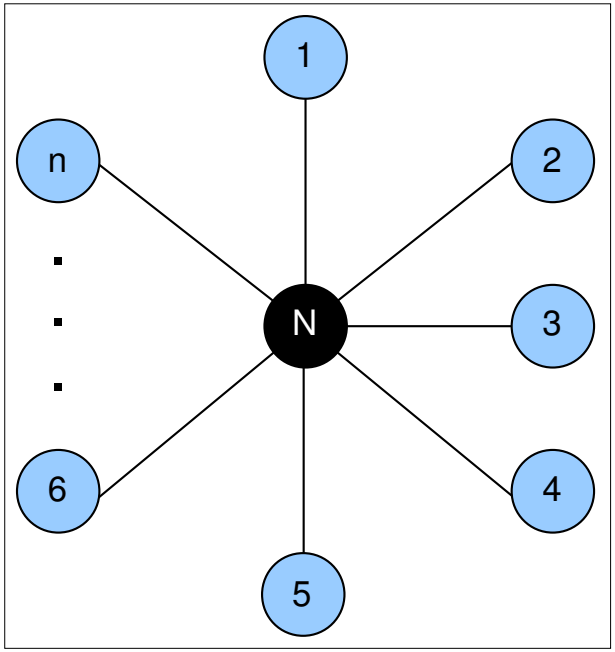
- Neighbor-Joining (NJ) technique [7]:
  - Heuristic algorithm for tree construction
  - Define the tree topology and branches length
  - Builds an un-rooted tree
  - Selects the closest pair of documents and joins them into a hypothetical ancestor
  - Overall running time $O(n^3)$
  - With text:
    - Leaves: data point.
    - Internal nodes: ancestor hypothetical doc.
    - Edges' lengths: distance between docs.



Neighbor–Joining (NJ) [10]

- Starts with a start-like tree, with n leaves connected to a single internal node

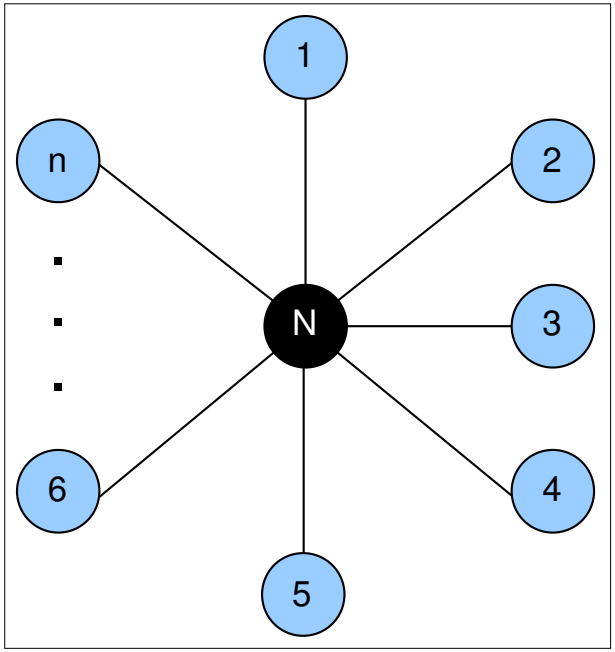
	1	2	3	4	5	6	7	...	n
1	0								
2		0							
3			0						
4				0					
5					0				
6						0			
7							0		
...								0	
n									0



1-Neighbor-Joining (NJ) [10]

- Starts with a start-like tree, with n leaves connected to a single internal node

	1	2	3	4	5	6	7	...	n
1	0								
2		0							
3			0						
4				0					
5					0				
6						0			
7							0		
...								0	
n									0

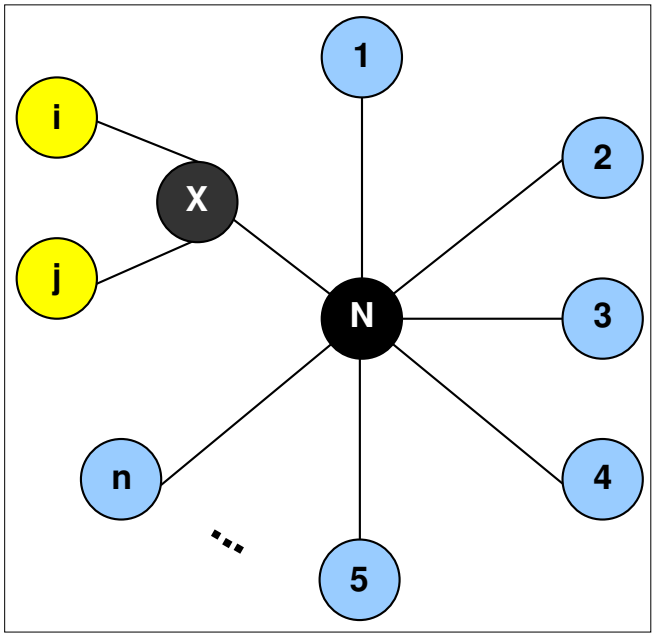


1 – Neighbor-Joining (NJ) [10]

- Selects the smallest sum of branch lengths S_{ij}

$$S_{ij} = \frac{1}{2(n-2)} \sum_{k \neq i,j} (D_{ik} + D_{jk}) + \frac{D_{ij}}{2} + \frac{1}{n-2} \sum_{k,l \neq i,j}^{k<l} D_{kl}$$

- Adds a node x to the tree, with i and j as children and connected to the common ancestor of i and j

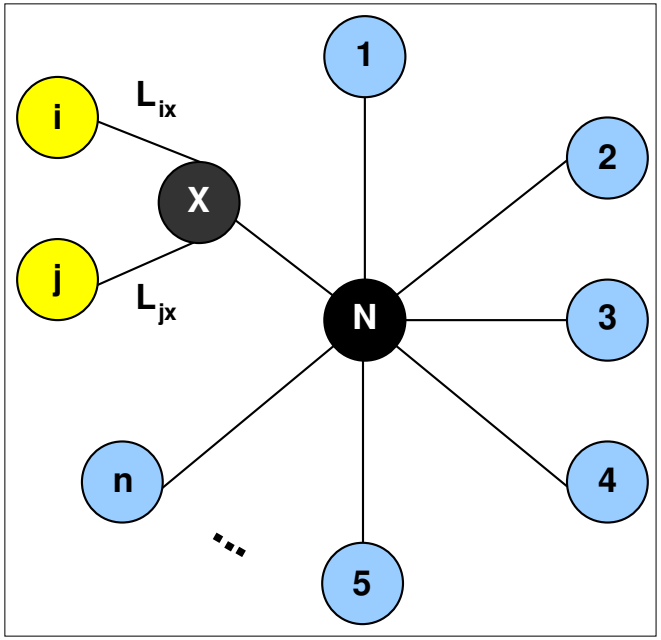


1 – Neighbor-Joining (NJ)

- Evaluates the branch lengths L_{ix} and L_{jx}

$$L_{ix} = \frac{D_{ij} + \frac{\sum_{k \neq j} D_{ik}}{n-2} - \frac{\sum_{k \neq i} D_{jk}}{n-2}}{2}$$

$$L_{jx} = \frac{D_{ij} + \frac{\sum_{k \neq i} D_{jk}}{n-2} - \frac{\sum_{k \neq j} D_{ik}}{n-2}}{2}$$



