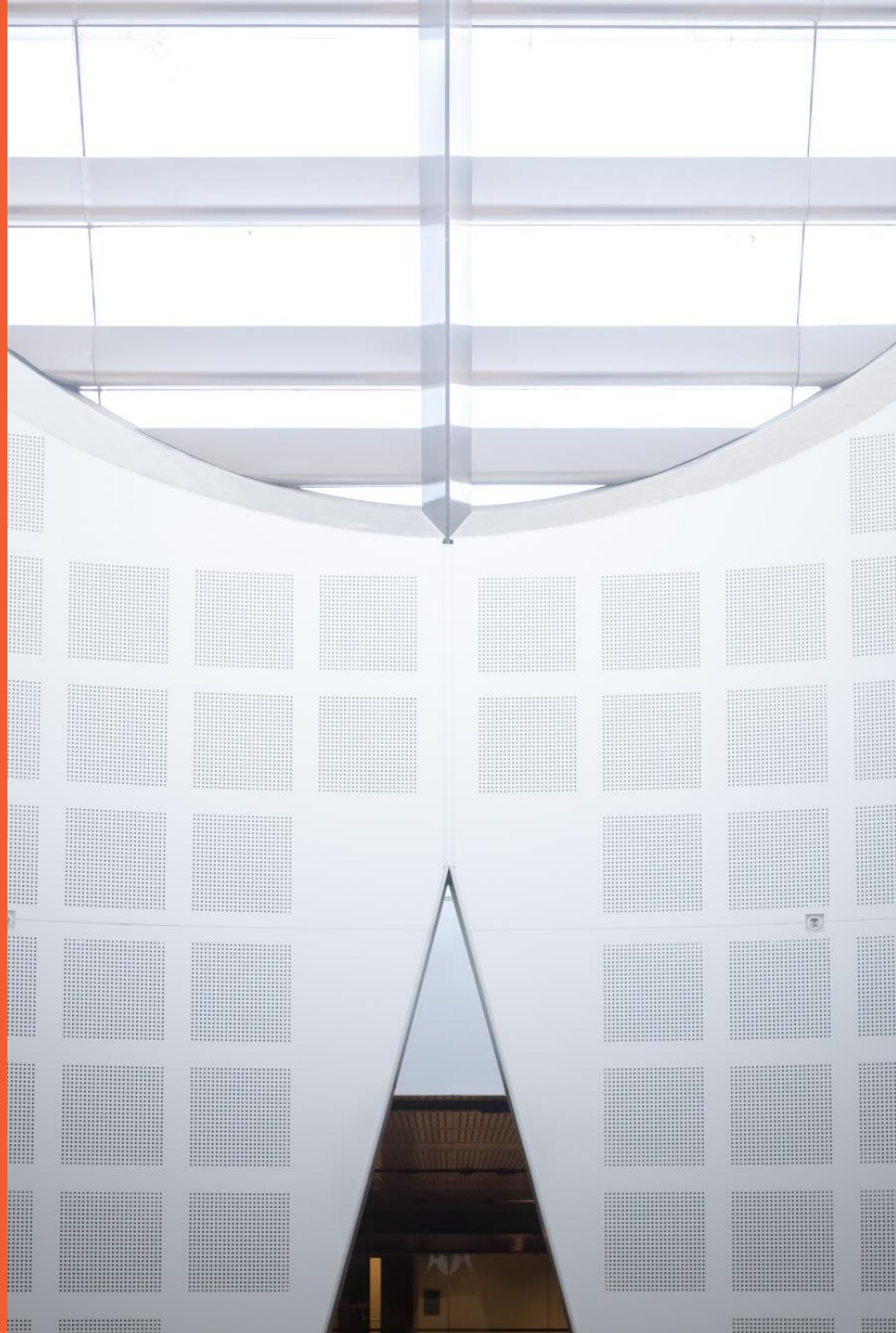
COMP5048
Visual Analytics

Week 7: Visual Analytic System (Examples)

Professor Seokhee Hong

School of Information Technologies

THE UNIVERSITY OF SYDNEY



Copyright warning

Visual Analytic System

For example, we plan to ***design/implement/evaluate*** a VA system for ***XXX data***, to support ***YYY tasks***.

Data: (eg) IMDB, Youtube, Twitter, Facebook etc.

Task: (eg) Find temporal patterns/outlier/trends/groups etc.

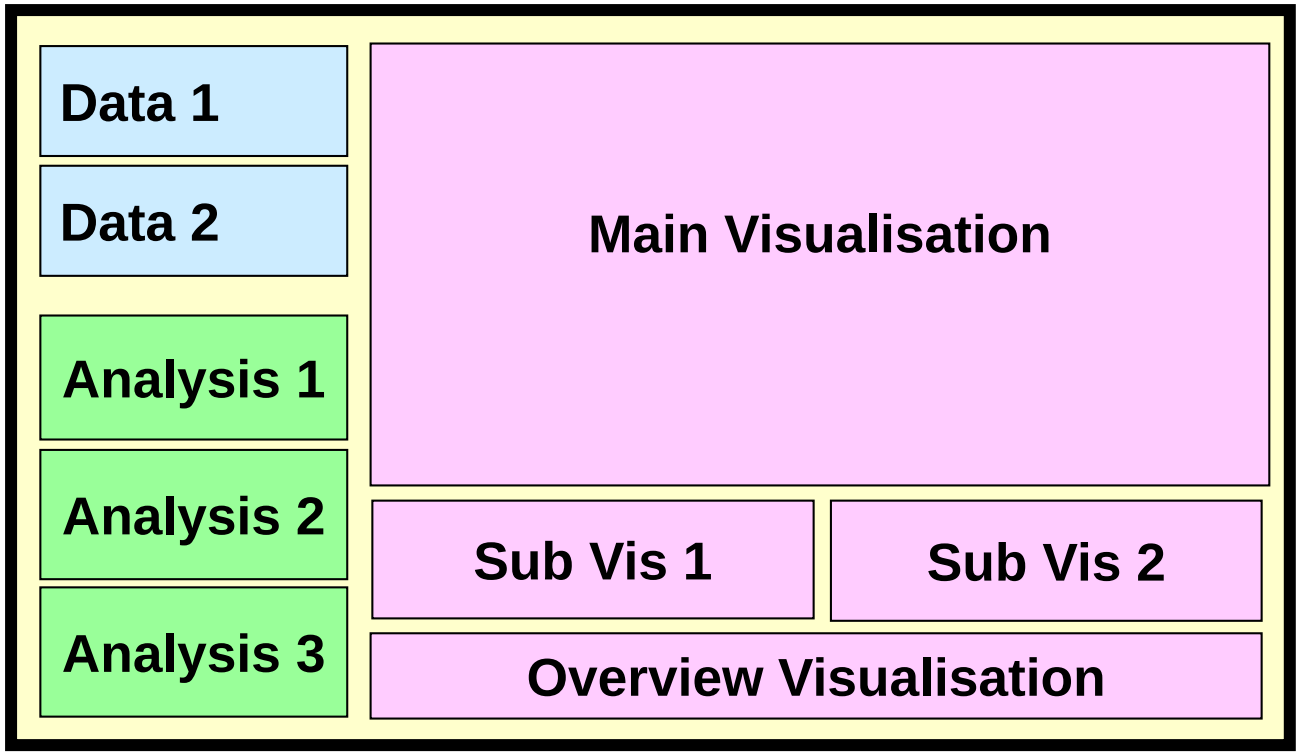
Data processing:
- filter/derive: (eg) collaboration network, co-citation network etc

Design
- ***Analysis***: (eg) clustering, centrality analysis etc
- ***Visualisation***: (eg) parallel coordinates, treemap, spring algorithm etc

Implement
- *Hardware spec,* ***Softwares/Tools****, programming languages,* ***GUI***

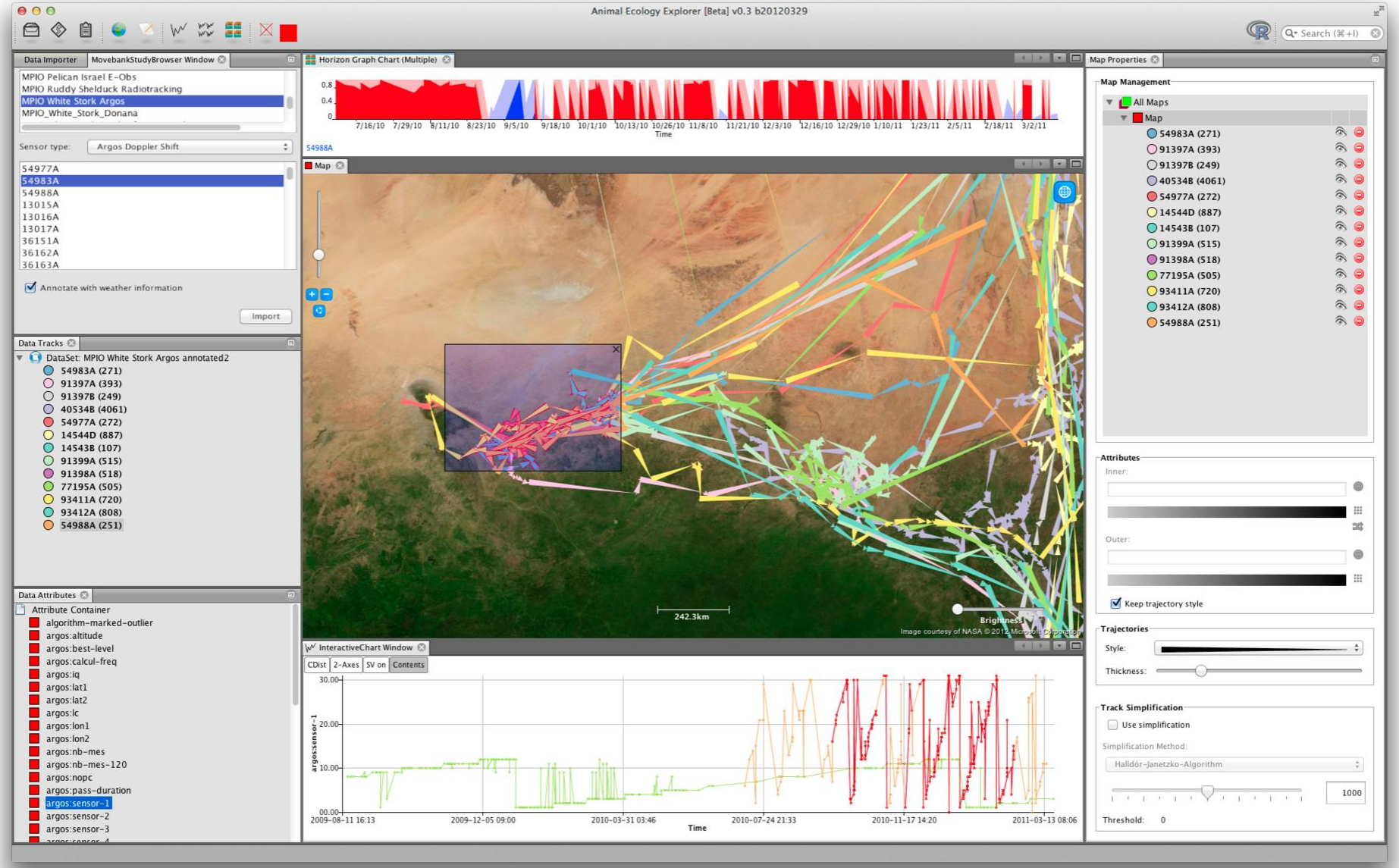
Evaluation
- (eg) Case study, Use case scenario, controlled experiment, feedback

Data: (eg) IMDB, WOS, Twitter, Facebook etc.
Task: (eg) Find temporal patterns/outlier/trends/groups etc.
Data processing: (eg) collaboration network, co-citation network etc
Design
- ***Analysis***: (eg) clustering, centrality analysis etc
- ***Visualisation***: (eg) parallel coordinates, treemap, spring algorithm etc



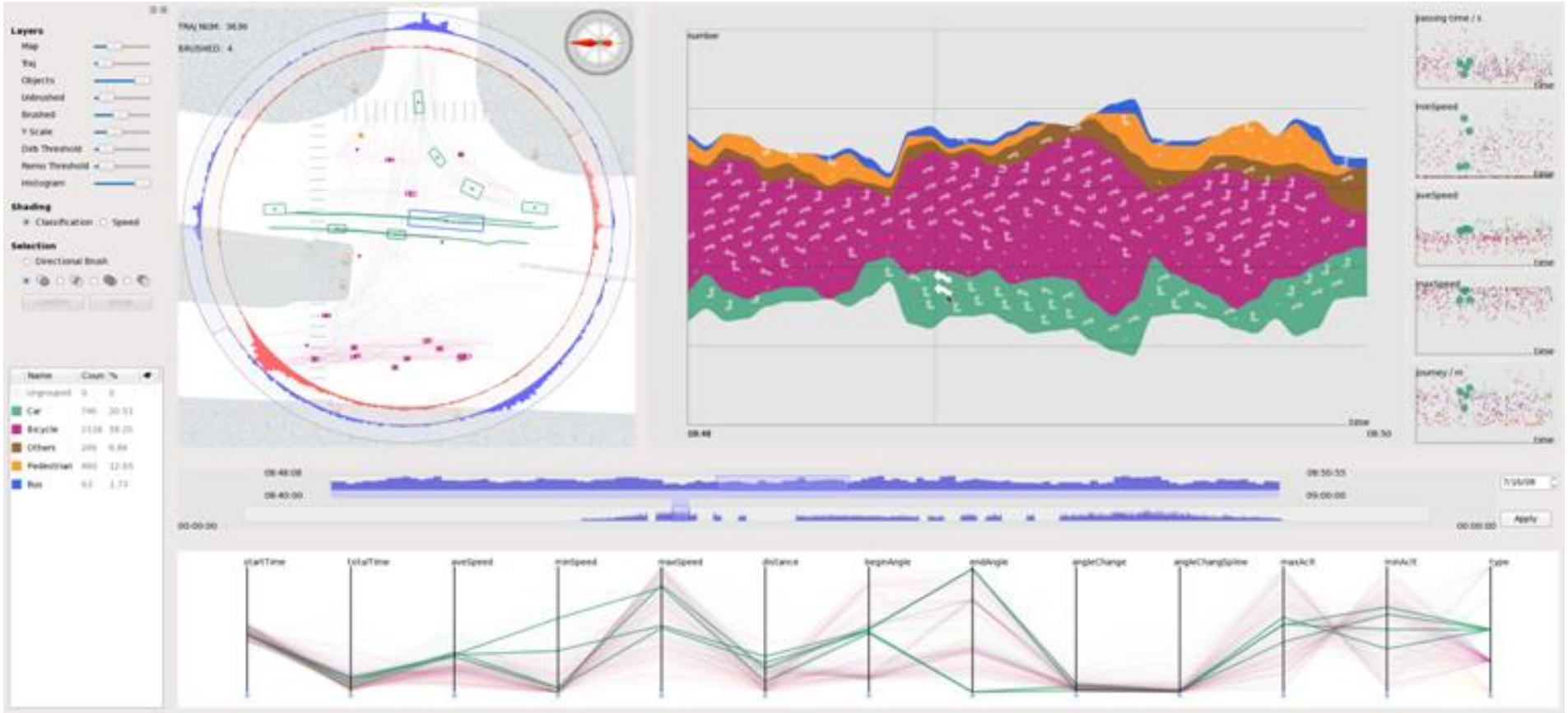
Trajectory VA system [Keim et al.]

Task: spatial-temporal analysis of movement data



Trajectory VA System [Yuan et al.]

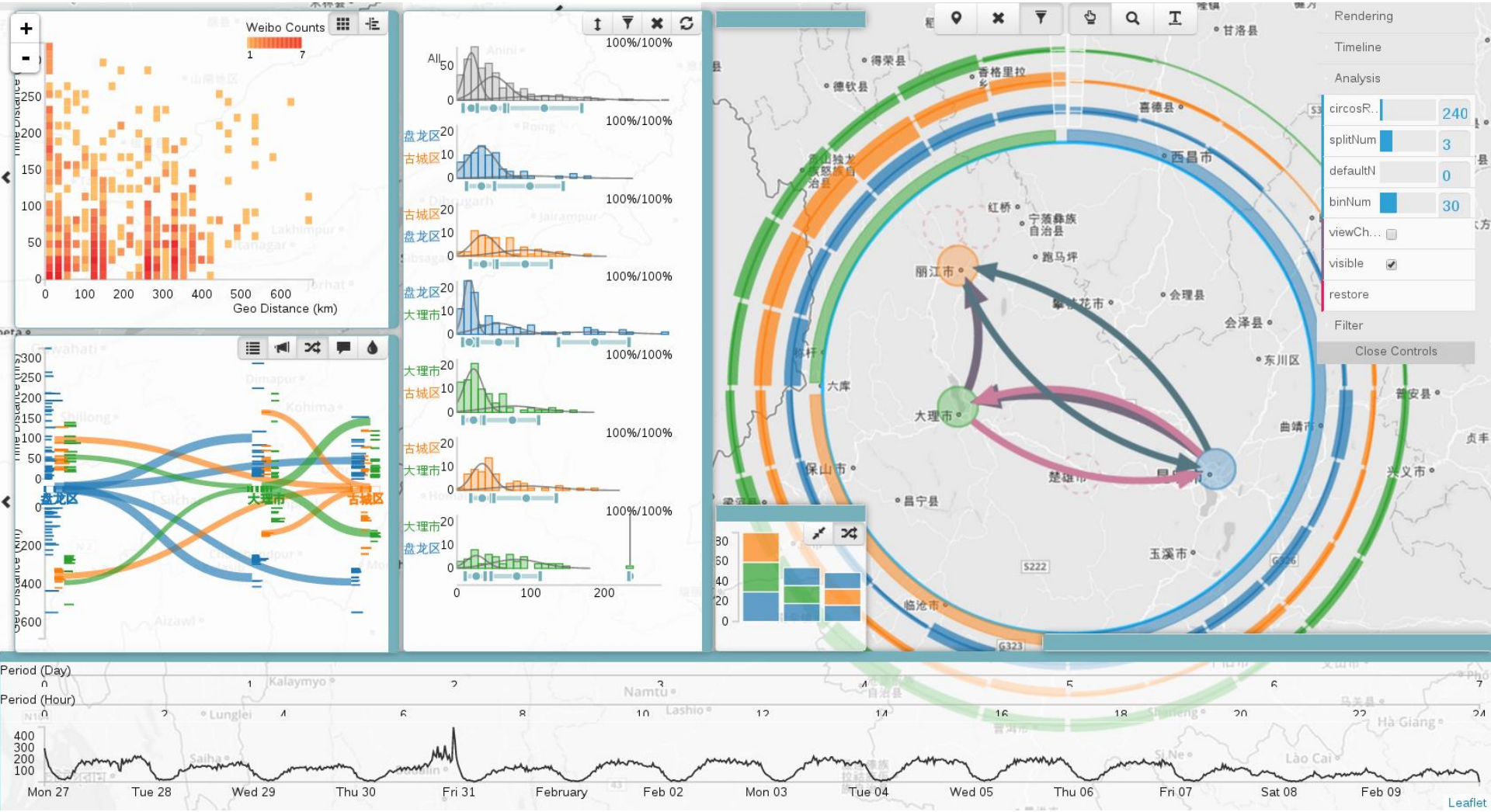
Task: Trajectory behavior analysis



Weibo VA system [Yuan et al.]

