Building Abstraction Hierarchies from Unbounded-arity Relations

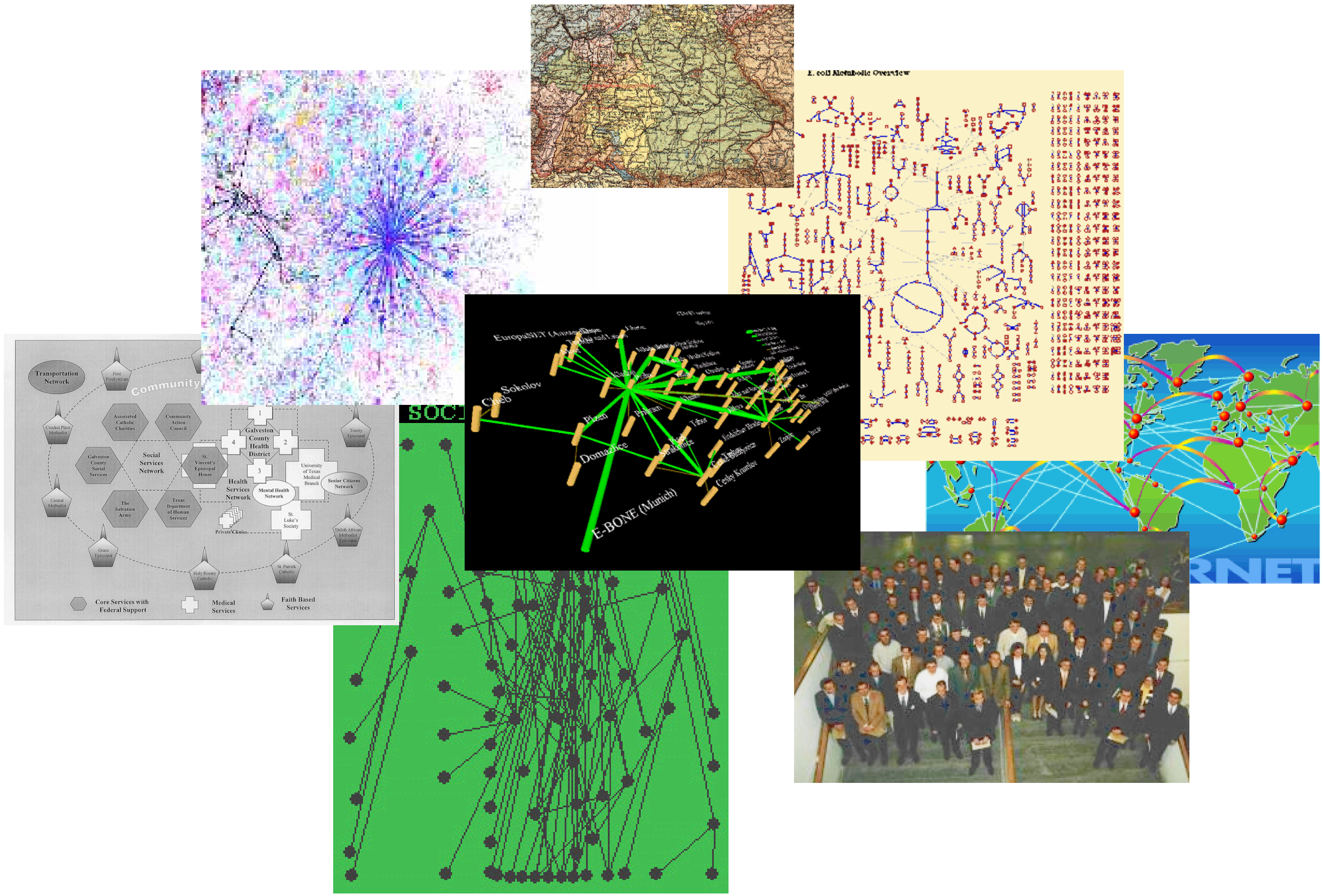
Klaus Holzapfel

Lehrstuhl für Effiziente Algorithmen

Fakultät für Informatik an der Technischen Universität München

Annual Meeting, March 26–28, 2003 (Tübingen)