1 – Neighbor–Joining (NJ) [10]

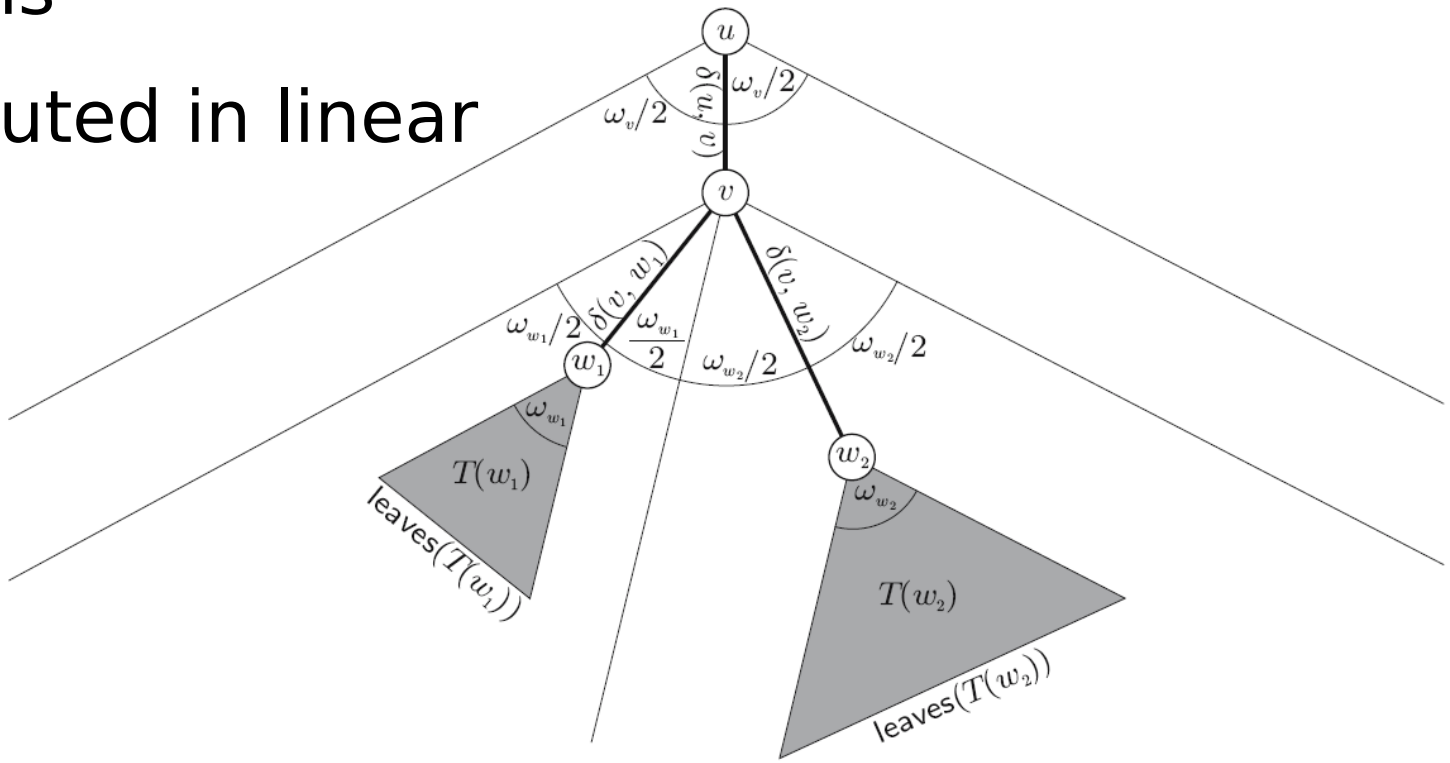
- Replaces i and j by x in the distance matrix evaluating the D_{xy} for every y in the matrix

$$D_{xy} = \frac{D_{iy} + D_{jy}}{2}$$

- Repeats this steps until there is only two nodes remaining in the matrix

2 – Display Radial layout [2]

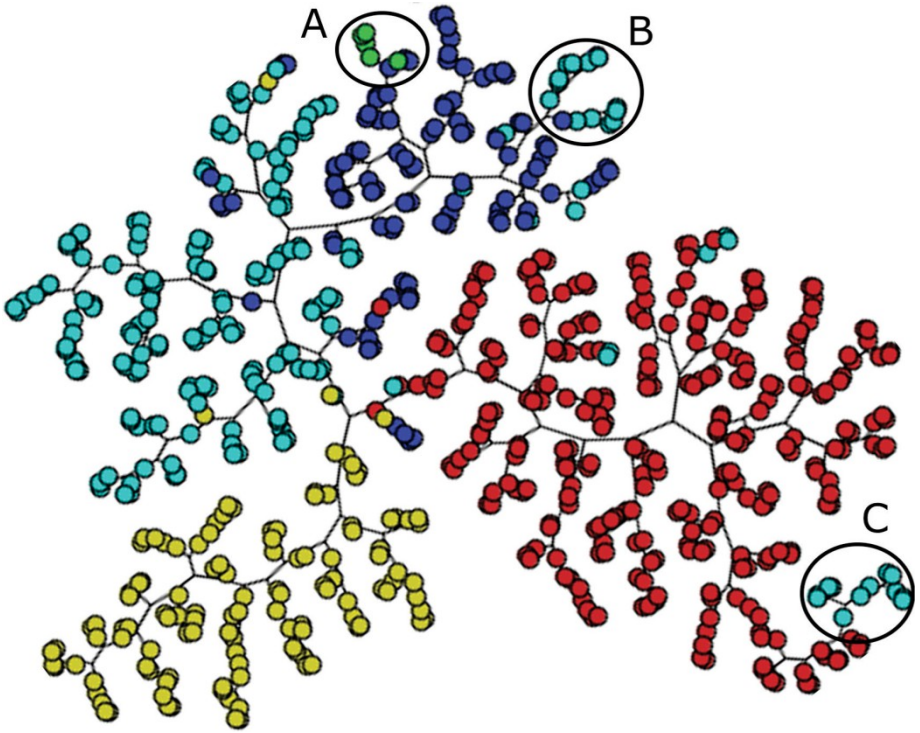
- Preserves edge lengths
- Computed in linear time



Results

- **Mapping Scientific data sets**
  - CBR+ILP+IR+SON, 680 files, Scientific papers

- **Case-based Reasoning**
- **Inductive Logic Programming**
- **Information Retrieval**
- **Sonification**