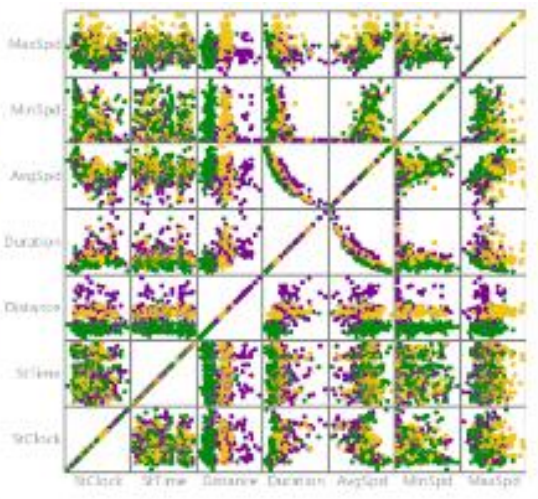
Data: Social media data with geotags

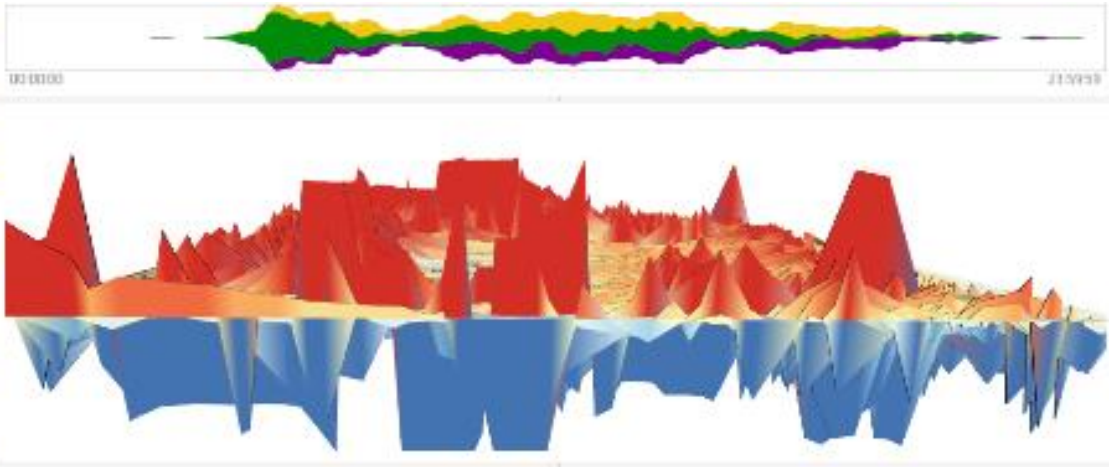
Task: movement pattern analysis

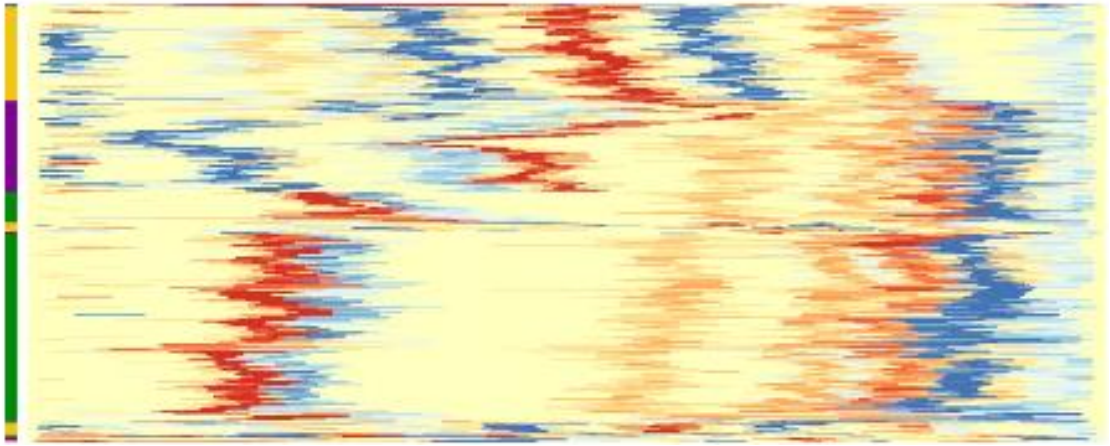


Trajectory VA System [Yuan et al.]

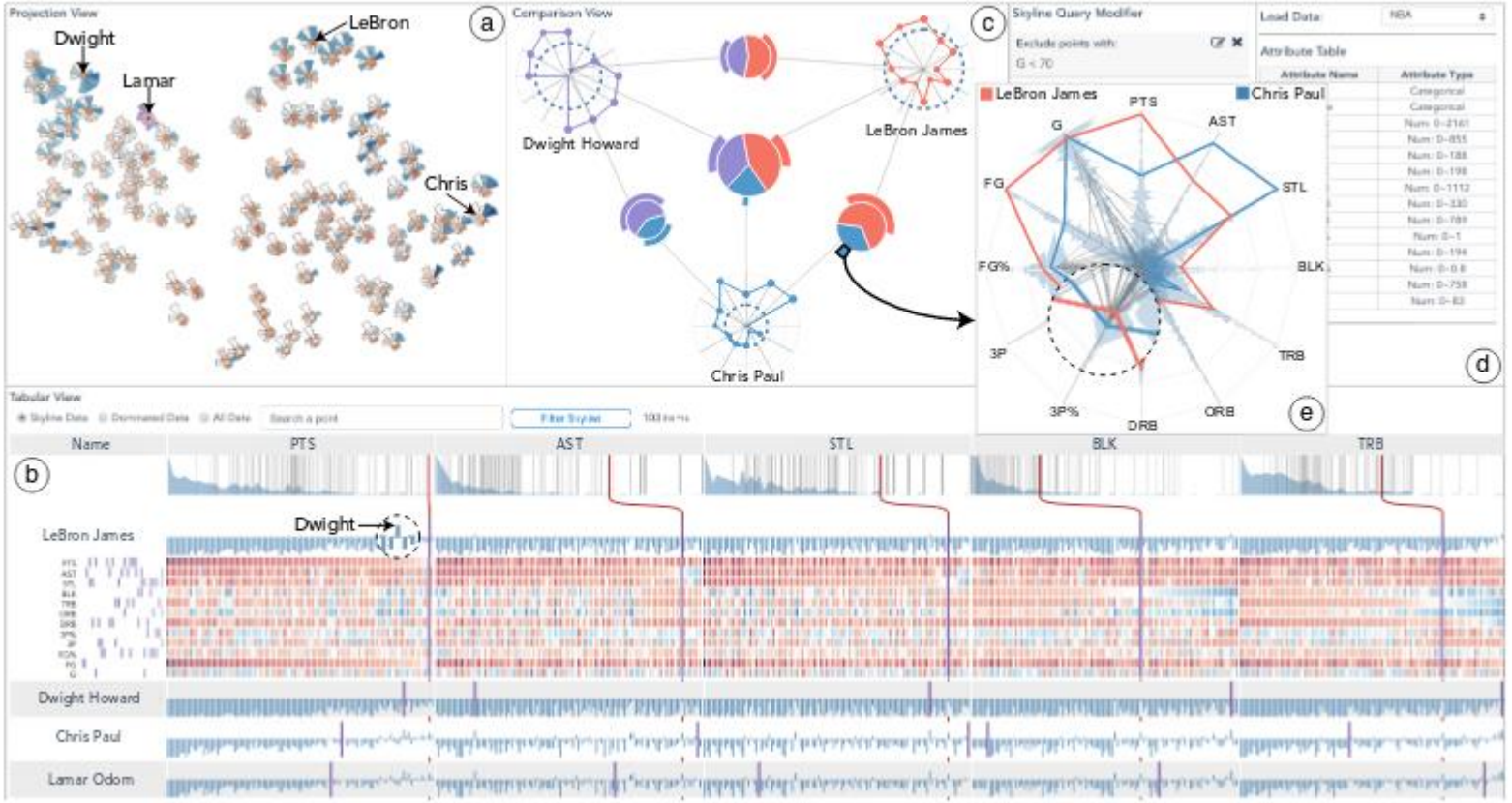
Task: Comparison of Trajectories using Timeline







SkyLens:
Visual Analysis of Skyline on Multi-dimensional Data



Source: http://www.cse.ust.hk/~huamin/tvcg_xun_skyline_2017.pdf

How to **Evaluate** VA Systems

Define Tasks

- show the ***structure***
- discover *new* ***knowledge/insight***
- find *regular/abnormal* ***patterns***
- generate/confirm/reject ***hypothesis***
- confirm ***expected*** and discover ***unexpected***
- predict the ***future***

Evaluation methods (Details: Week 9)

- case study
- expert feedback
- controlled human experiments
- questionnaire
- cognitive walkthrough

…

Approaches to **Design** VA System

1. ***Overview first, then Details on demand***

2. ***Overlay analysis using visual variables (data-ink ratio)***

3. ***Integrate a number of analysis and visualisation methods***

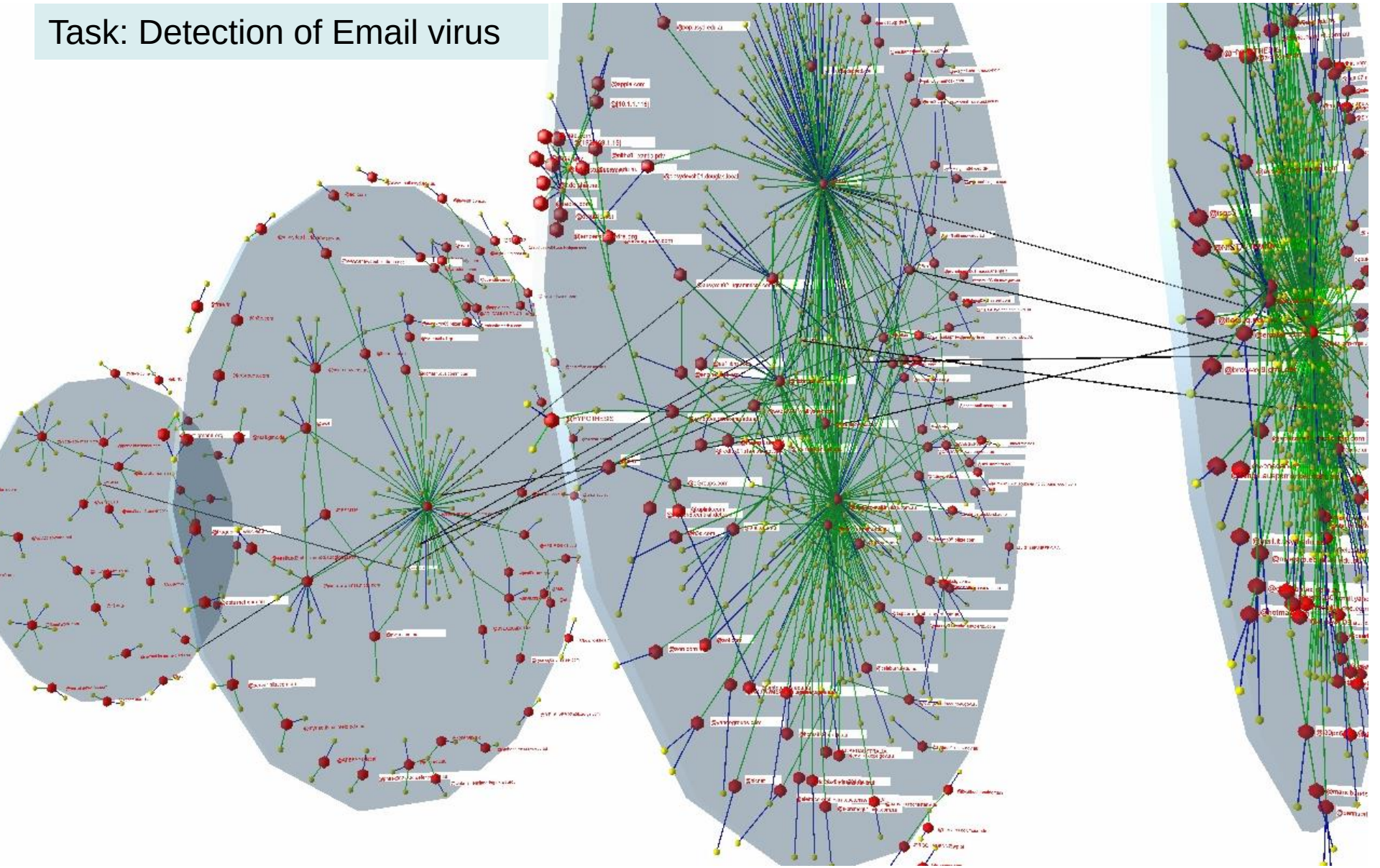
4. ***If the data is big/complex, reduce the data set***

5. ***Storytelling with the data: narrative visualisation***

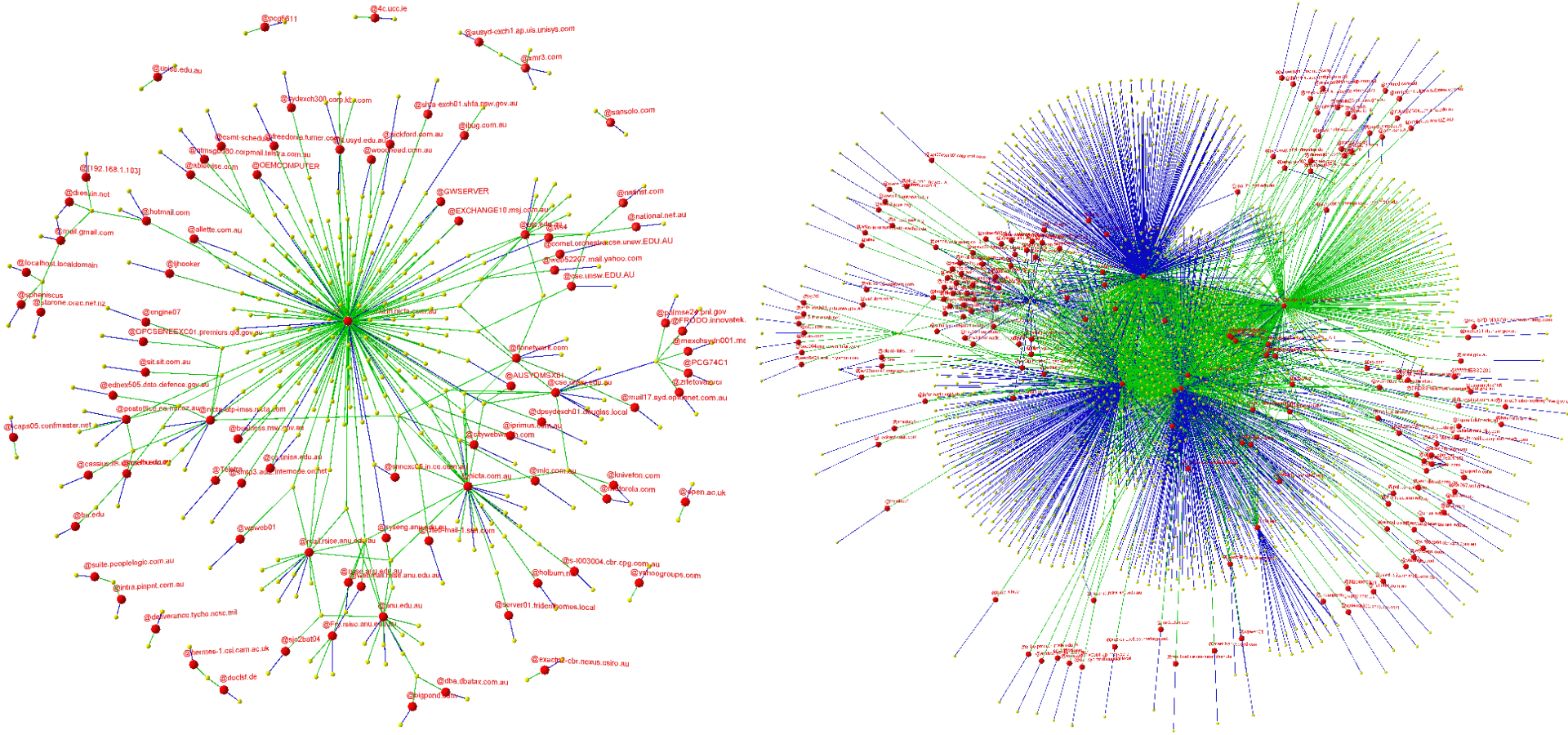
1. *Overview* first, *zoom* in, then *Details* on demand