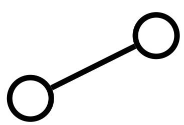
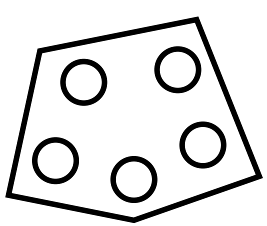
Real World Networks



Basic Concepts of Dealing with Real World Networks

Modeling the World

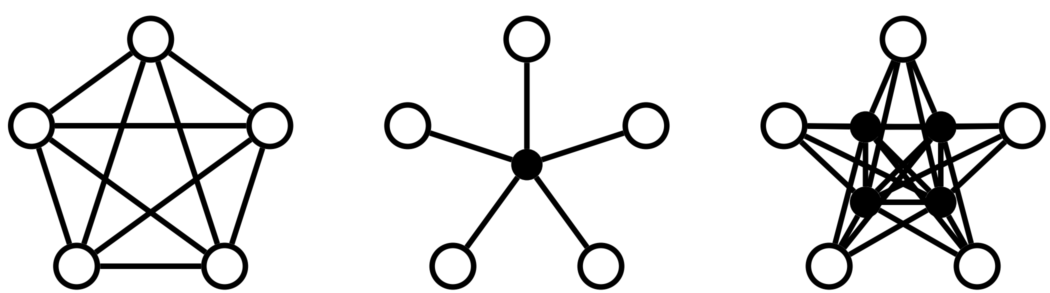
real world networks
=
entities and relations

two-arity 	symmetric: • who-knows-who asymmetric: • hyperlink graph • institutional hierarchy
unbounded-arity 	• Hollywood graph • social groups • co-authorship (LEABIB)

Modeling Relations

relations		hyper graphs		simple graphs
entities subsets	$\Rightarrow$	vertices hyperedges	$\Rightarrow$	vertices edges

- Transformation of entities to vertices (multiple occurrence, spelling)
- Introduction of weights for the hyperedges (size of subset, elements of subset)
- Transformation of hyperedge to edge



How to Deal with Large Data

Input properties:

- large-scale (10^5 authors, 10^{10} web pages)
- fuzzy data (noisy, incomplete)

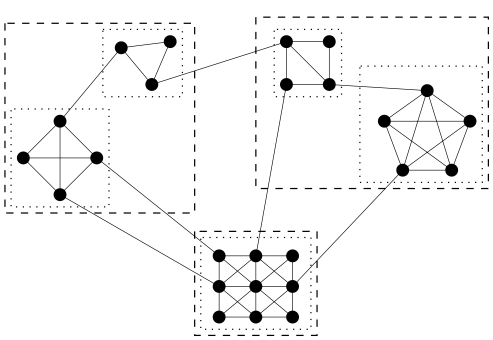
Opportunities:

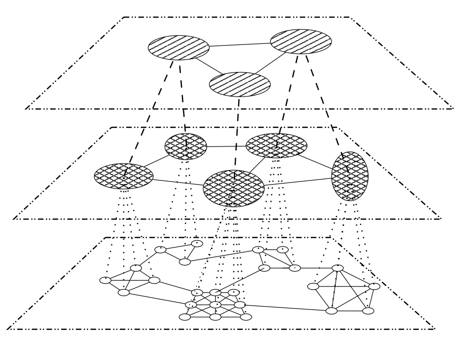
- reduction of the data

  $\rightarrow$ concentrating on a subset of entities

- abstraction of the data

  $\rightarrow$ representing groups of entities





Semantics of Hierarchies on Relations

Aim:

- groups of entities should represent strong relations between their elements
- relations between those groups should represent the relations between its group members

Applications:

- general multi-arity relations

  $\rightarrow$ social centers / structure

  $\rightarrow$ project management
- co-citation relation (LEABIB)

  $\rightarrow$ group detection (research, referees)

  $\rightarrow$ quality

  $\rightarrow$ query enhancement

How to detect groups?