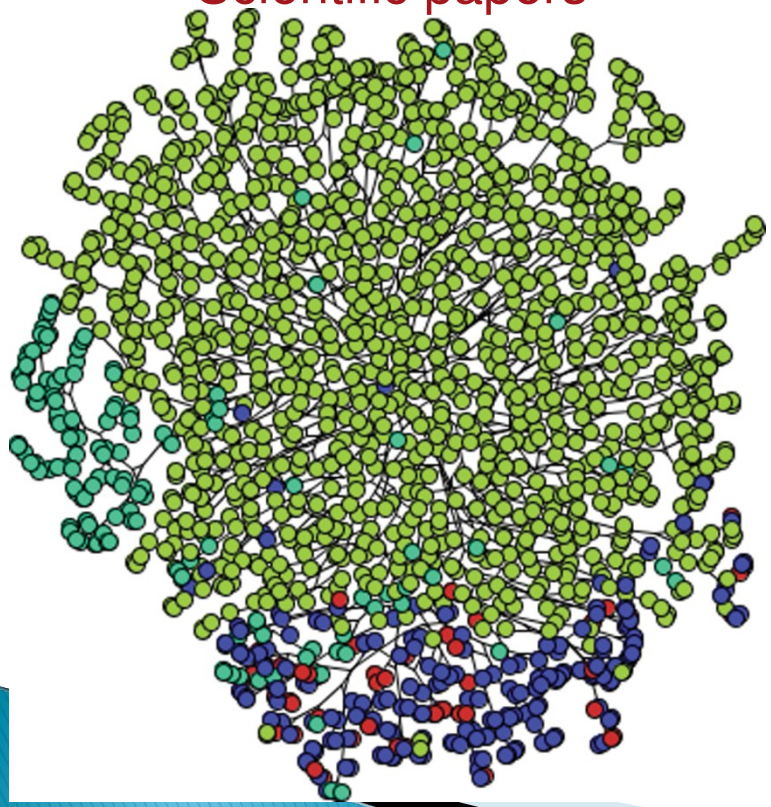


Results

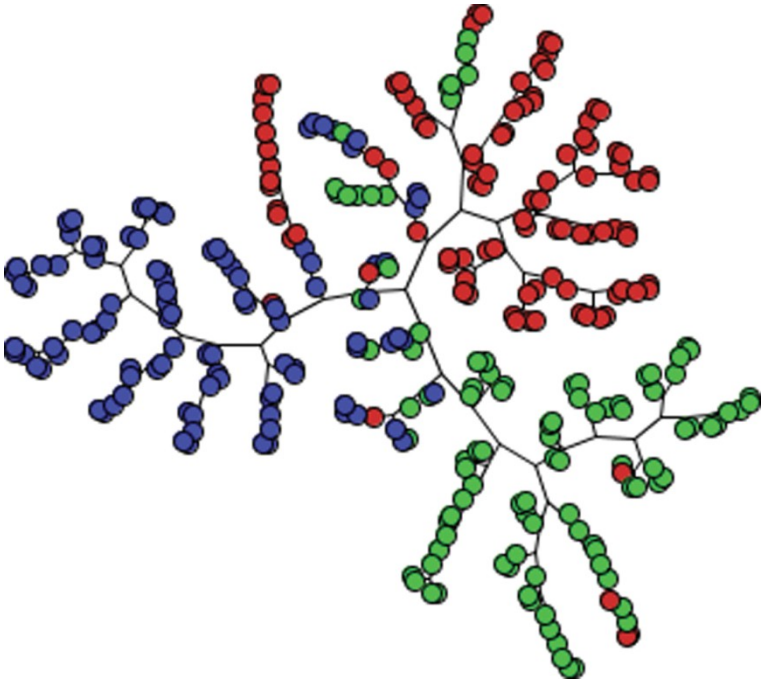
- **Mapping data sets**

  KDVis, 1,624 files,

  Scientific papers



- MESSAGES, 300 files, Discussion groups

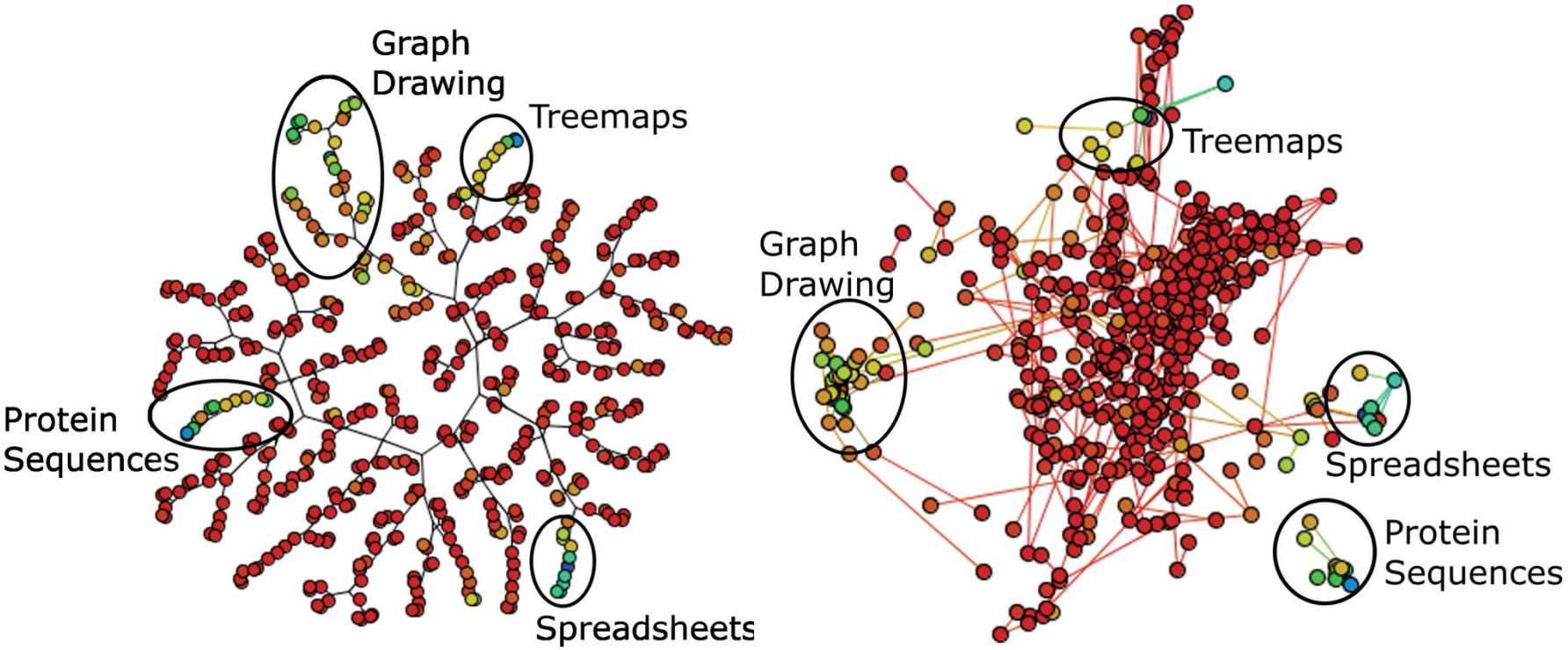


LCAD

Results