Overview: Email Virus Visualisation

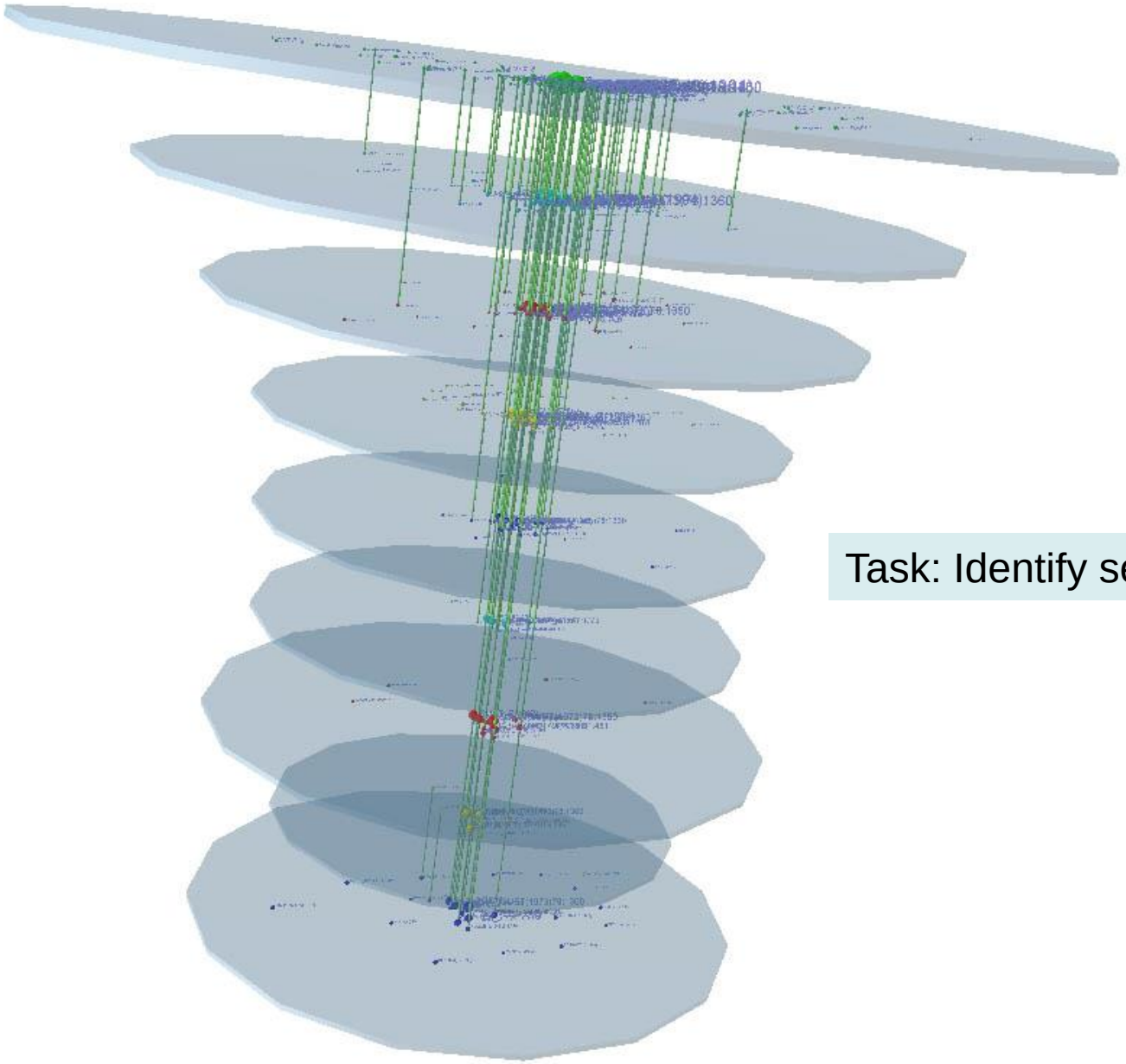
Task: Detection of Email virus



Details: Email Virus Detection



Overview: Evolution of citation network (WOS)

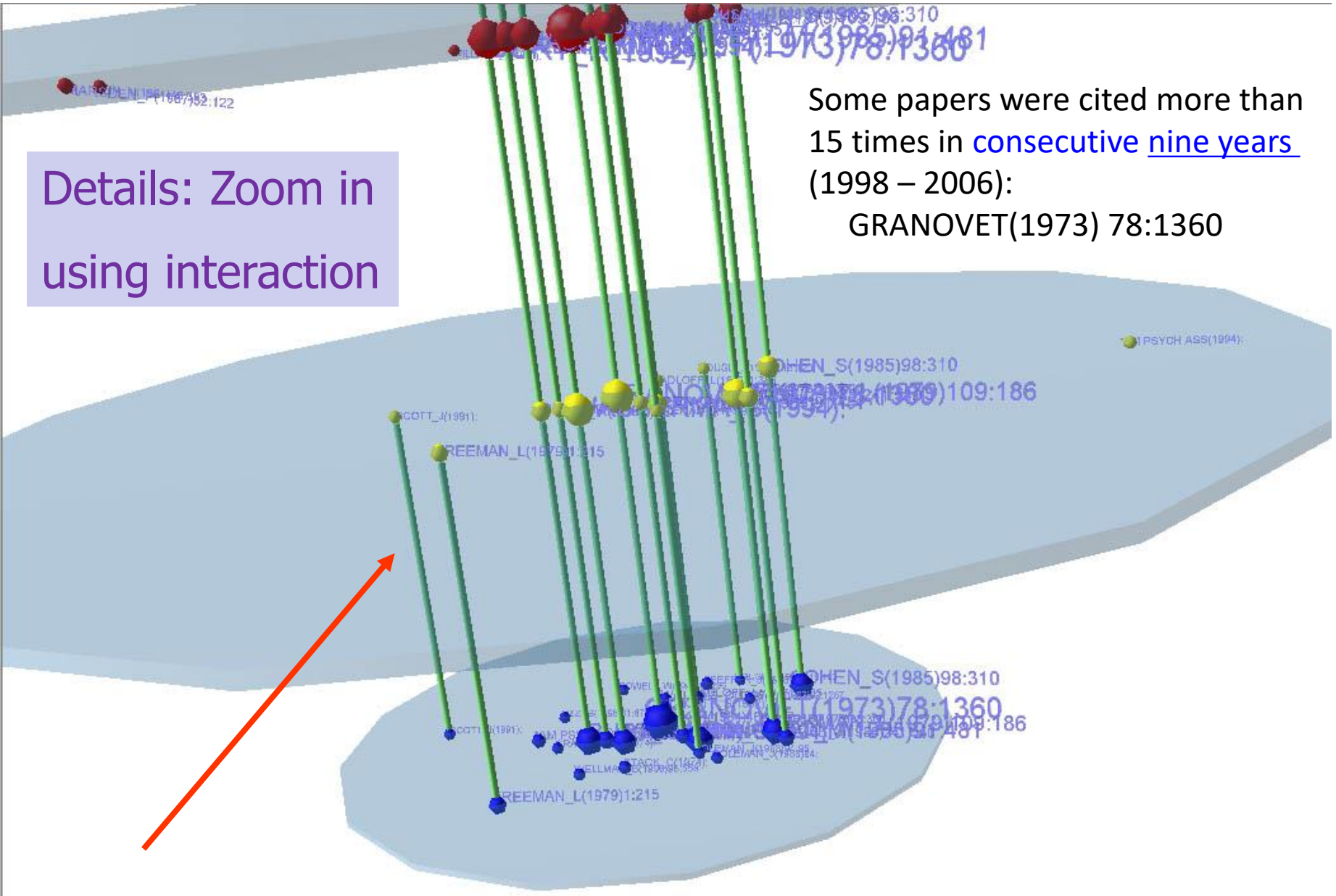


Task: Identify seminal papers

Details: Zoom in using interaction

Some papers were cited more than 15 times in consecutive nine years (1998 – 2006):

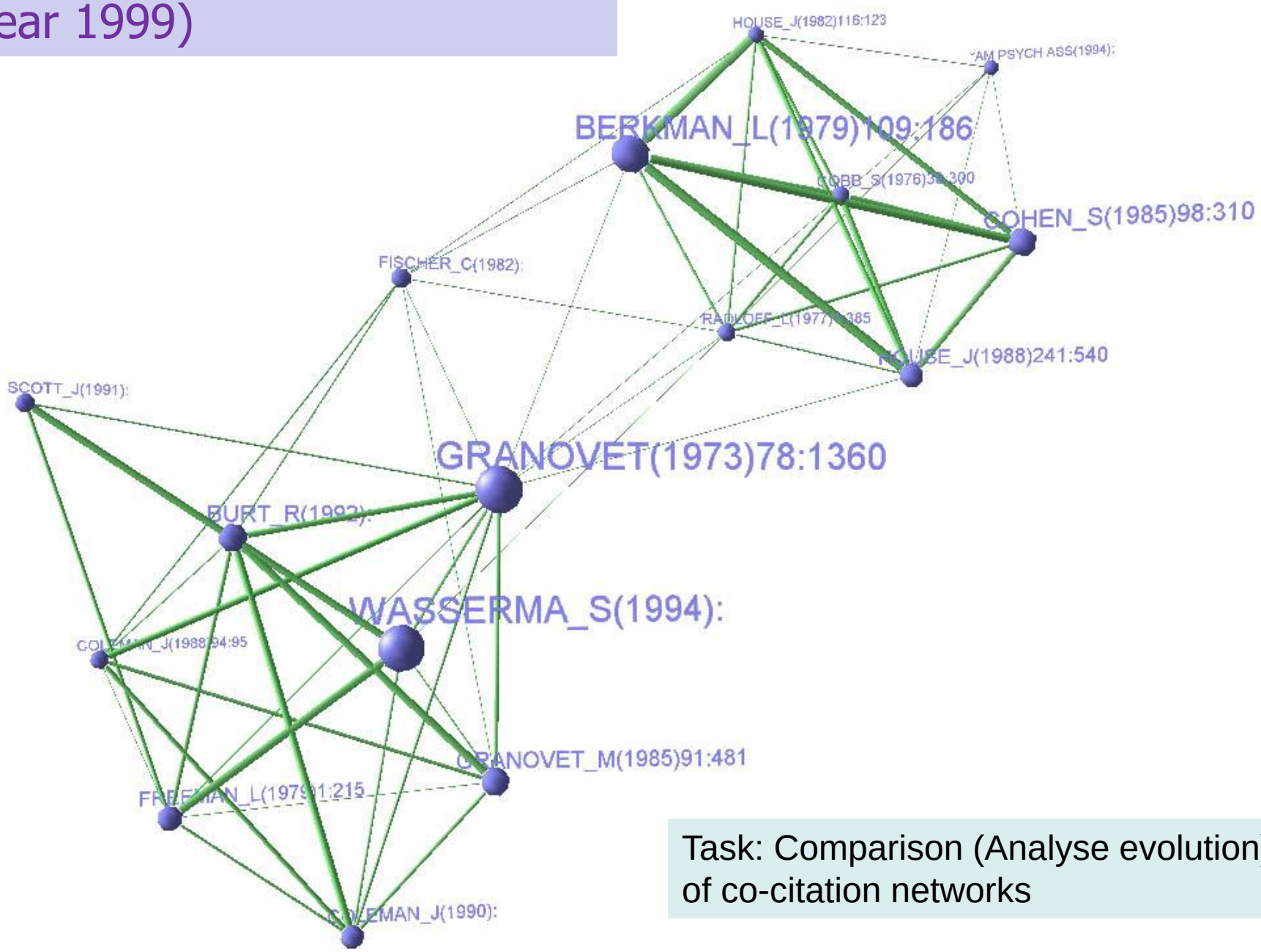
GRANOVET(1973) 78:1360



Some papers were cited more than 15 times in consecutive two years:

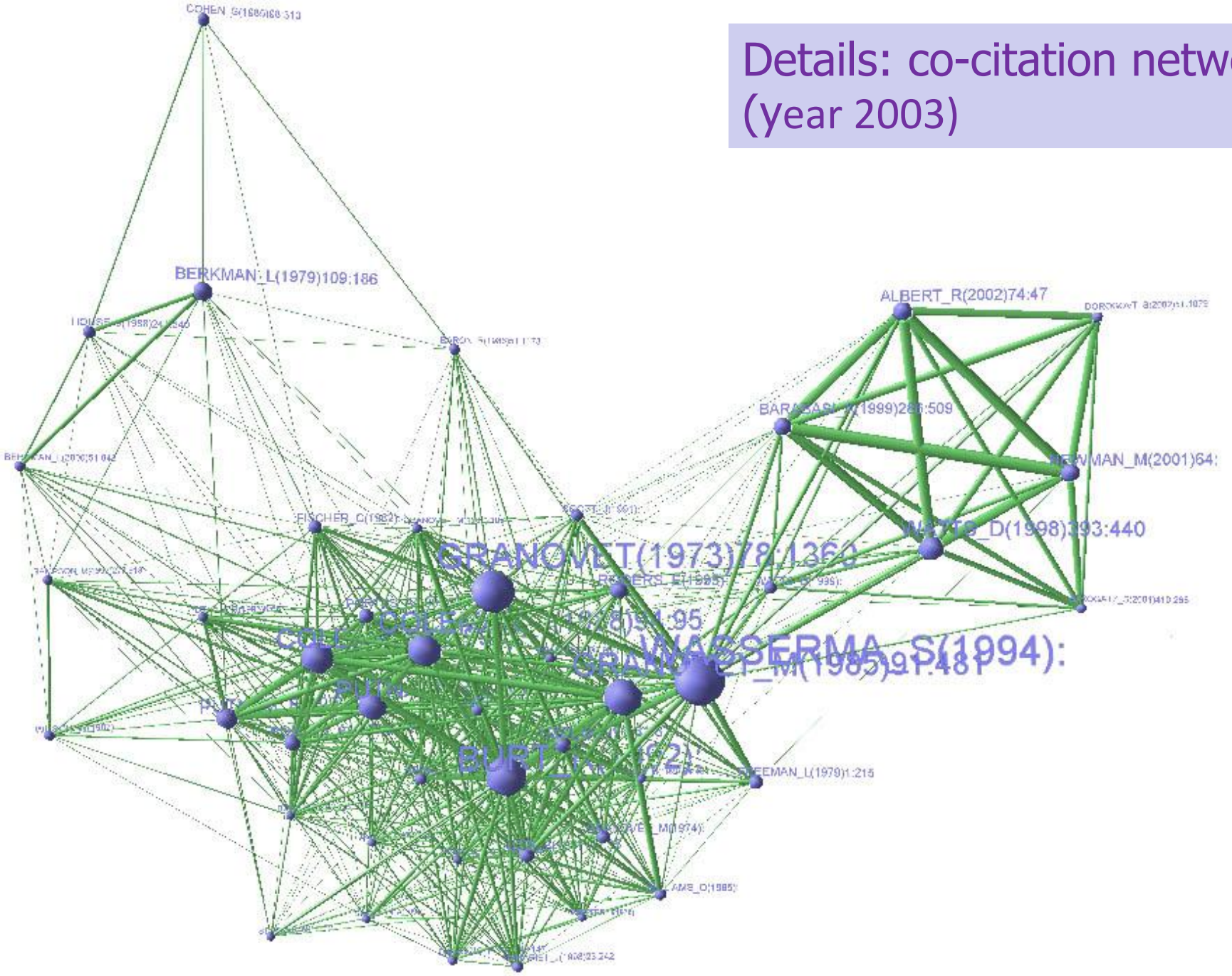
SCOTT_J(1991):, FREEMAN_L(1979)1:215 (both were cited in 1998 and 1999)

Details: Co-citation network (year 1999)



Task: Comparison (Analyse evolution) of co-citation networks

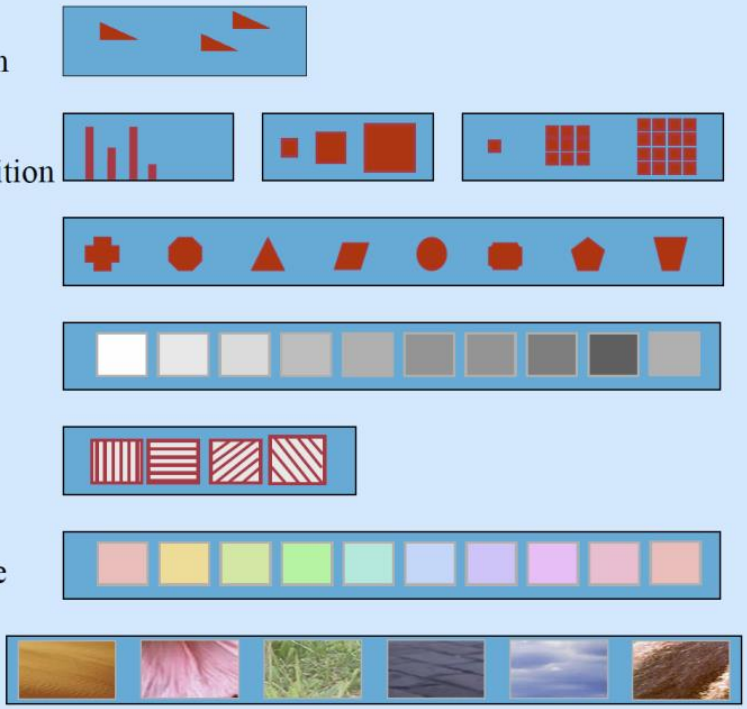
Details: co-citation network (year 2003)