- connectivity

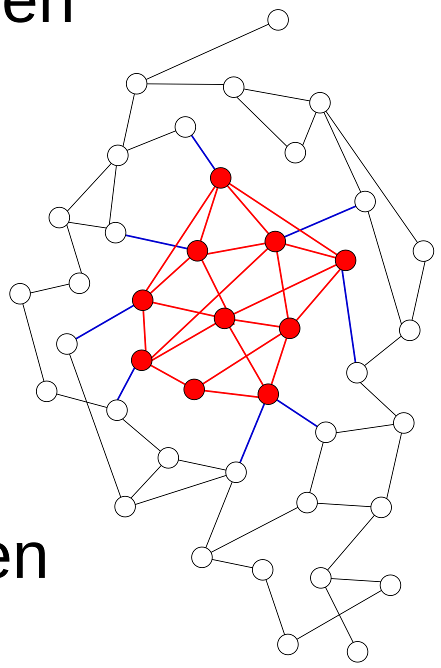
  $\rightarrow$ optimize number of disjoint paths between any two vertices in the group

- weighted Min-Cut (Ratio-Cut)

  $\rightarrow$ optimize the quantity: $\frac{\mathrm{Cut\_Size}(A,B)}{\|A\| \cdot \|B\|}$

- density based

  $\rightarrow$ optimize the number of relations between the entities in the group (e.g. $\|E\|$)



Density Based-Clustering(1)

Motivation:
- small-world characteristics (sparse graphs with dense substructures)
- looking for regions with significantly higher average degree

Problem:	γ-DENSE SUBGRAPH-PROBLEM (γ-DSP)
Input:	Graph G, $k \in \mathbb{N}$
Output:	Does there exist a subgraph G' of size k having at least $\gamma(k)$ edges

- $\gamma(k) = \binom{k}{2}$ γ-DSP $=$ CLIQUE $\in$ **NP-C**
- $\gamma(k) = 0$ γ-DSP $\in$ **P**

Complexity-Results

[H. et al. CIAC 2003]

Theorem. *Let* $\gamma : \mathbb{N} \to \mathbb{N}$ *be a function that is computable in polynomial time:*

1. *If* $\gamma(k) = k + O(1)$ *then* γ-DSP *is in* **P**.
2. *If* $\gamma(k) = k + \Omega(k^\epsilon)$, *for some rational number* $0 < \epsilon < 2$, *then* γ-DSP *is* **NP**-*complete.*

Theorem. *Let* $\gamma : \mathbb{N} \to \mathbb{N}$ *be a function that is computable in polynomial time.*

If $\gamma(k) \in k + \Theta(\log k)$ *and* γ-DSP *in* **NP**-*c, then*

NP$\subseteq$ **DTIME**$(n^{O(\log(n))})$.