- **Mapping data sets for NJ and projection techniques**

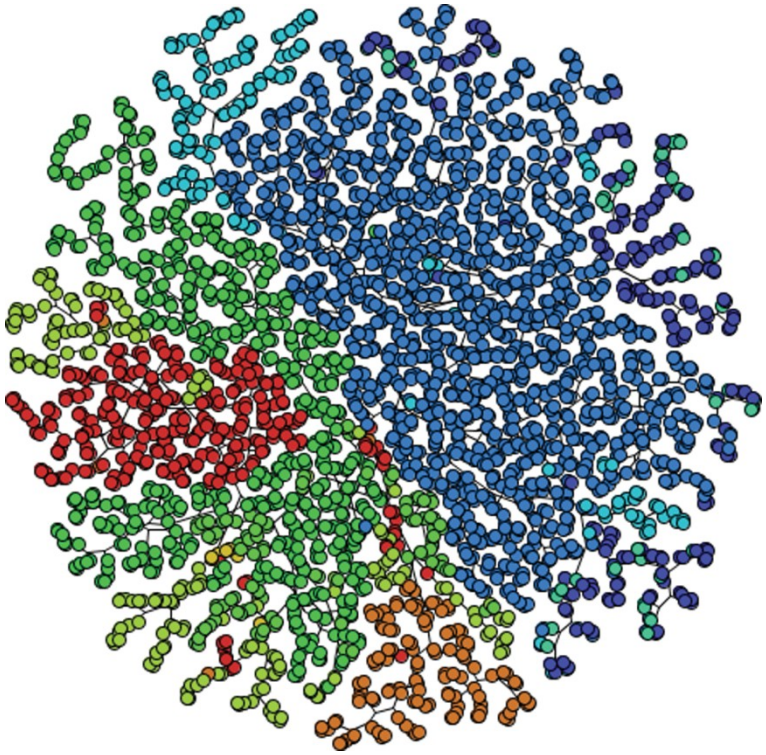
2004 IEEE Infovis Contest, 515 files, Scientific papers



NJ LSP

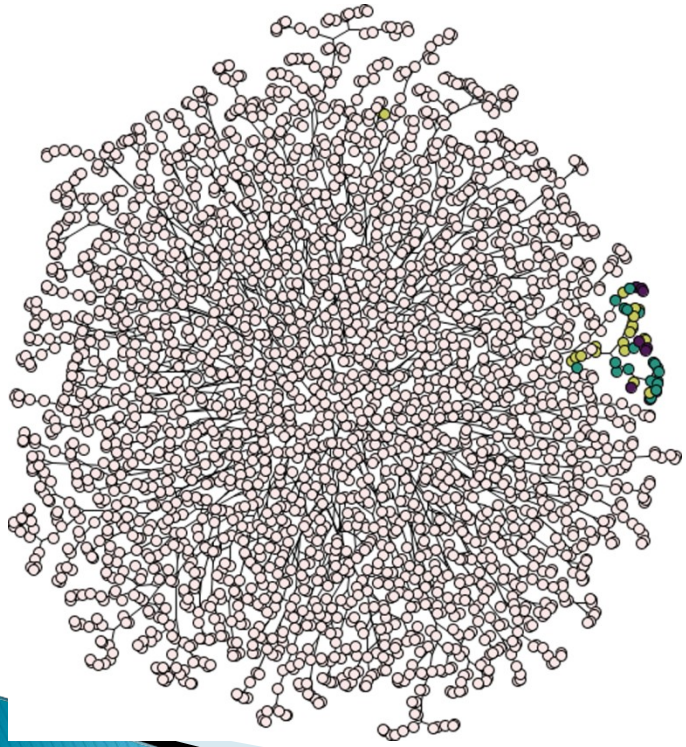
Results

- All scientific data set together using NCD similarity [8]