2. Overlay Analysis using Marks and Visual Variables

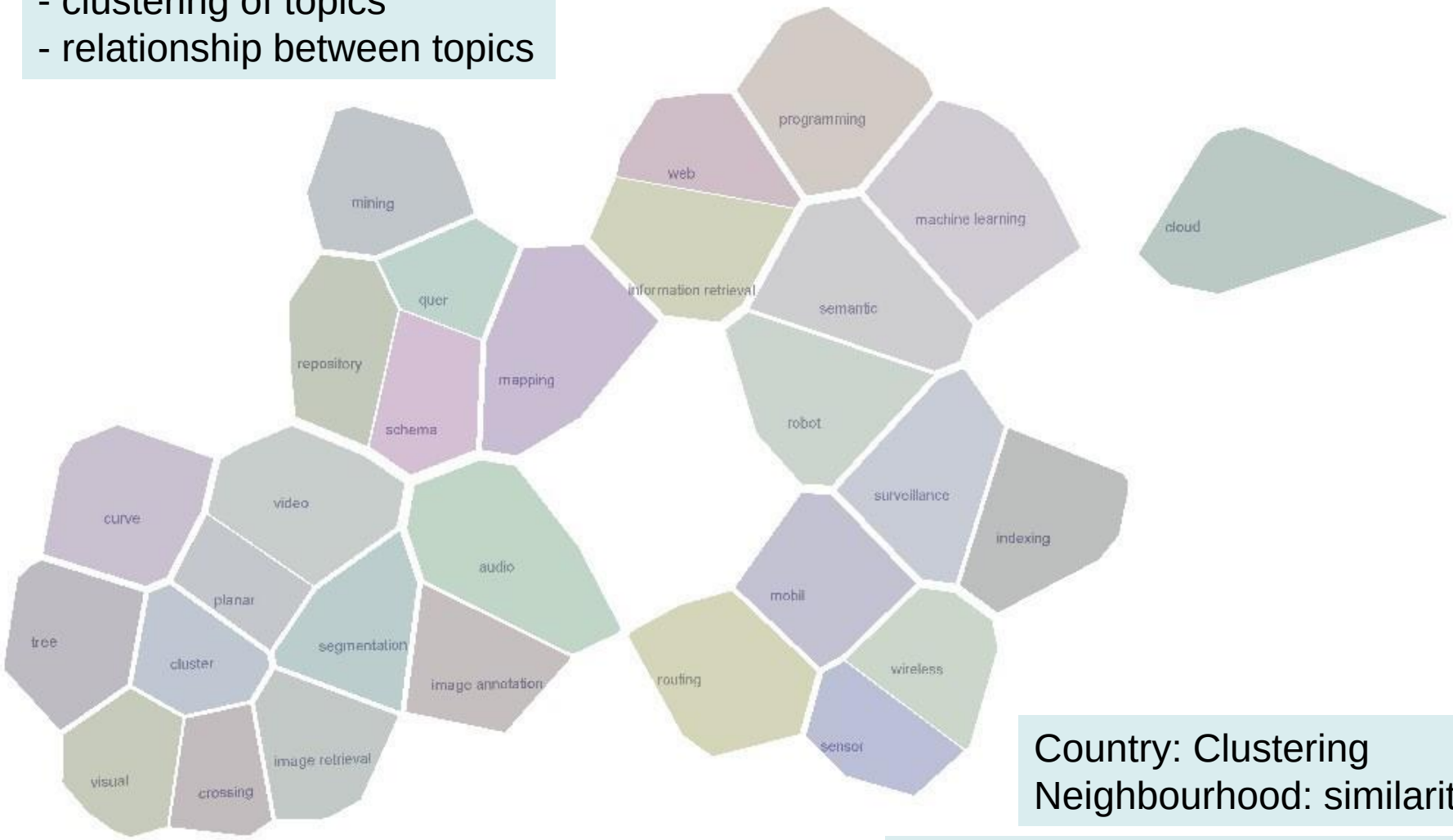
Marks
- line
- point
- area (volume)

- **position**
  - changes in the x, y, (z) location
- **size**
  - change in length, area or repetition
- **shape**
  - infinite number of shapes
- **value**
  - changes from light to dark
- **orientation**
  - changes in alignment
- **colour**
  - changes in hue at a given value
- **texture**
  - variation in pattern
- **motion**



Overview: Map of DBLP

Task:
- clustering of topics
- relationship between topics

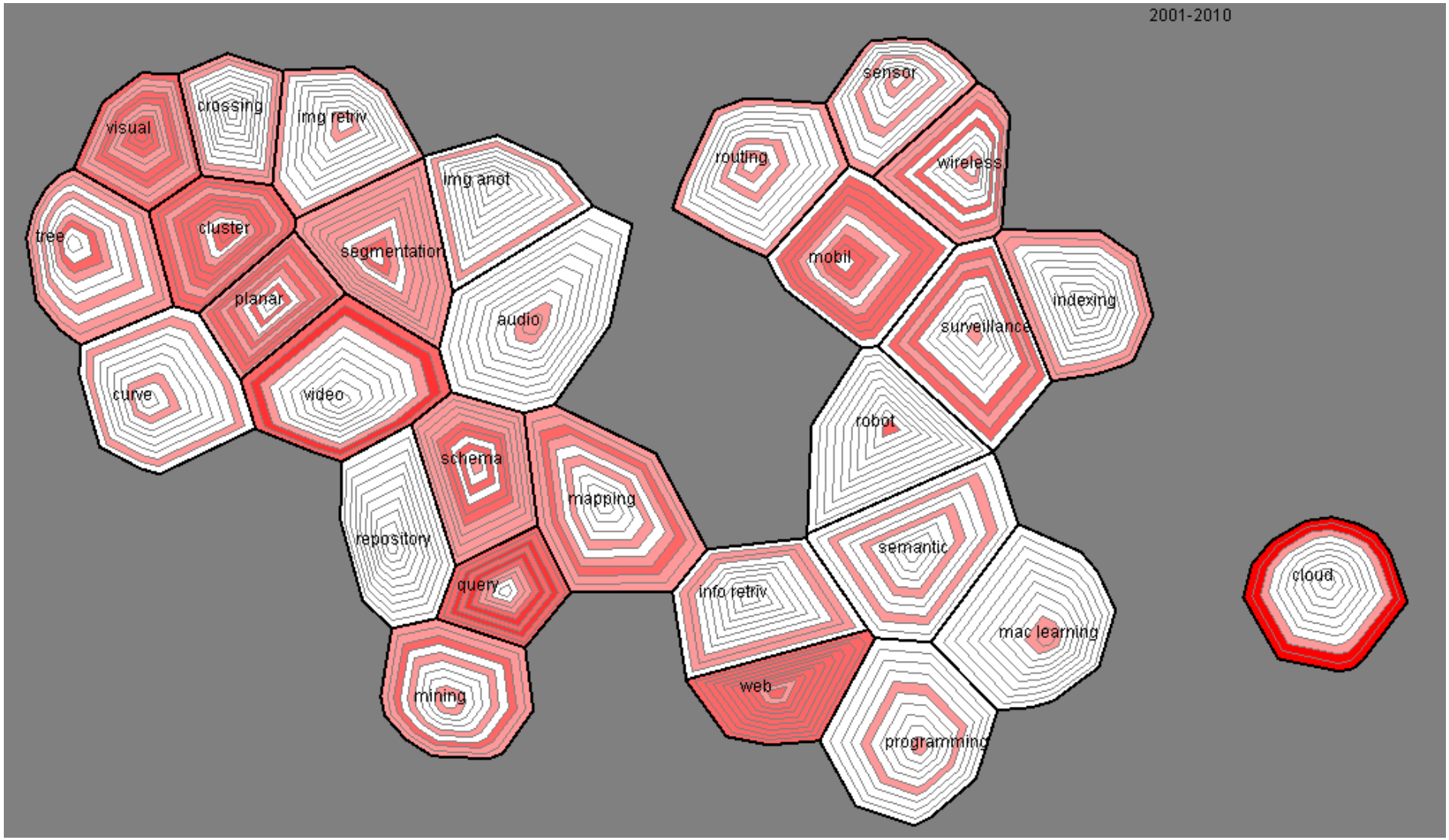


Country: Clustering
Neighbourhood: similarity

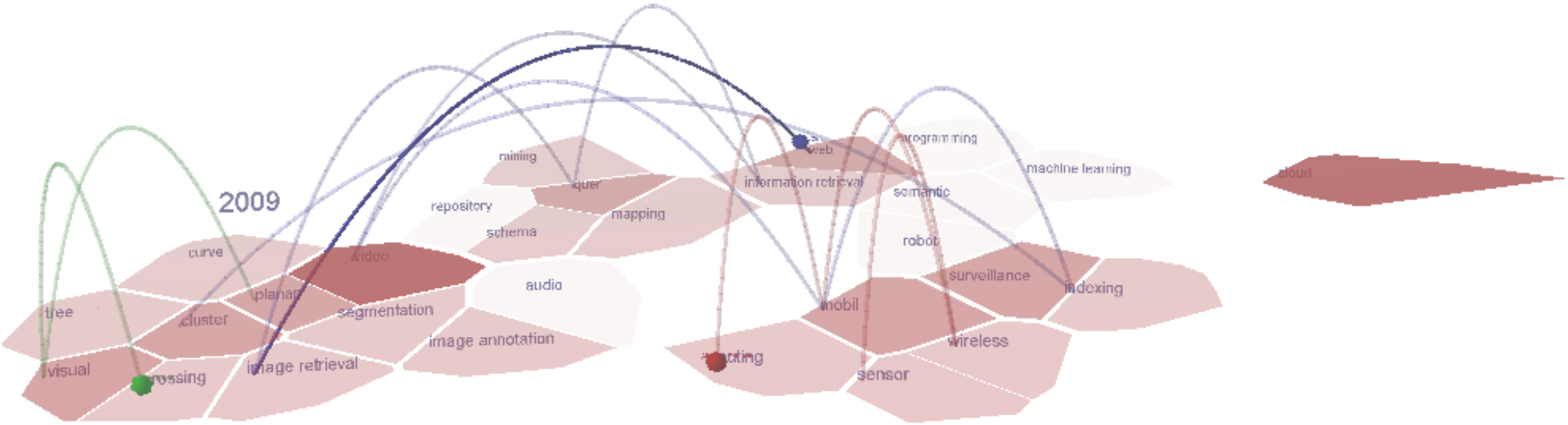
Position: MDS+ Voronoi Daigram

Temporal Map

task: trend analysis (evolution of research topics)