An Abstraction Hierarchy for Unbounded-arity Relations

Our Approach

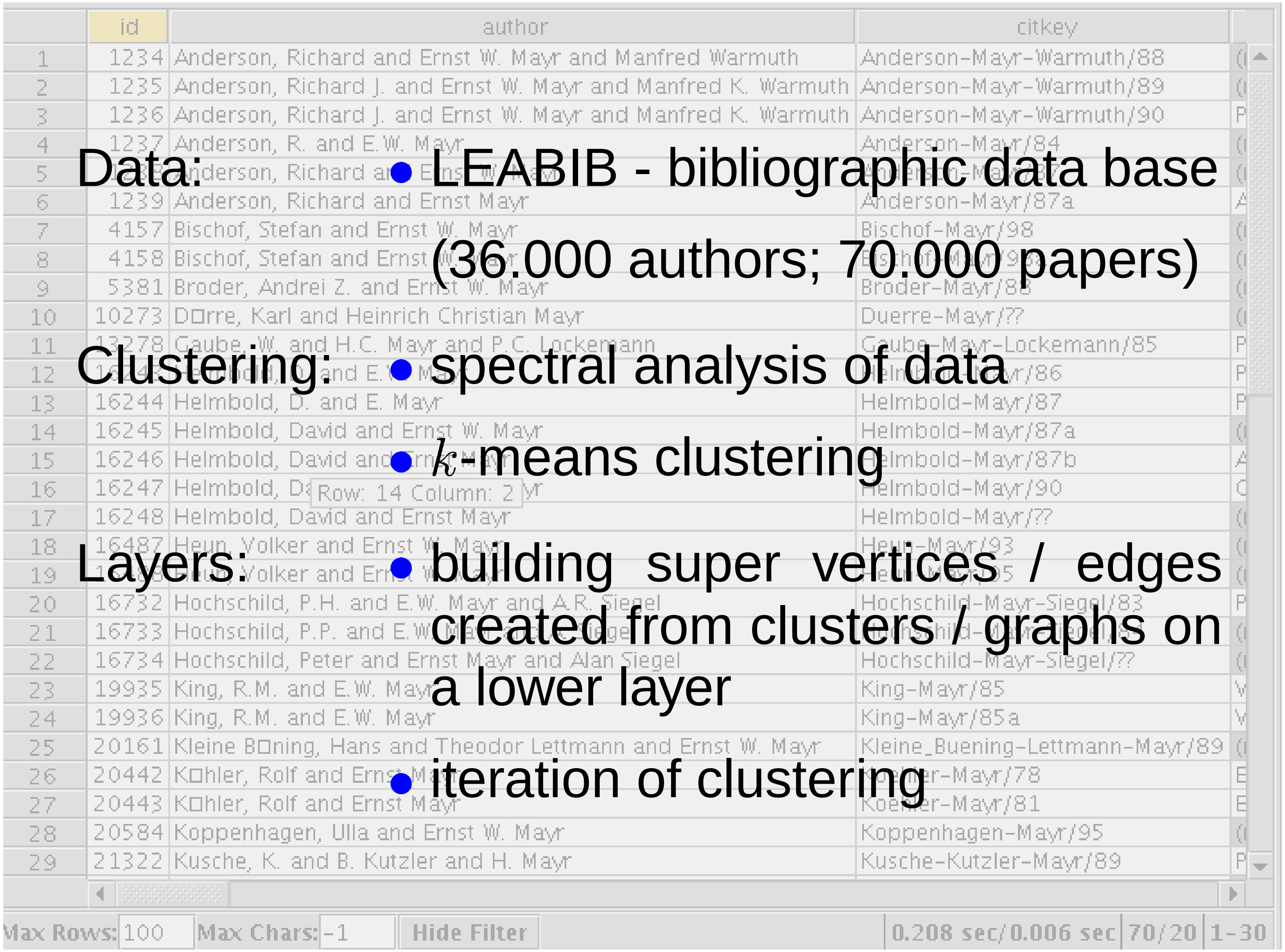
Data: • LEABIB - bibliographic data base (36.000 authors; 70.000 papers)

Clustering: • spectral analysis of data

• k-means clustering

Layers: • building super vertices / edges created from clusters / graphs on a lower layer