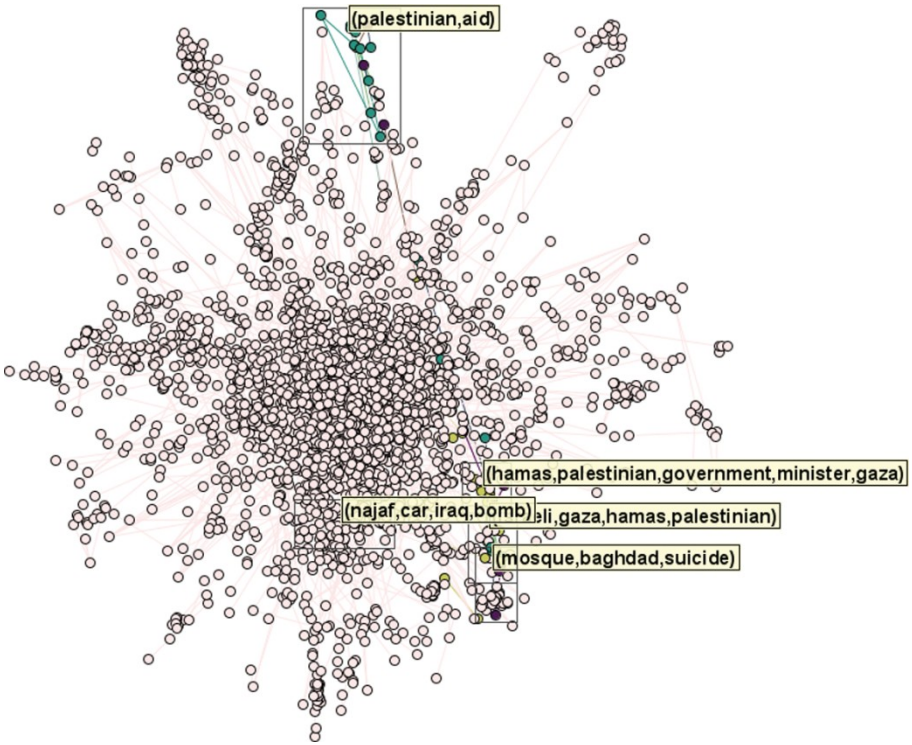


Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - NEWS, 2,684 files

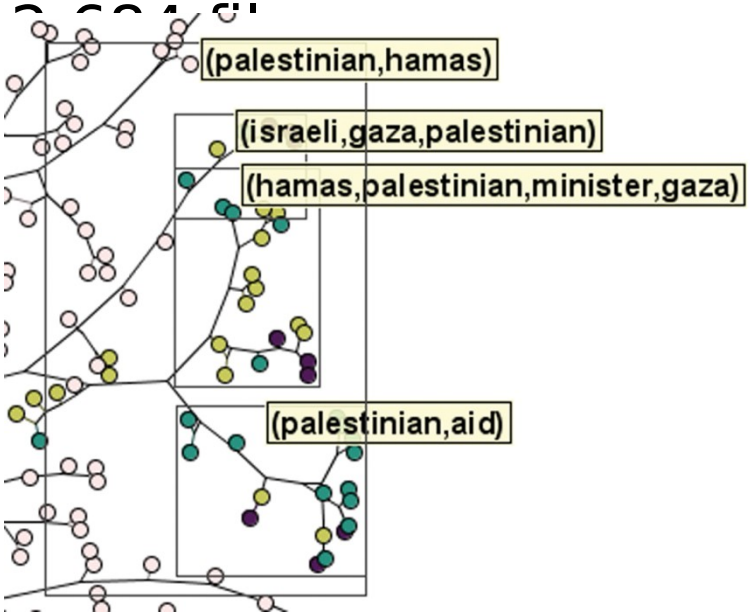


NJ



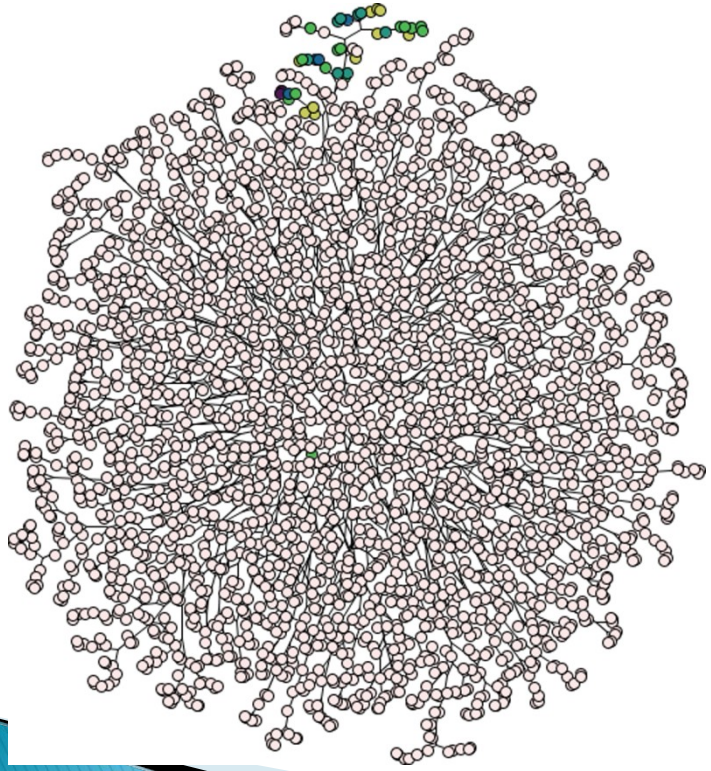
Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, [illegible]