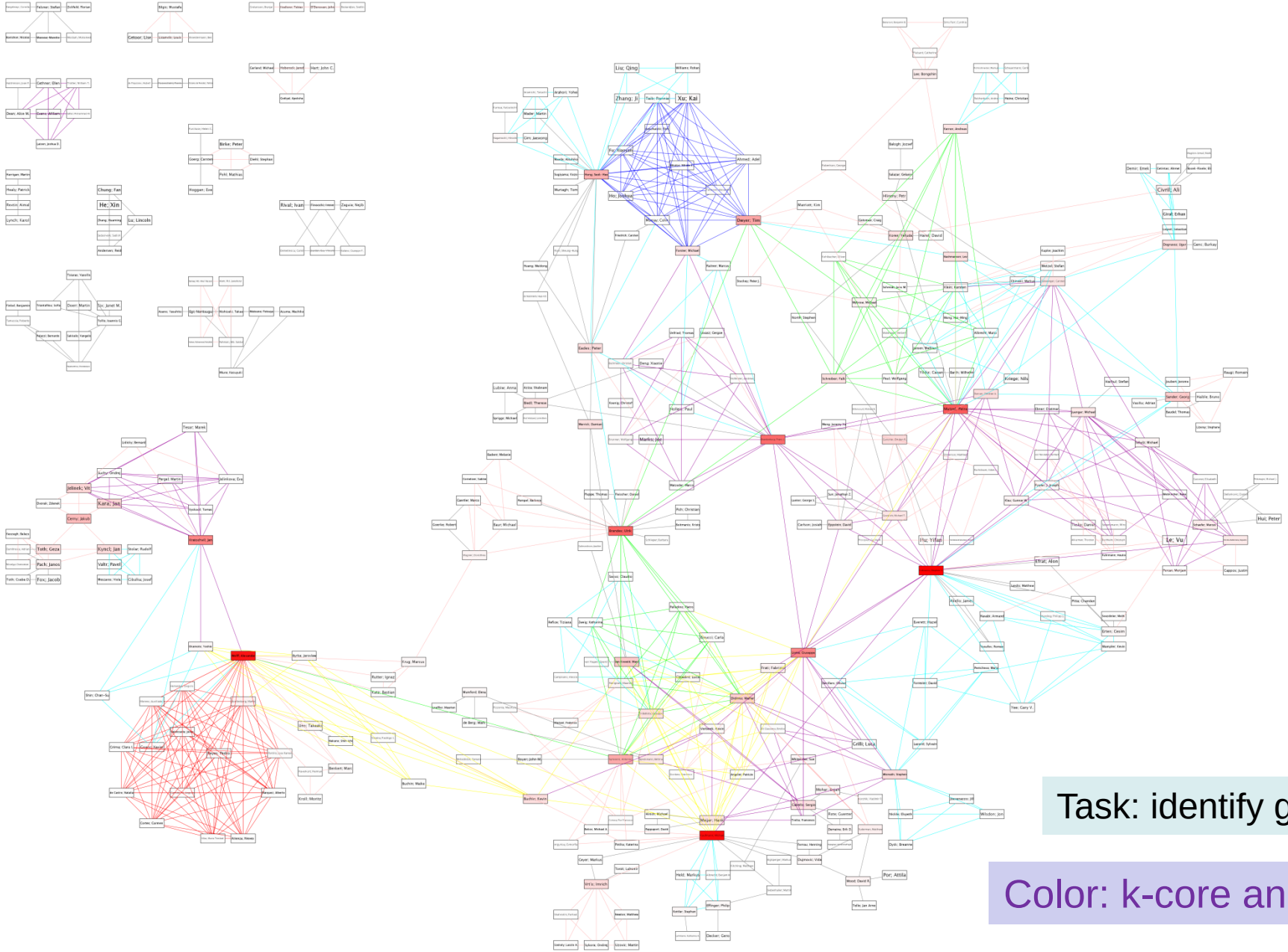


Task: Comparison of Individual Trajectories



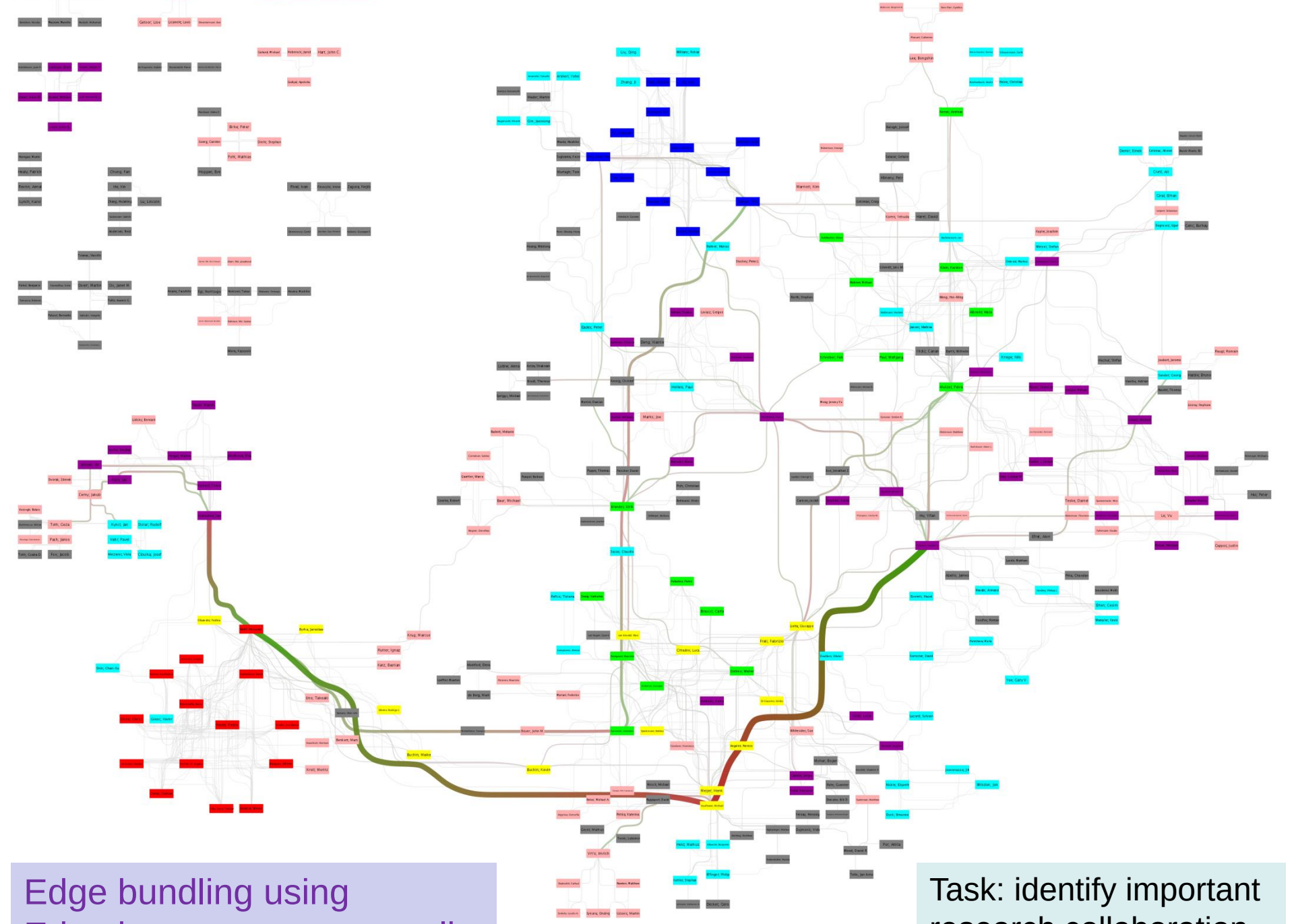
2.5D Visualisation
Animation

GD Research Collaboration Network



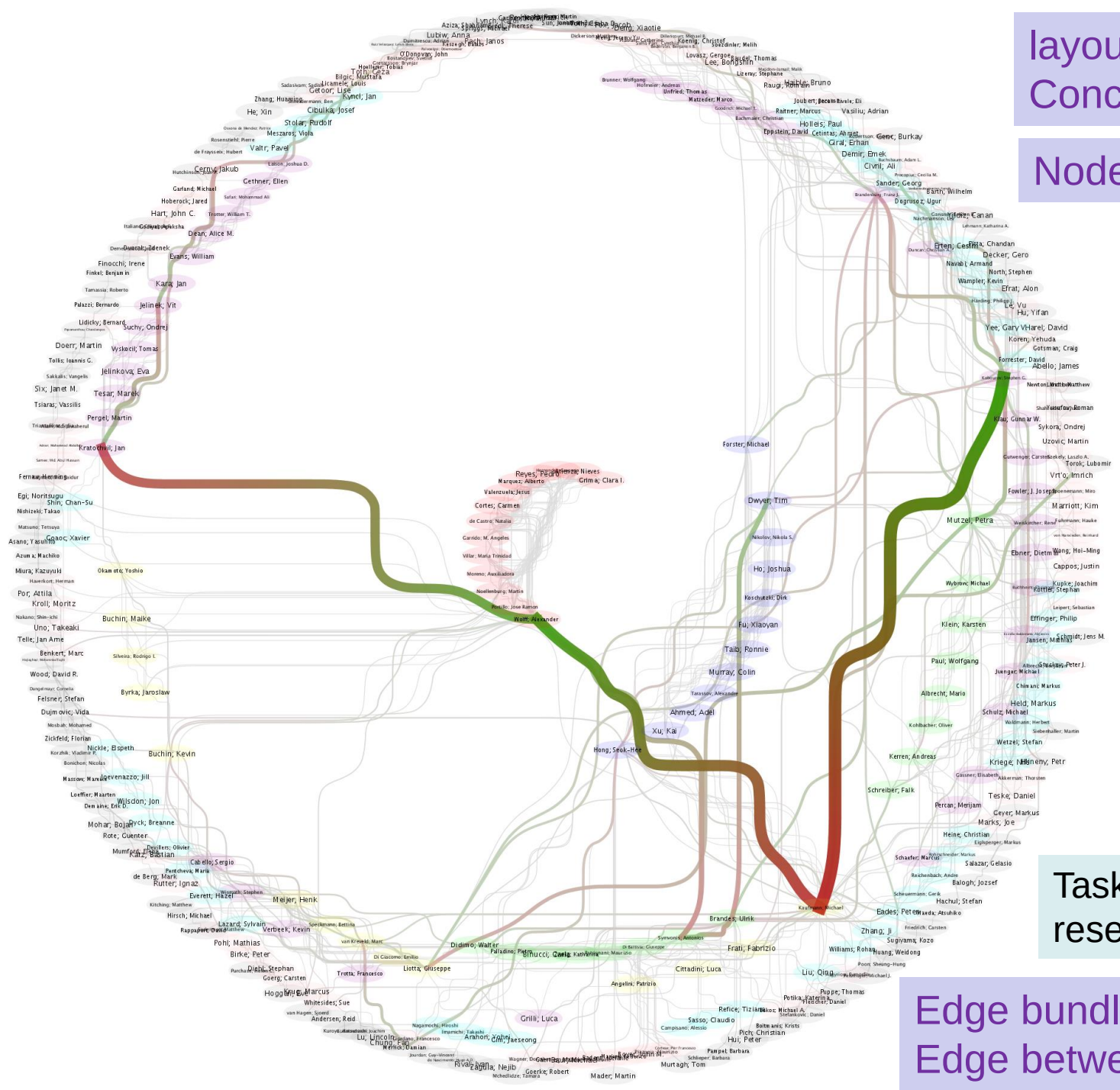
Task: identify groups

Color: k-core analysis



Edge bundling using
Edge betweenness centrality

Task: identify important research collaboration



layout: k-core analysis
Concentric circles

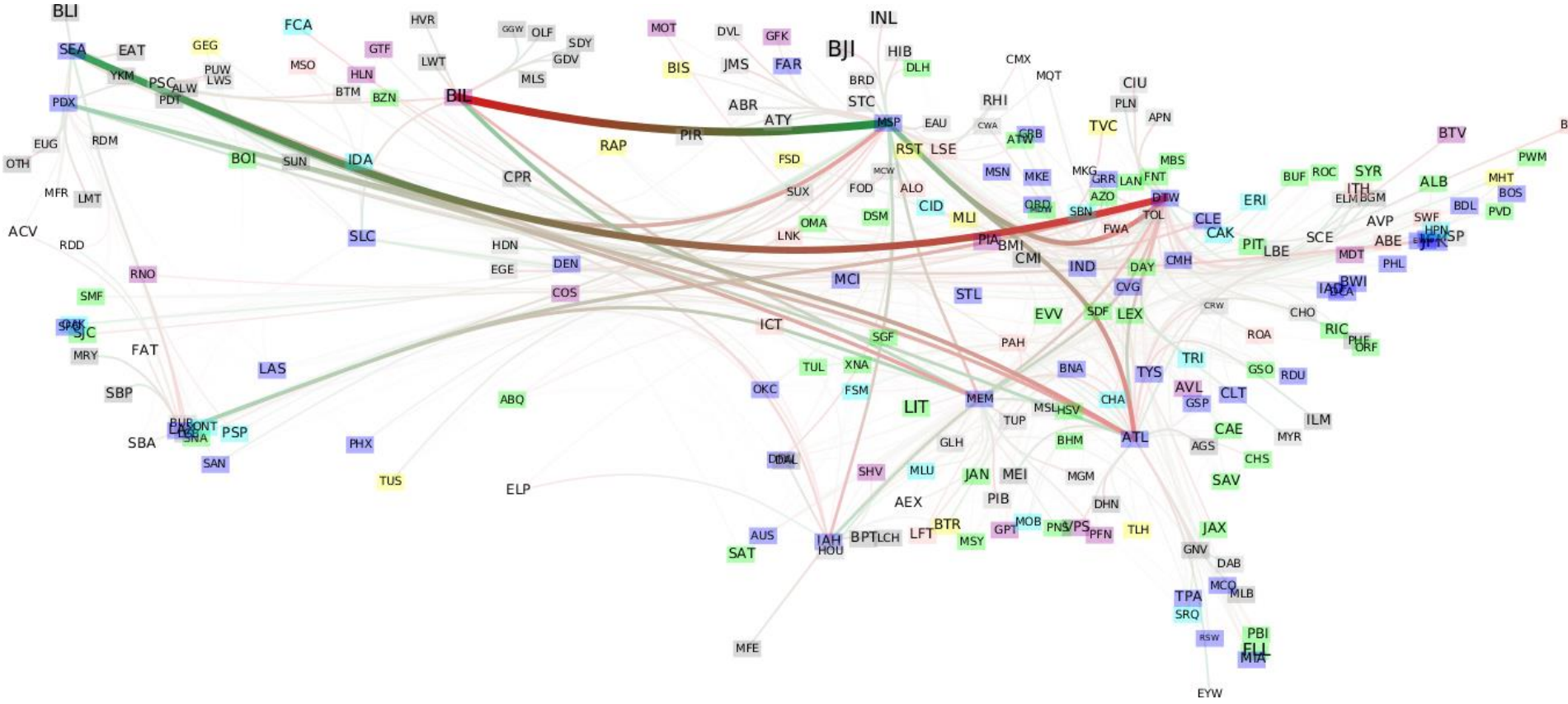
Node color: clustering

Task: identify important groups

Task: identify important research collaboration

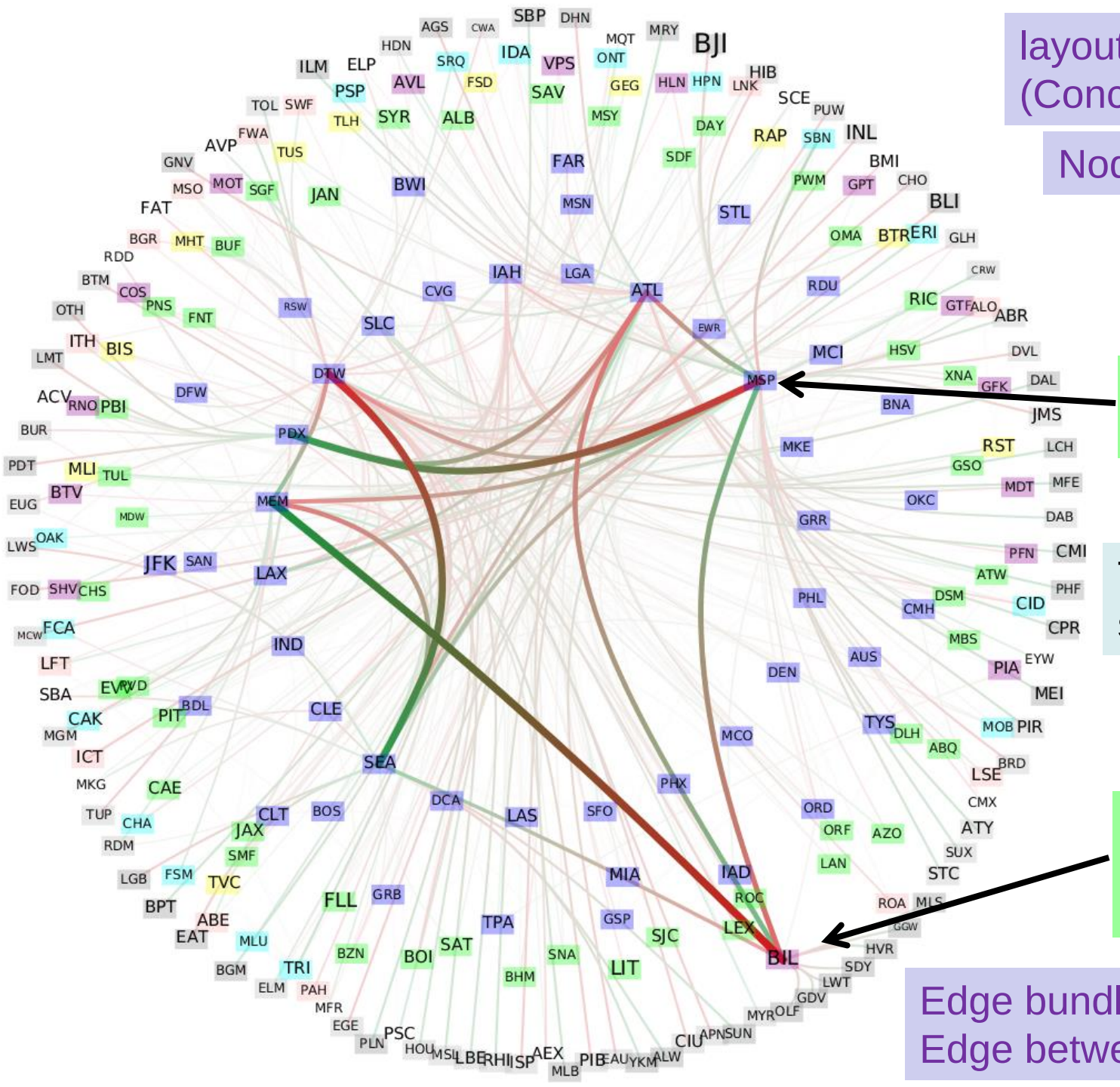
Edge bundling using
Edge betweenness centrality

US Airline Network: Traffic analysis



Task: identify significant delays

Edge bundling/thickness: Edge betweenness centrality
Node color: k-core analysis



layout: k-core analysis (Concentric circles)

Node color: k-core

Confirm expected: Hubs have delays

Task: identify significant delays

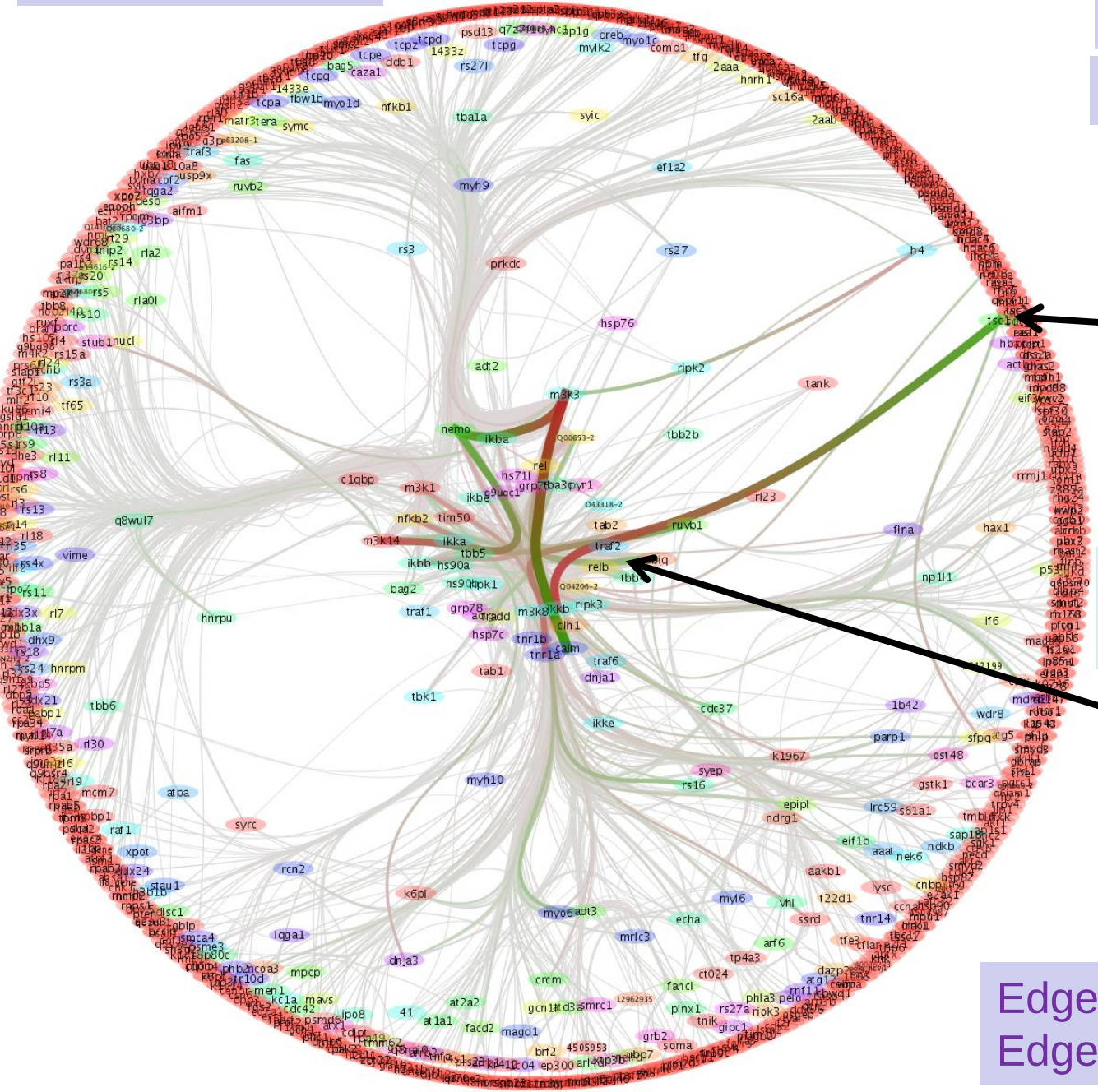
Discover Unexpected: Outlier detection

Edge bundling using Edge betweenness centrality

PPI Network

layout: k-core analysis
(Concentric circles)

Node color: clustering



Discover unexpected:
New unknown protein
- Outlier detection
- New hypothesis

Task: group analysis
(node/edge clustering)