• iteration of clustering



	id	author	citkey
1	1234	Anderson, Richard and Ernst W. Mayr and Manfred Warmuth	Anderson-Mayr-Warmuth/88
2	1235	Anderson, Richard J. and Ernst W. Mayr and Manfred K. Warmuth	Anderson-Mayr-Warmuth/89
3	1236	Anderson, Richard J. and Ernst W. Mayr and Manfred K. Warmuth	Anderson-Mayr-Warmuth/90
4	1237	Anderson, R. and E.W. Mayr	Anderson-Mayr/84
5	[illegible]	[illegible]	[illegible]
6	1239	Anderson, Richard and Ernst Mayr	Anderson-Mayr/87a
7	4157	Bischof, Stefan and Ernst W. Mayr	Bischof-Mayr/98
8	4158	[illegible]	[illegible]
9	5381	Broder, Andrei Z. and Ernst W. Mayr	Broder-Mayr/88
10	10273	D□rre, Karl and Heinrich Christian Mayr	Duerre-Mayr/??
11	13278	Gaube, W. and H.C. Mayr and P.C. Lockemann	Gaube-Mayr-Lockemann/85
12	[illegible]	[illegible]	[illegible]
13	16244	Helmbold, D. and E. Mayr	Helmbold-Mayr/87
14	16245	Helmbold, David and Ernst W. Mayr	Helmbold-Mayr/87a
15	16246	[illegible]	Helmbold-Mayr/87b
16	16247	[illegible]	Helmbold-Mayr/90
17	16248	Helmbold, David and Ernst Mayr	Helmbold-Mayr/??
18	16487	Heun, Volker and Ernst W. Mayr	Heun-Mayr/93
19	[illegible]	[illegible]	[illegible]
20	16732	Hochschild, P.H. and E.W. Mayr and A.R. Siegel	Hochschild-Mayr-Siegel/83
21	16733	[illegible]	[illegible]
22	16734	Hochschild, Peter and Ernst Mayr and Alan Siegel	Hochschild-Mayr-Siegel/??
23	19935	King, R.M. and E.W. Mayr	King-Mayr/85
24	19936	King, R.M. and E.W. Mayr	King-Mayr/85a
25	20161	Kleine B□ning, Hans and Theodor Lettmann and Ernst W. Mayr	Kleine_Buening-Lettmann-Mayr/89
26	20442	[illegible]	[illegible]
27	20443	K□hler, Rolf and Ernst Mayr	Koehler-Mayr/81
28	20584	Koppenhagen, Ulla and Ernst W. Mayr	Koppenhagen-Mayr/95
29	21322	Kusche, K. and B. Kutzler and H. Mayr	Kusche-Kutzler-Mayr/89

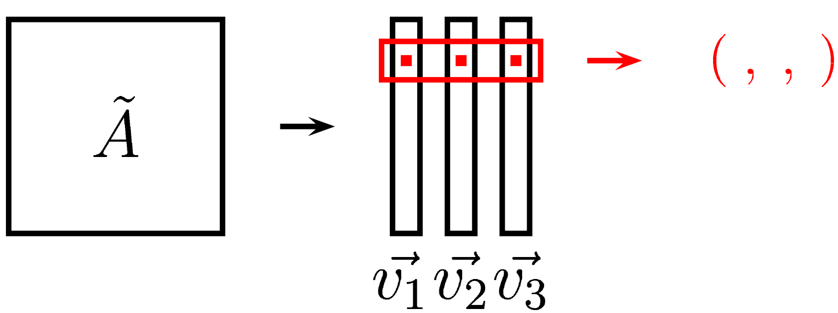
Clustering with Spectral Partitioning

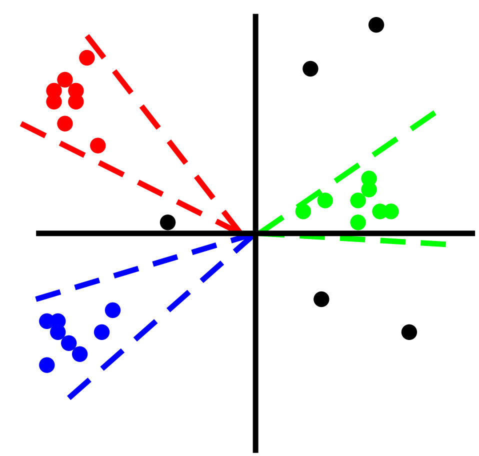
Idea:

- using SVD for low-rank approximation of (weighted) adjacency matrix
- transformation of n-dimensional data into a low-dimensional space

Advantage:

- fast approximation algorithms
- spectral partitioning and ratio-cut partitioning are correlated