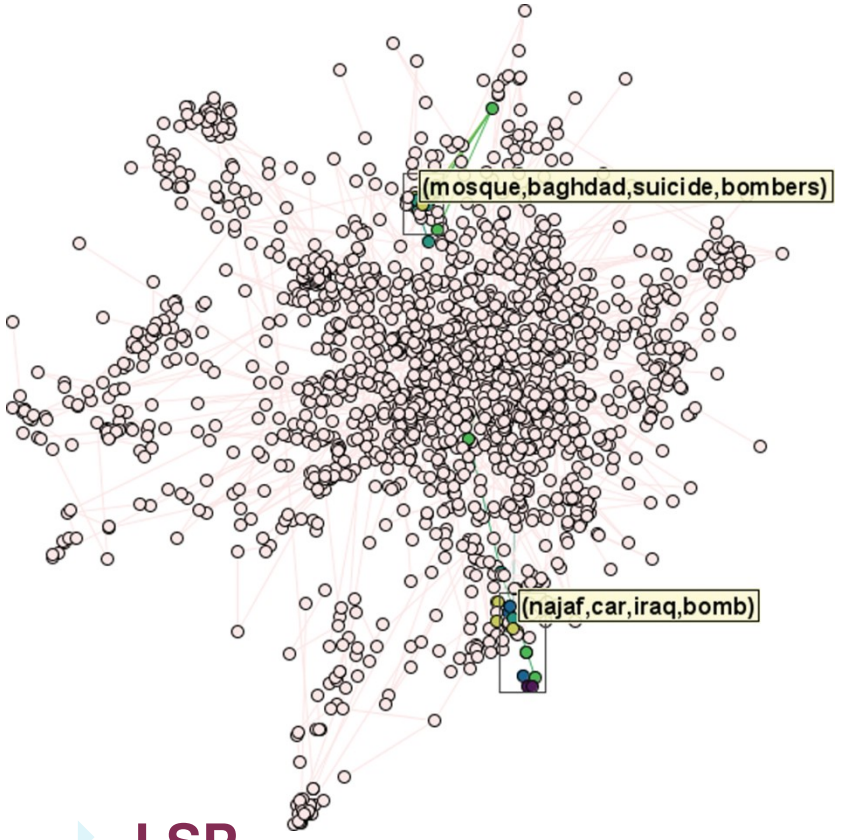


Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, 2,684 files



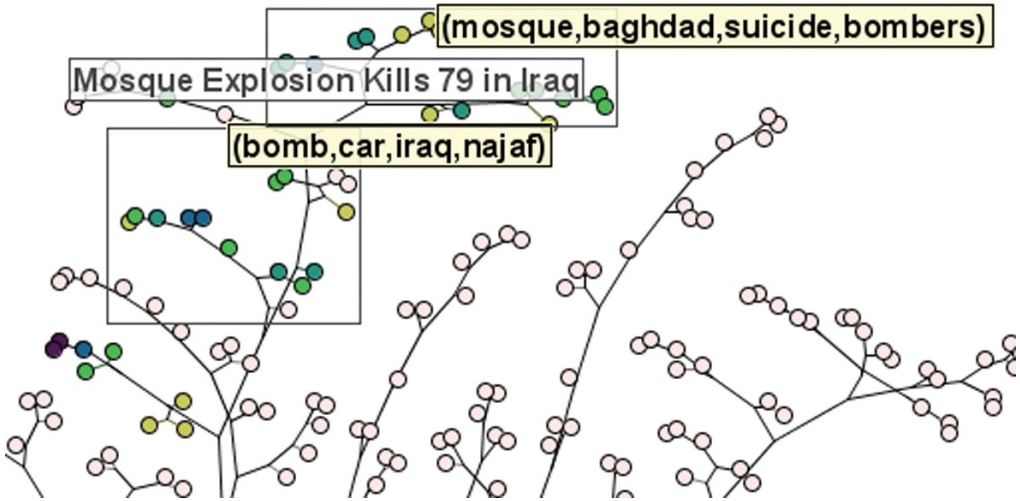
NJ



LSP

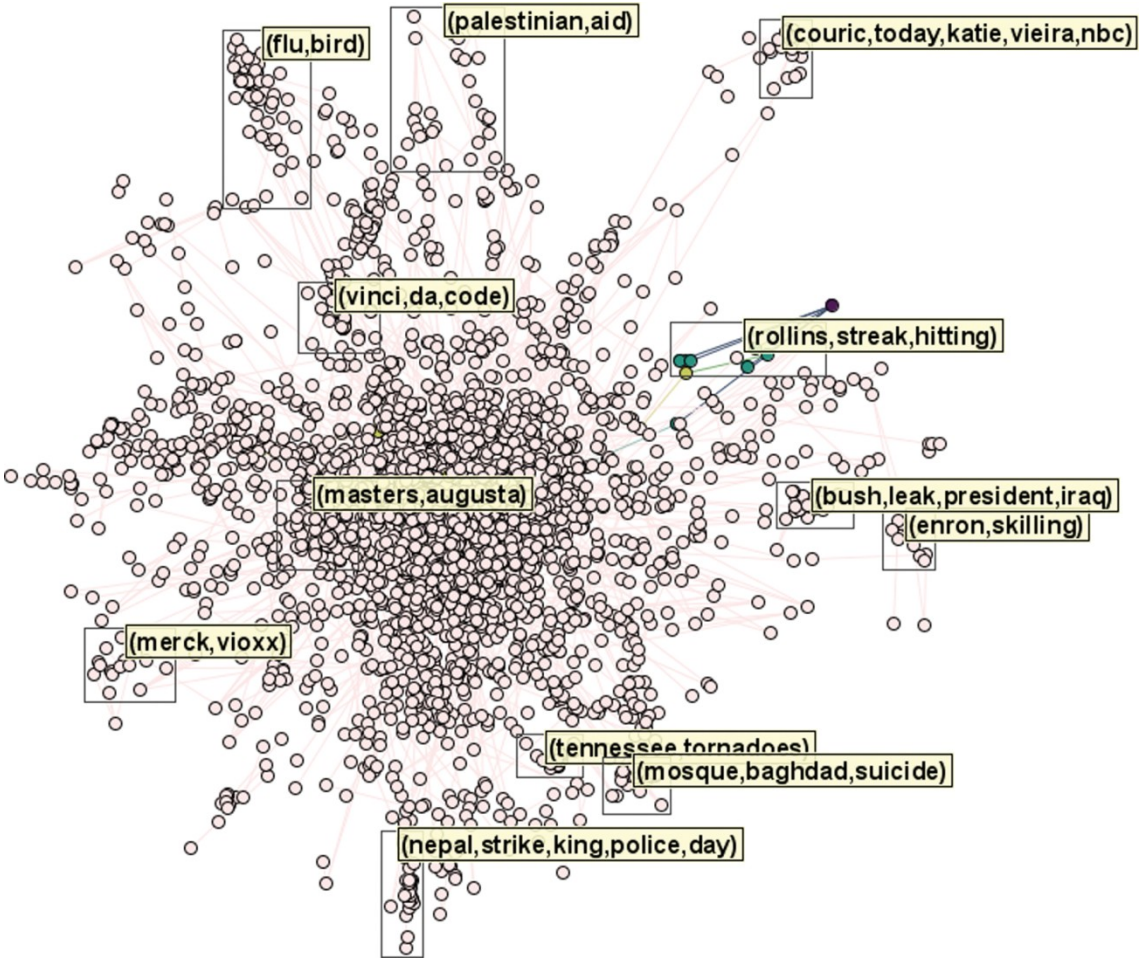
Results

- **Exploring RSS feeds of flash news (Associated Press, BBC, CNN, and Reuters)**
  - Corpus NEWS, 2,684 files