Confirm expected:
Well-known/studied
core proteins

Edge bundling using
Edge betweenness centrality

3. Integrate a number of analysis and visualisation methods

IEEE InfoVis Competition 2004 Student Winning Entry

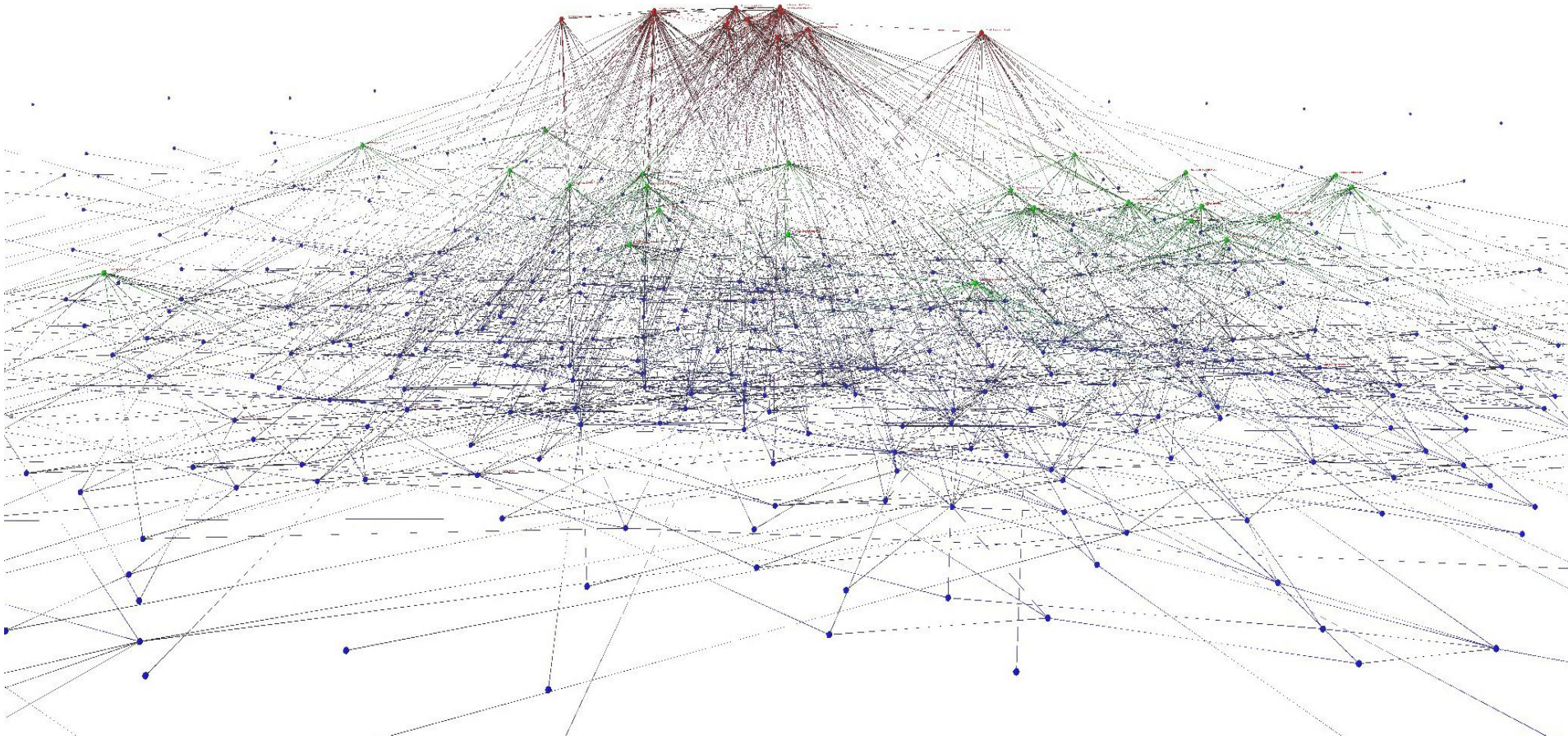
1. Citation Network

2. Collaboration Network

3. Evolution of Research Area

4. Structural Zooming and Panning (Navigation)

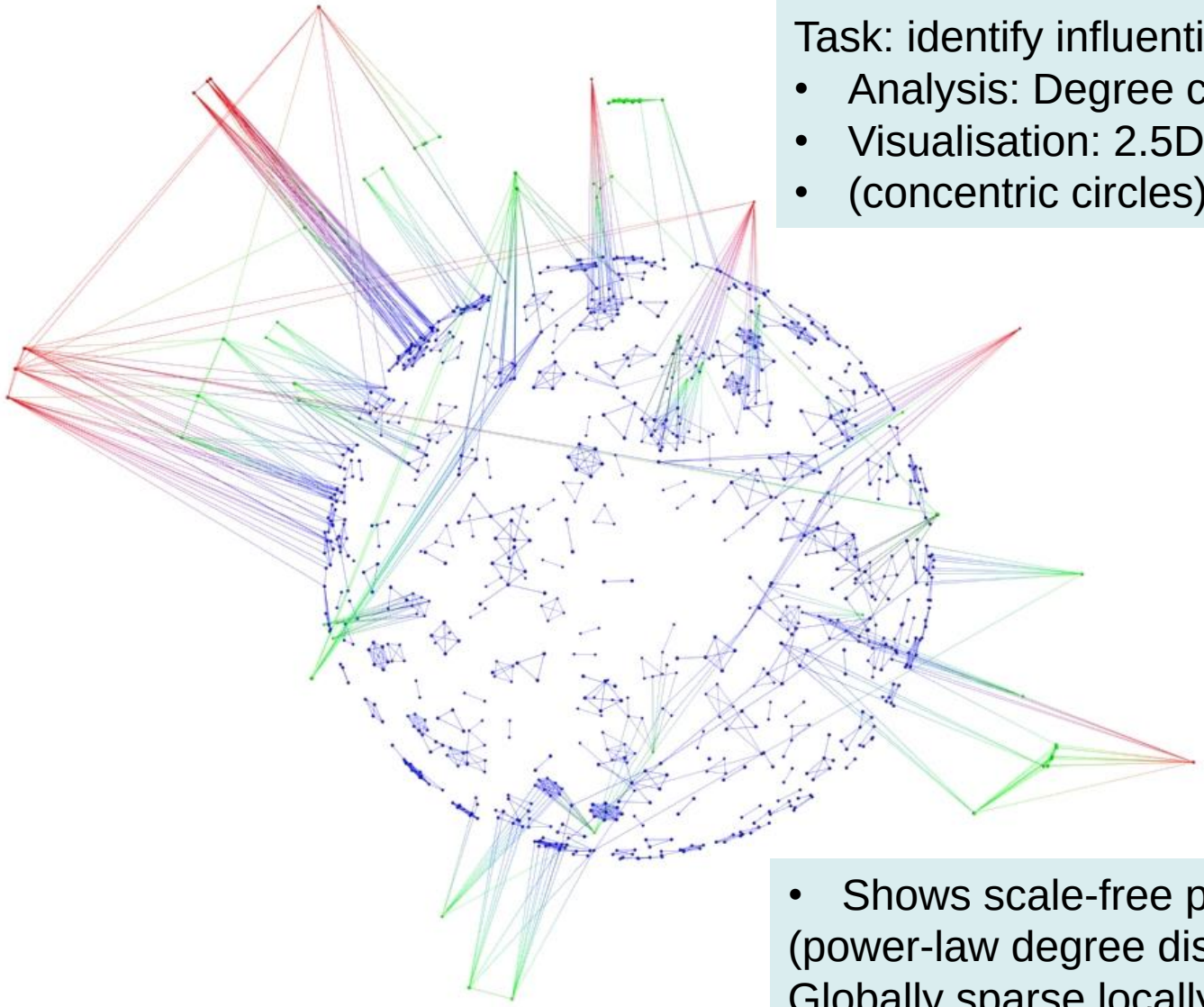
Citation Network



Task: identify seminal papers

- Analysis: Degree centrality (color, geometry)
- Visualisation: 2.5D FADE layout (parallel planes)
- Shows scale-free property (power-law degree distribution)

Collaboration Network



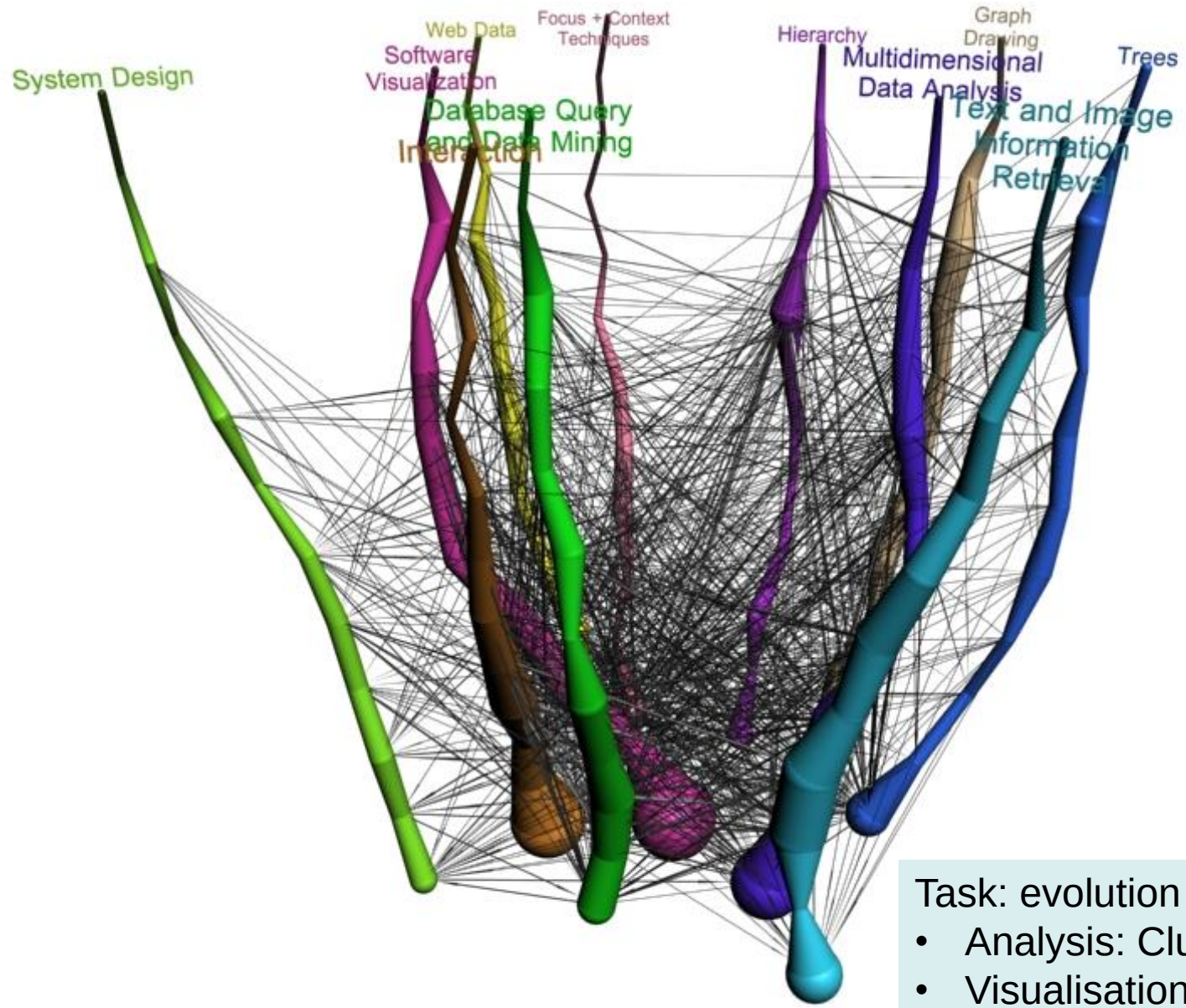
Task: identify influential researchers

- Analysis: Degree centrality (color, geometry)
- Visualisation: 2.5D FADE layout
- (concentric circles)

- Shows scale-free property

(power-law degree distribution, small diameter, Globally sparse locally dense clustering)

Overview: Evolution of Research Area



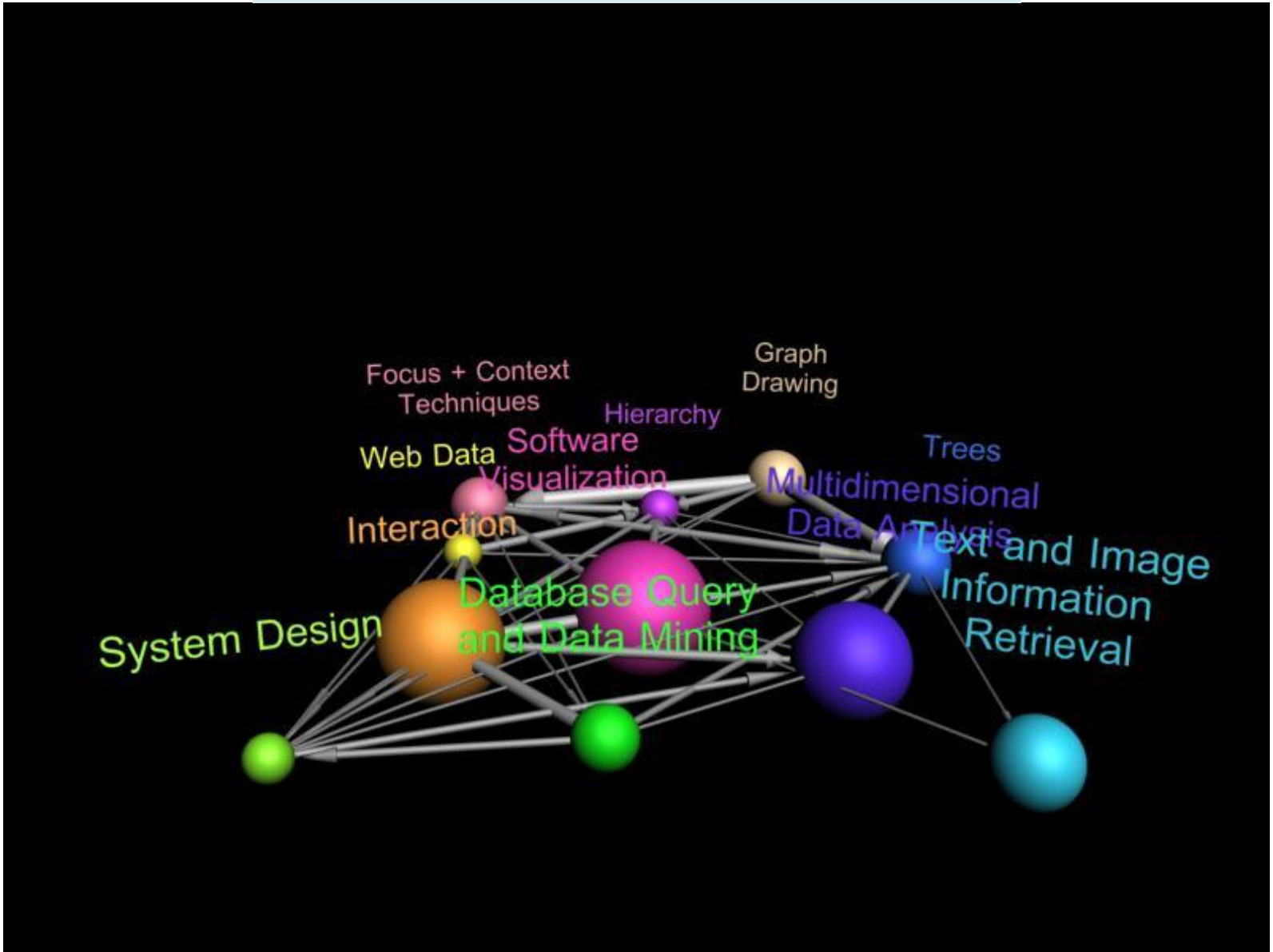
Task: evolution of research topics

- Analysis: Clustering (color)
- Visualisation: 2.5D spring algorithm

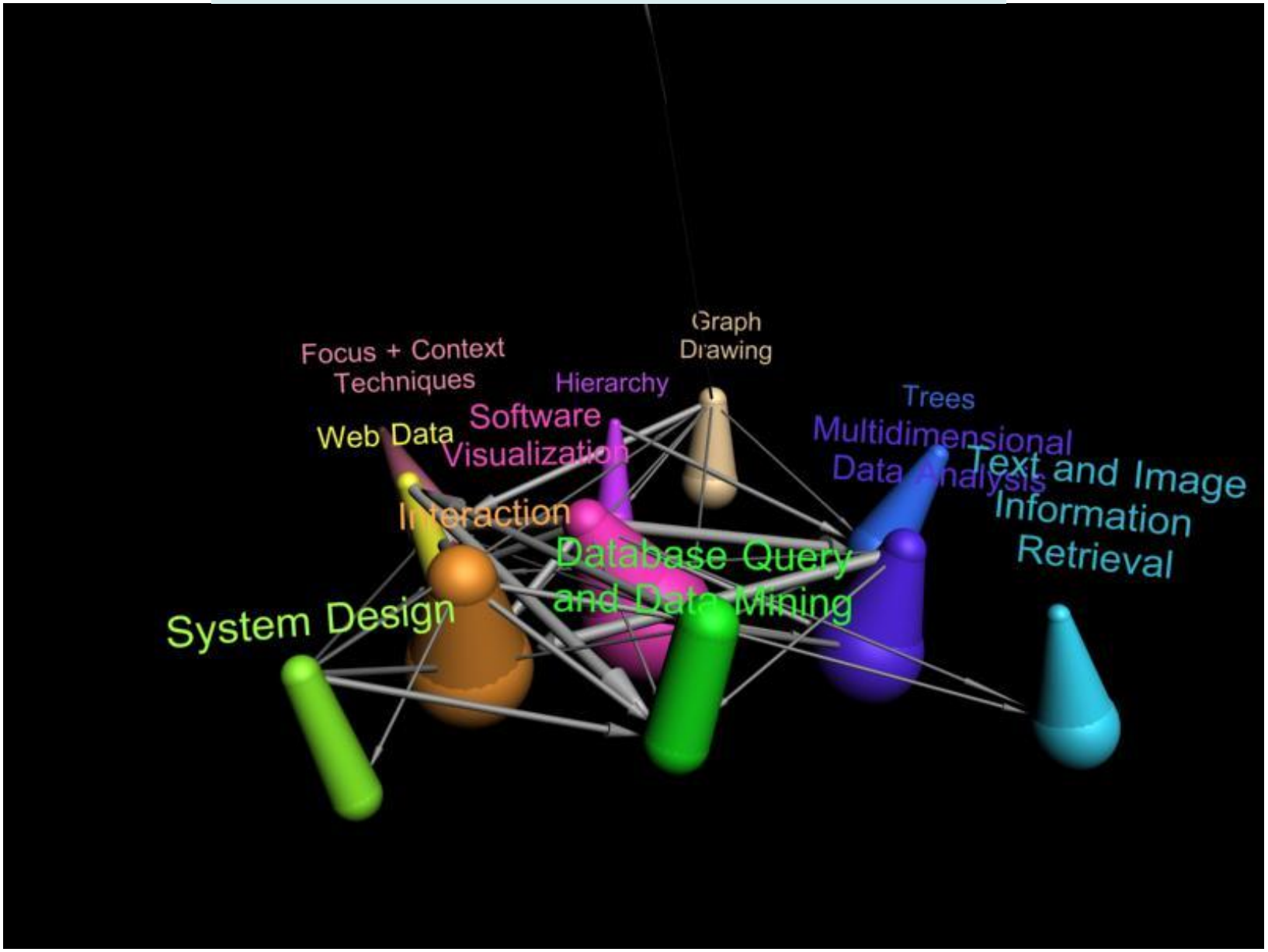
(time: parallel planes)