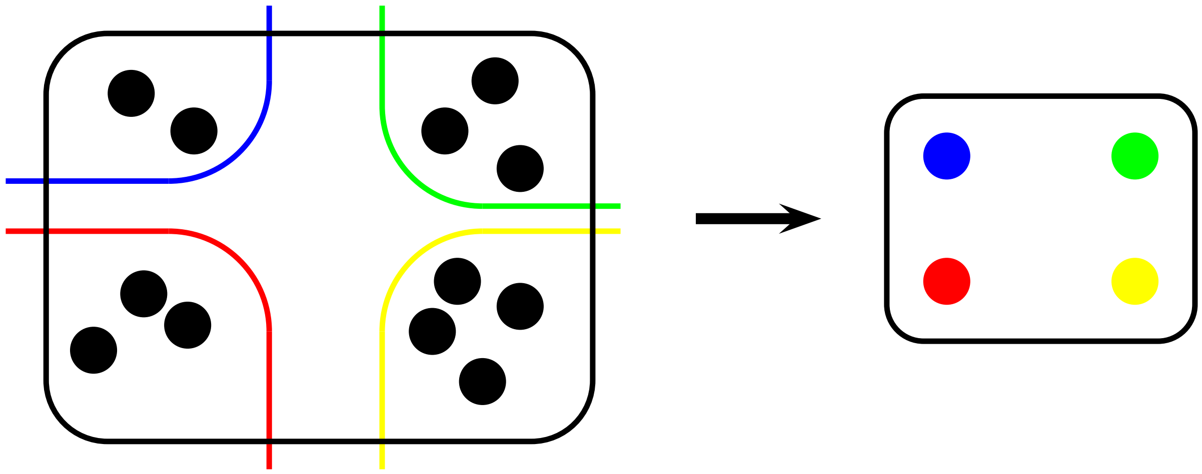




Building Levels

Input:	Hyper graph G, clustering C
Output:	Hyper graph G' at the next higher level