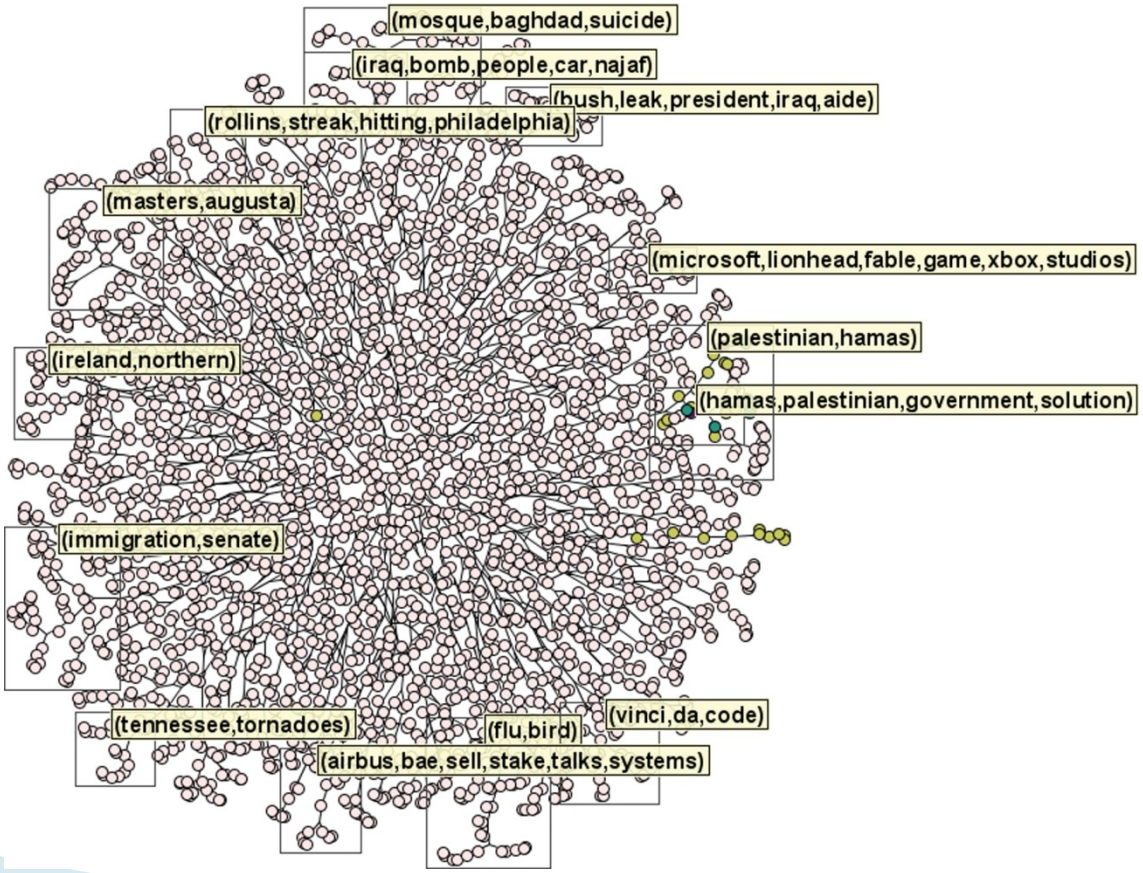
Results

- **Grouping by topic (LSP)**



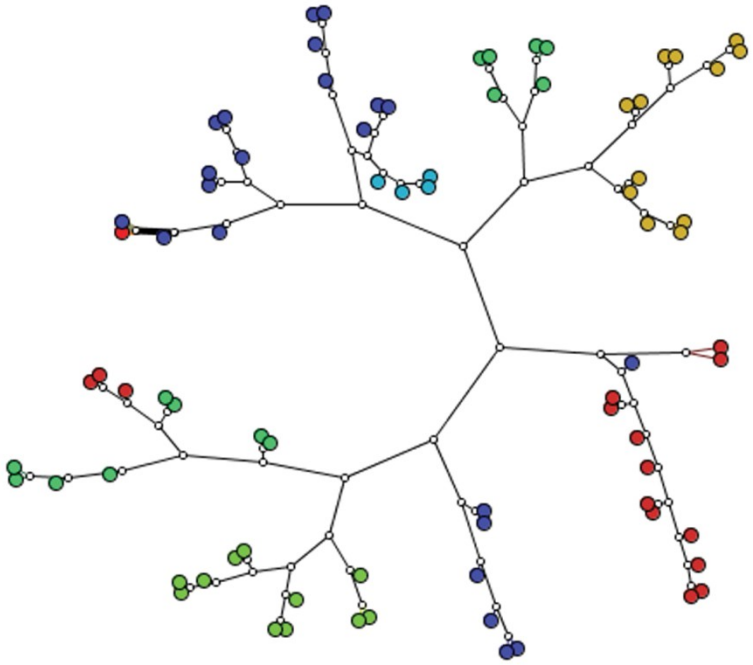
Results

- **Grouping by topic (NJ)**



Results

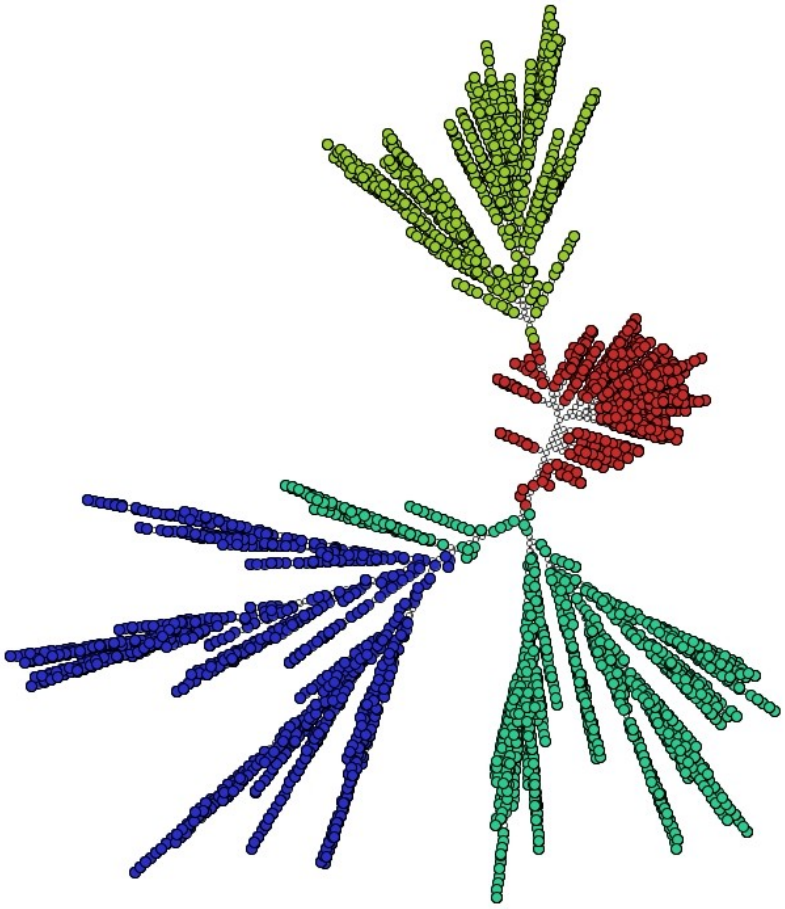
- **Stream-flow in hydroelectric plants of Paraná River (Brazil)**
  - Color is sub-basin of the river



Results

- **Quadrupeds mammals data set**

10,000 data instances



Results

- Time in seconds to create maps in a 3.2 GHz Pentium 4

Data Set	NJ	Layout	Total
CBR+ILP+IR+SON	4,55	0,52	5,17
KDVis	66,20	1,26	67,46
INFOVIS04	1,83	0,45	2,28
ALL	454,66	2,17	456,83
MESSAGES	0,35	0,31	0,66
NEWS	359,63	1,70	361,33

Concluding remarks on NJ

- NJ
  - Reflects content relationship visually
  - Constructs a hierarchy
- Interpretation of display
  - Makes good use of the visual space
  - Complementary of the projections
- Same distance matrix always generates the same tree
  - Helps evaluating the similarity measurement

Remarks on NJ

- Continuing work
  - Other trees, of course
  - Tools for proper exploration of similarity trees
  - Reduce processing time
  - Hybrid and hierarchical approaches
  - Improvement of some of NJ drawbacks