Details: Year 1993

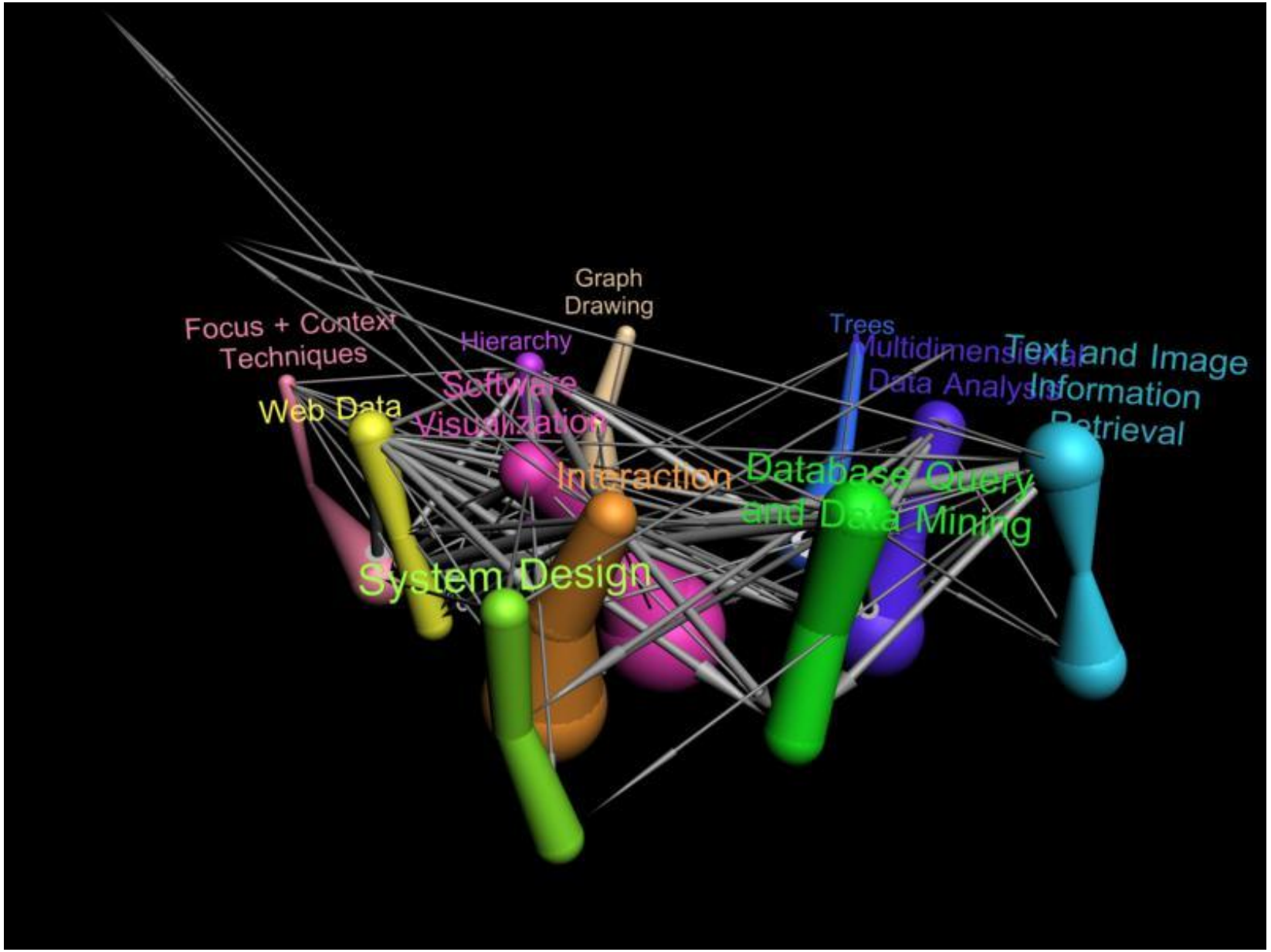
- color: clustering (11 research topics)
- size: # of papers (popularity of the topic)
- edge: citation



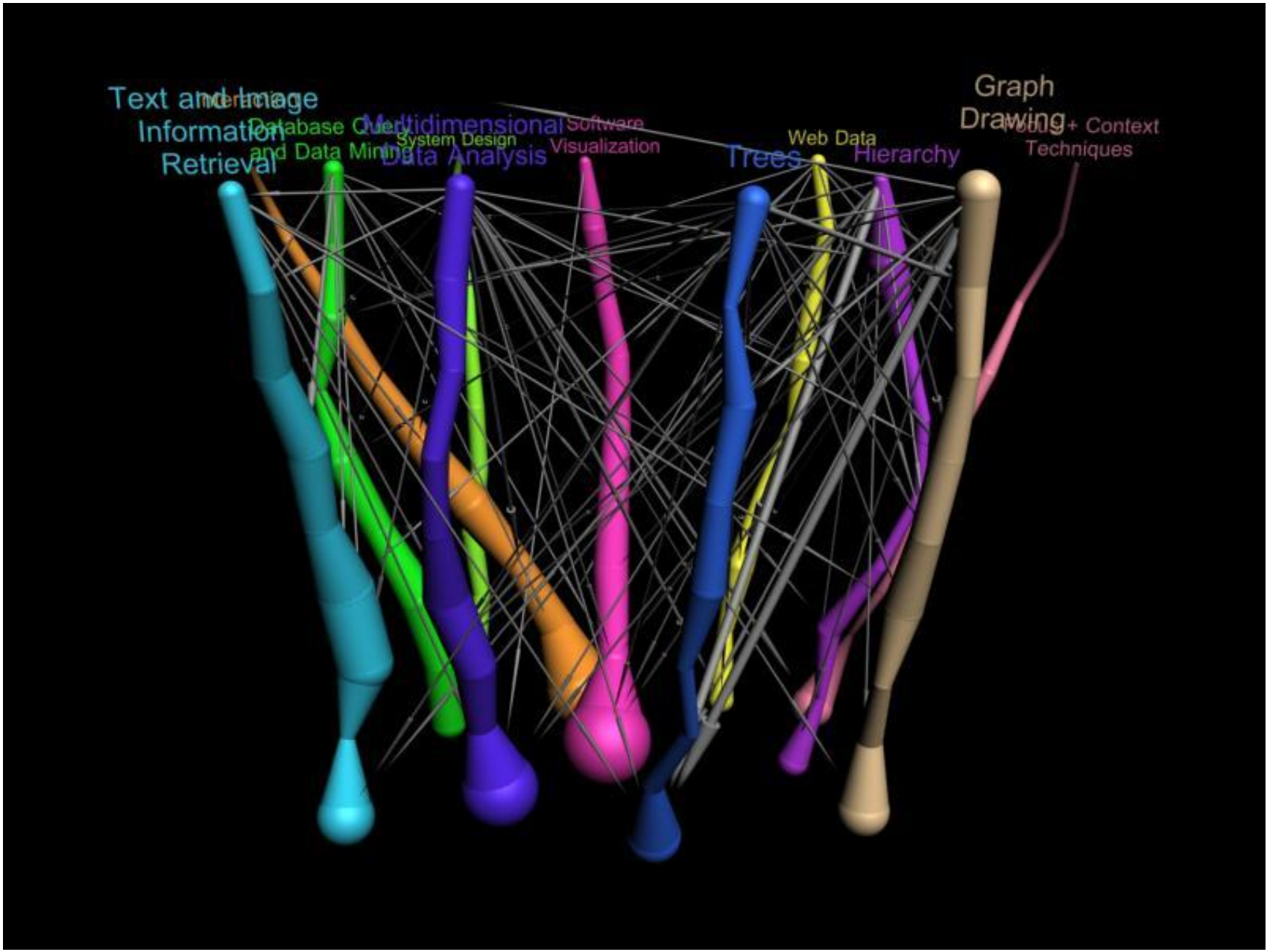
Details: Thickness of column
- popularity of research topic: trend analysis)
- weighted by the number of citations



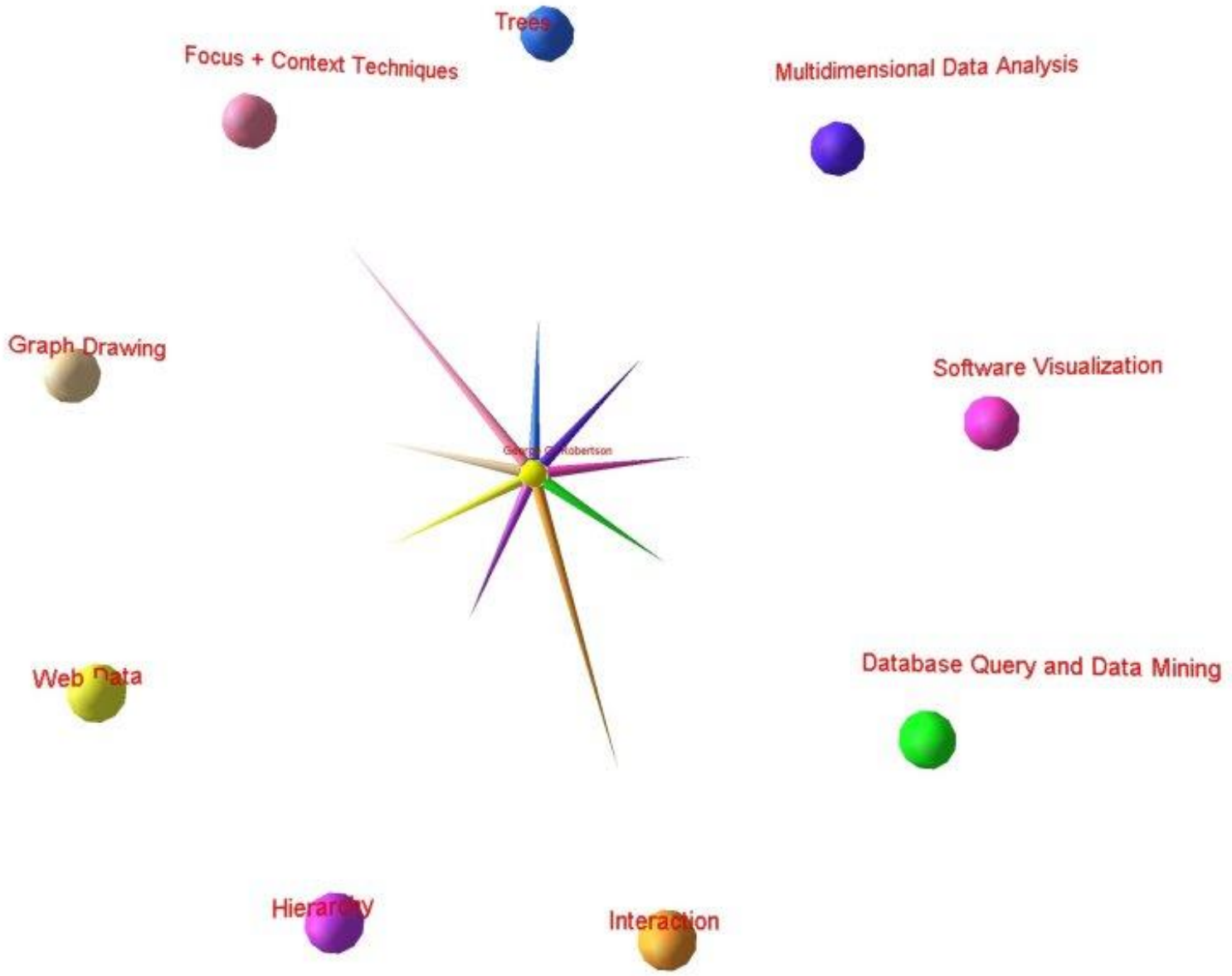
Year 1995: reference to future publication (to appear)



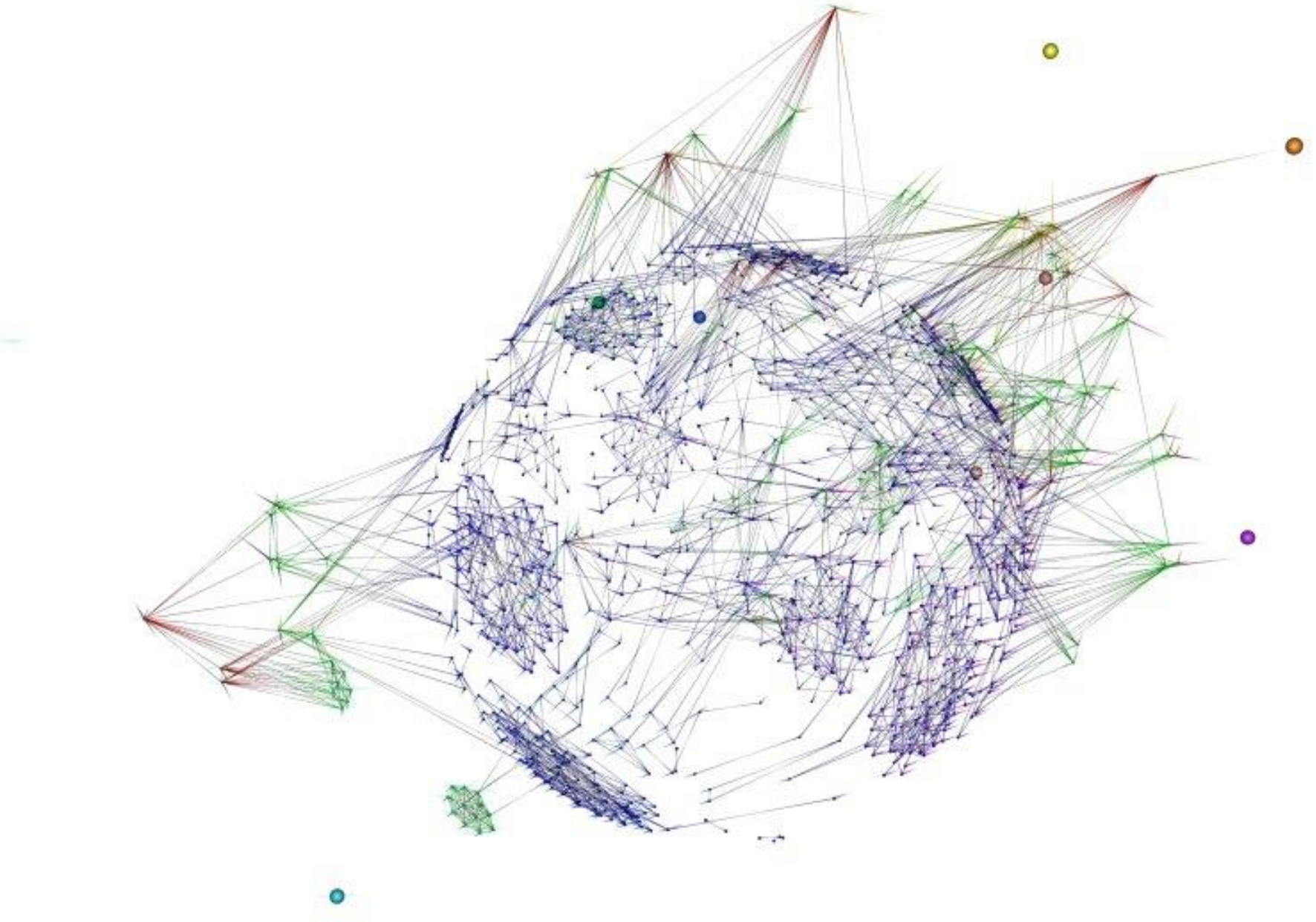
Direction of column: relations between research topics