- vertices: clusters
- hyperedges: shrink old hyperedges



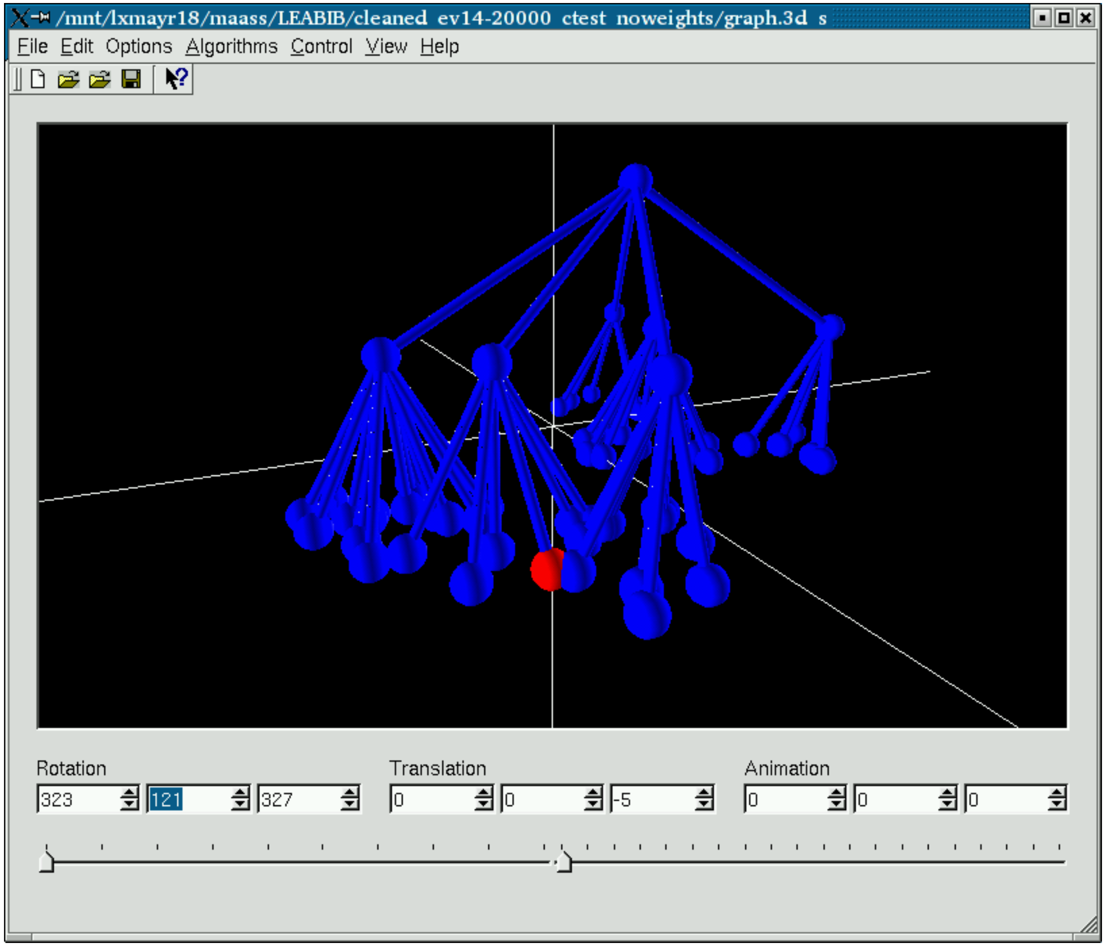
Analysis of Clusters

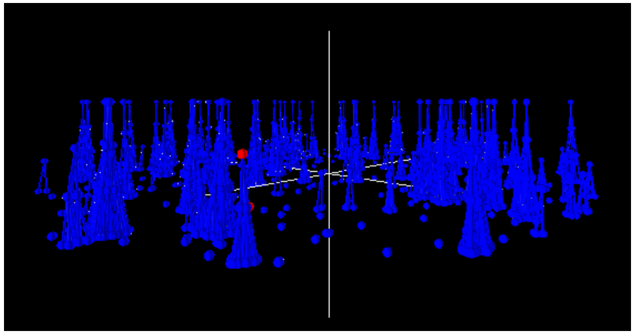
```
Cluster 191 on level 2:

Cluster   232 on level 1
Cluster   240 on level 1
Cluster   560 on level 1
Cluster   561 on level 1
Cluster   648 on level 1
Cluster   780 on level 1
```

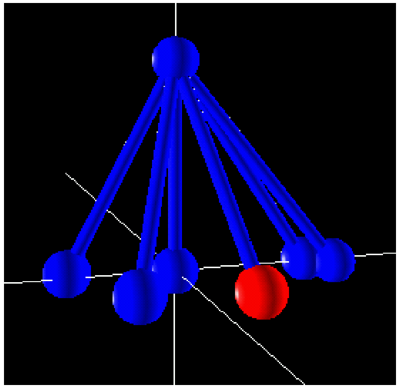
```
Cluster 232 on level 1:

 871  Mayr, E. W.
2564  Bischof, S.
6559  Heun, V.
7331  Lettmann, T.
7429  Koppenhagen, U.
```

```
Cluster 780 on level 1:

1110  Suen, S.
1216  Assaf, S.
1217  Upfal, E.
1304  Feige, U.
1348  Shamir, E.
1393  Broder, A. Z.
1394  Karlin, A. R.
3101  Frieze, A. M.
      ...
```

Visualization







Future Directions, What's Next(1)

Input Data

- difficulties: $\rightarrow$ noisy data
  $\rightarrow$ small data set
- tasks: $\rightarrow$ data cleaning
  $\rightarrow$ incorporation of further data

Modeling of hyper graphs

- representation of hyperedges
- usage of weights
- graph abstraction

Future Directions, What's Next(2)

$\boxed{\text{Dynamic hierarchies}}$

- modification of hierarchy
  - $\rightarrow$ union/split of clusters
  - $\rightarrow$ insertion/deletion of levels
- incorporation of further data
  - $\rightarrow$ insertion/deletion of vertices/edges
  - $\rightarrow$ modification of weight functions
- parametric clusters/hierarchies
  - $\rightarrow$ transformation/weights of hyperedges
  - $\rightarrow$ graph generation on each level