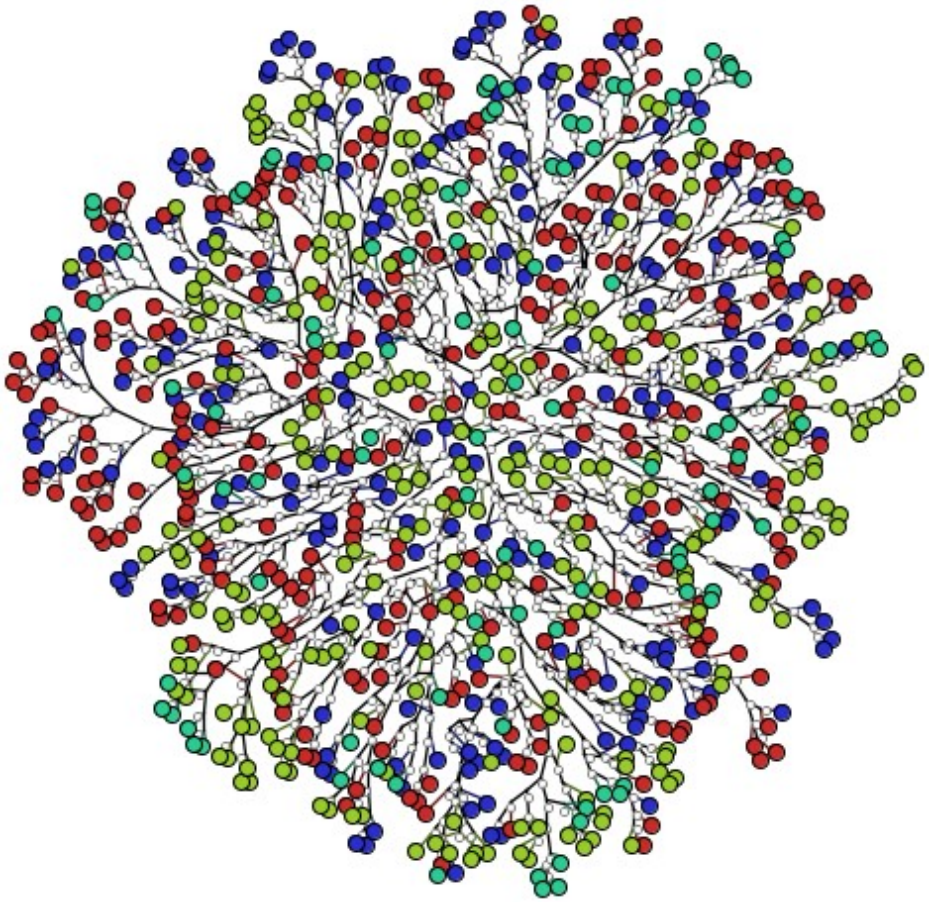
Minimun Spanning Tree

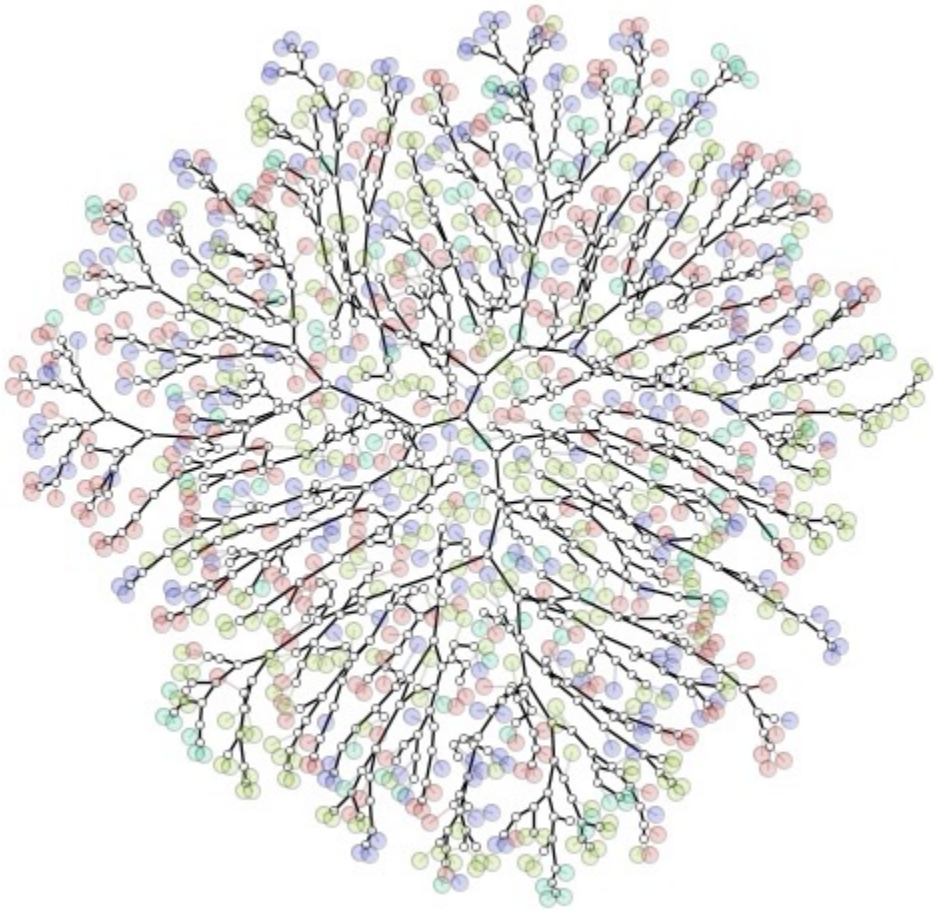
Minimun Spanning Tree

Problems?

NJ Trees

An open problem: Space Occupation

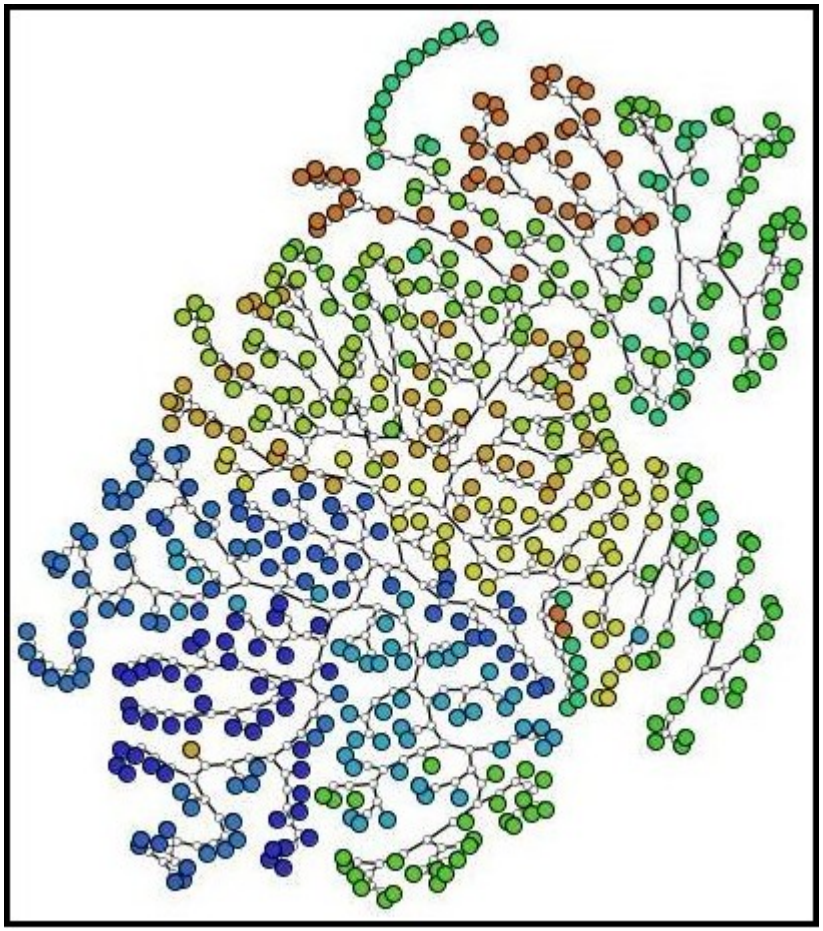




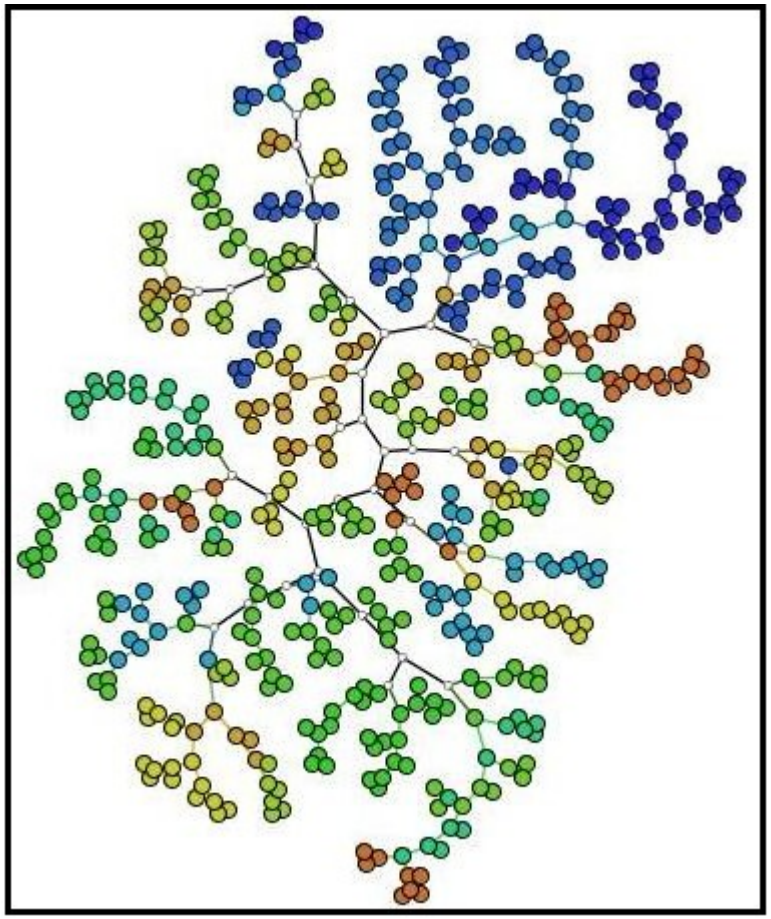
LCAD

NJ Trees

Slim NJ – 6 medical classes



NJ



Slim NJ

LCAD

Challenge 1:
Exploration of Data collections

- Exploratory tasks
  - Some queries are known, others are not
  - Questions arise
  - Model development is inspired by observations

- Test Cases
  - Image Collections
  - Volumetric (Scientific) Data Sets