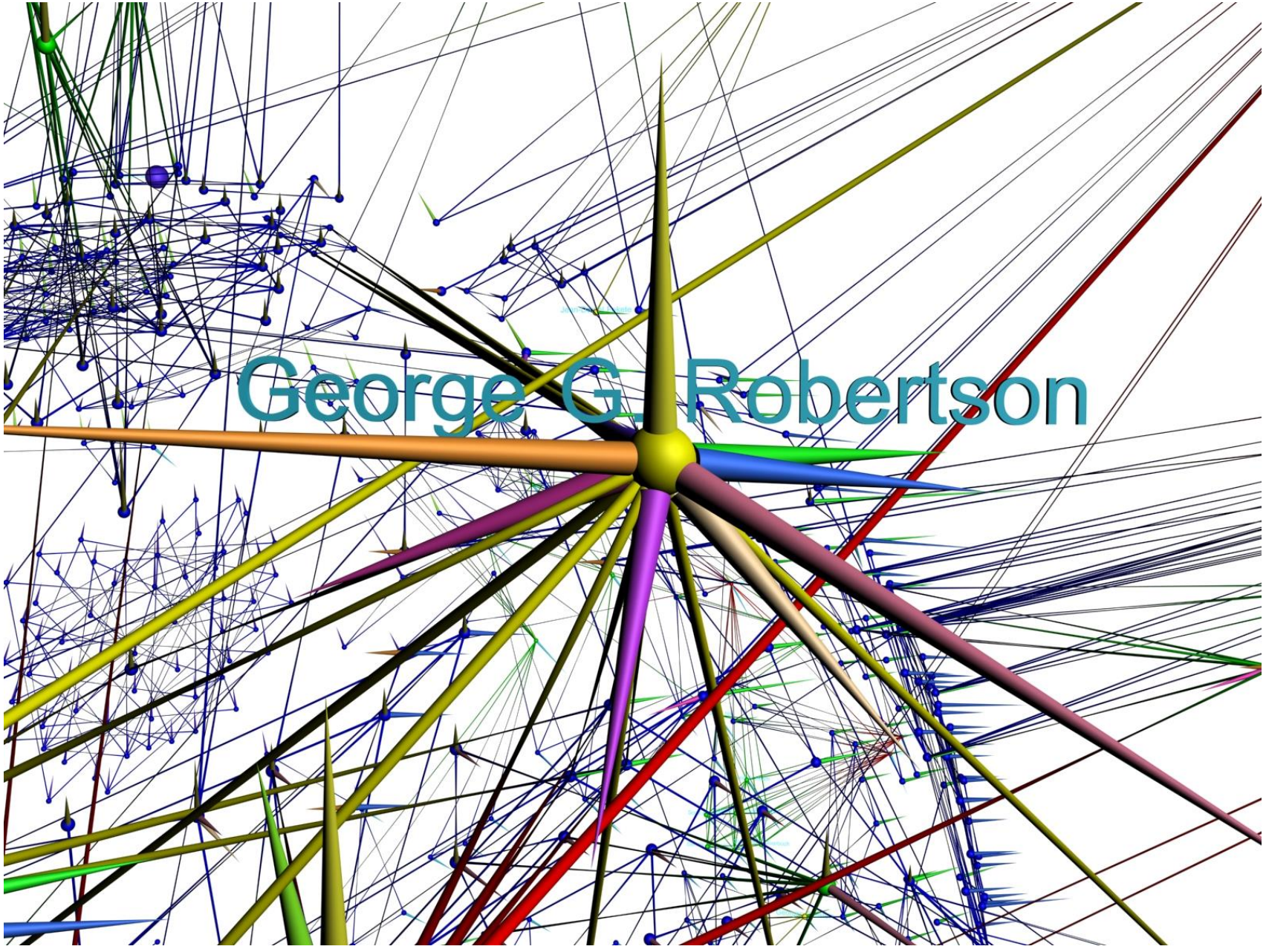
Glyph: Researcher with research topics



Collaboration network + Glyphs:
Important researchers with their research topics



Navigation: Structural Zooming and Panning



Movie: InfoVis 2004 contest

GD Contest 2006
History of World Cup
1st Place

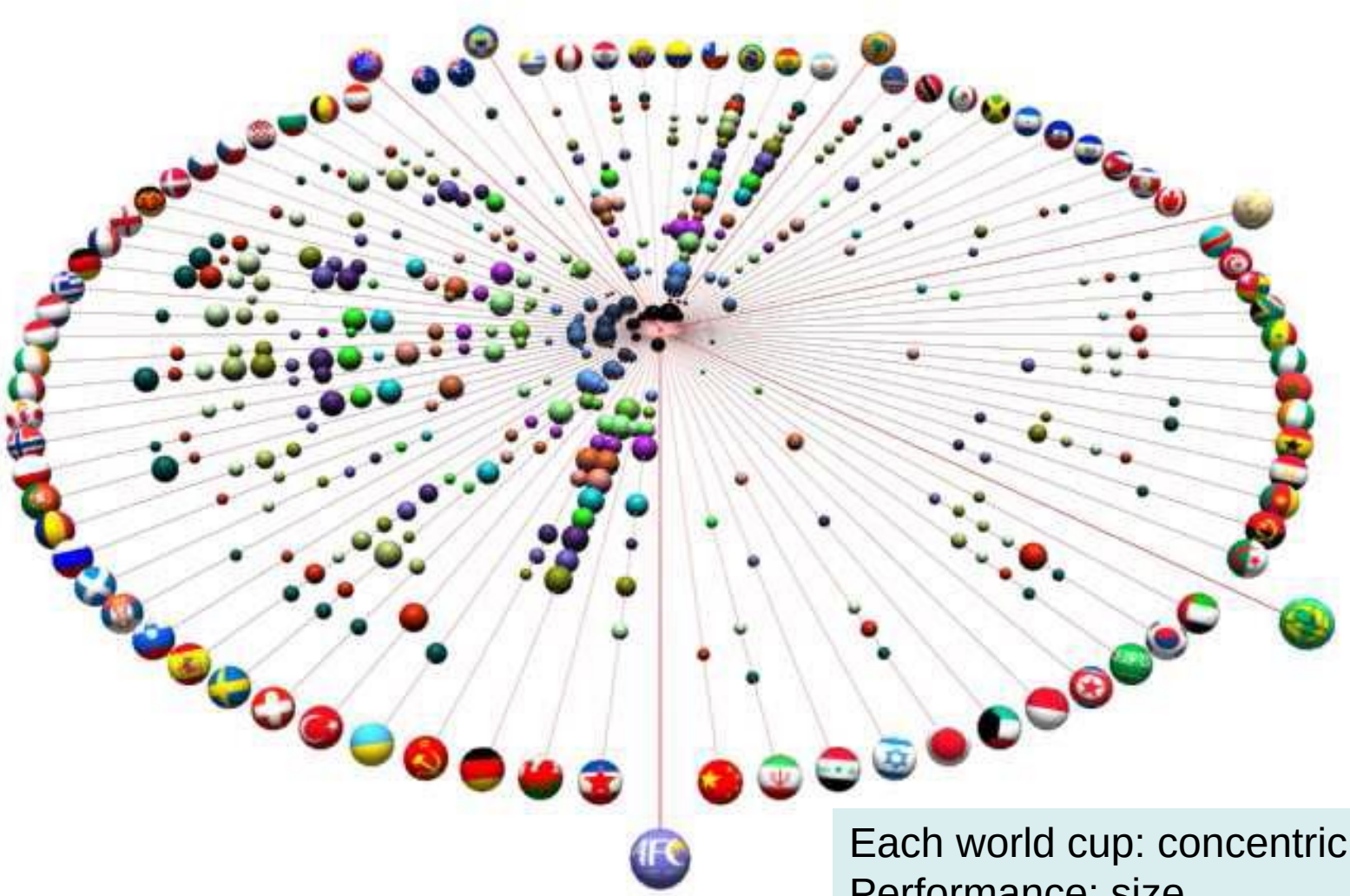
Task: Evolution

1. Wheel Layout

2. Radial Layout

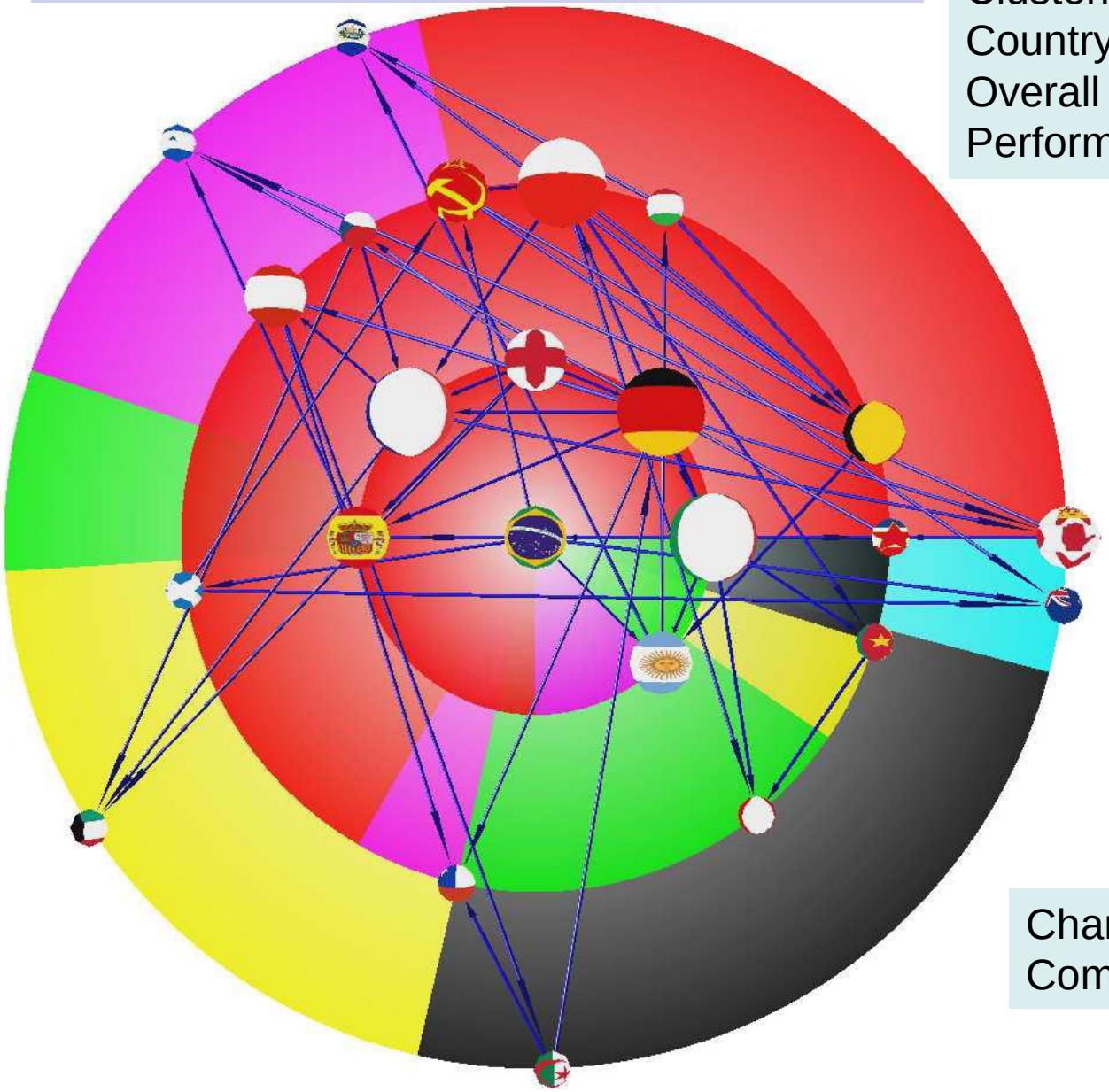
3. Hierarchical Layout

Overview: Wheel Layout



Each world cup: concentric circle
Performance: size
Clustering (continents): wedge
Country: texture

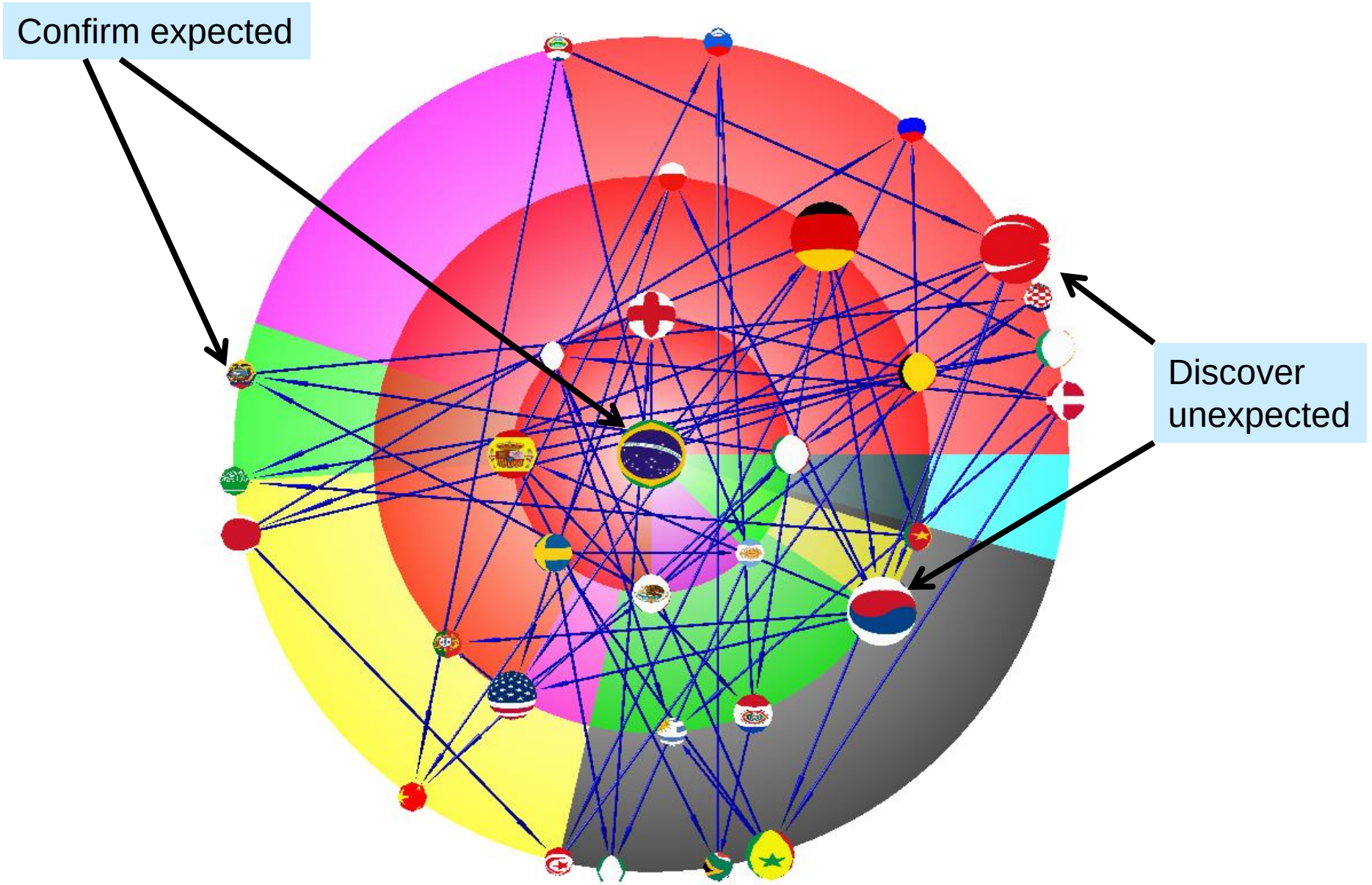
Details: Radial Layout



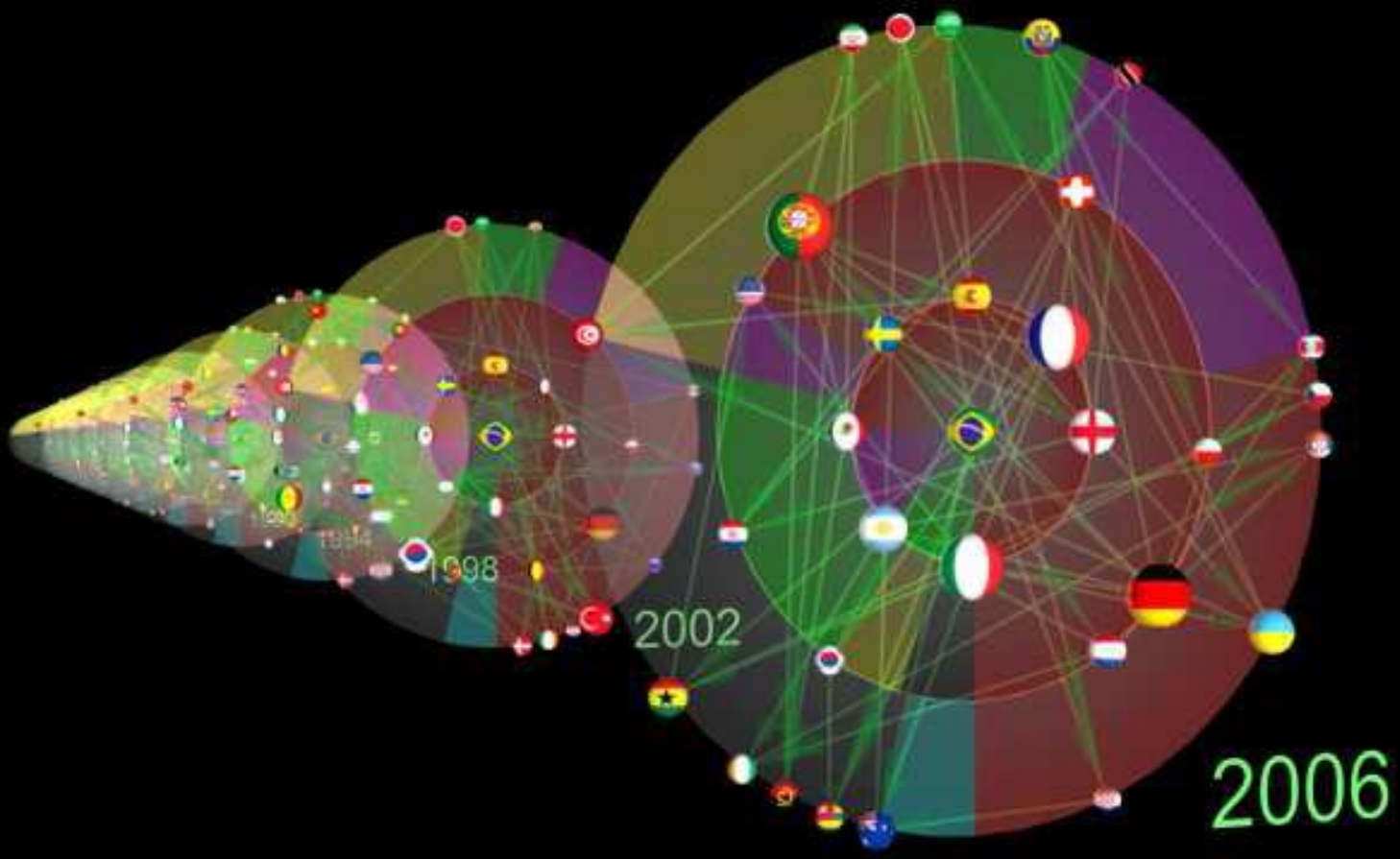
Position: Union graph approach
Concentric circles: Degree centrality
Clustering (continents): color
Country: texture
Overall performance: position
Performance of specific year: size

Change of rules: only 24 teams
Comparison between continents

Details: Year 2002



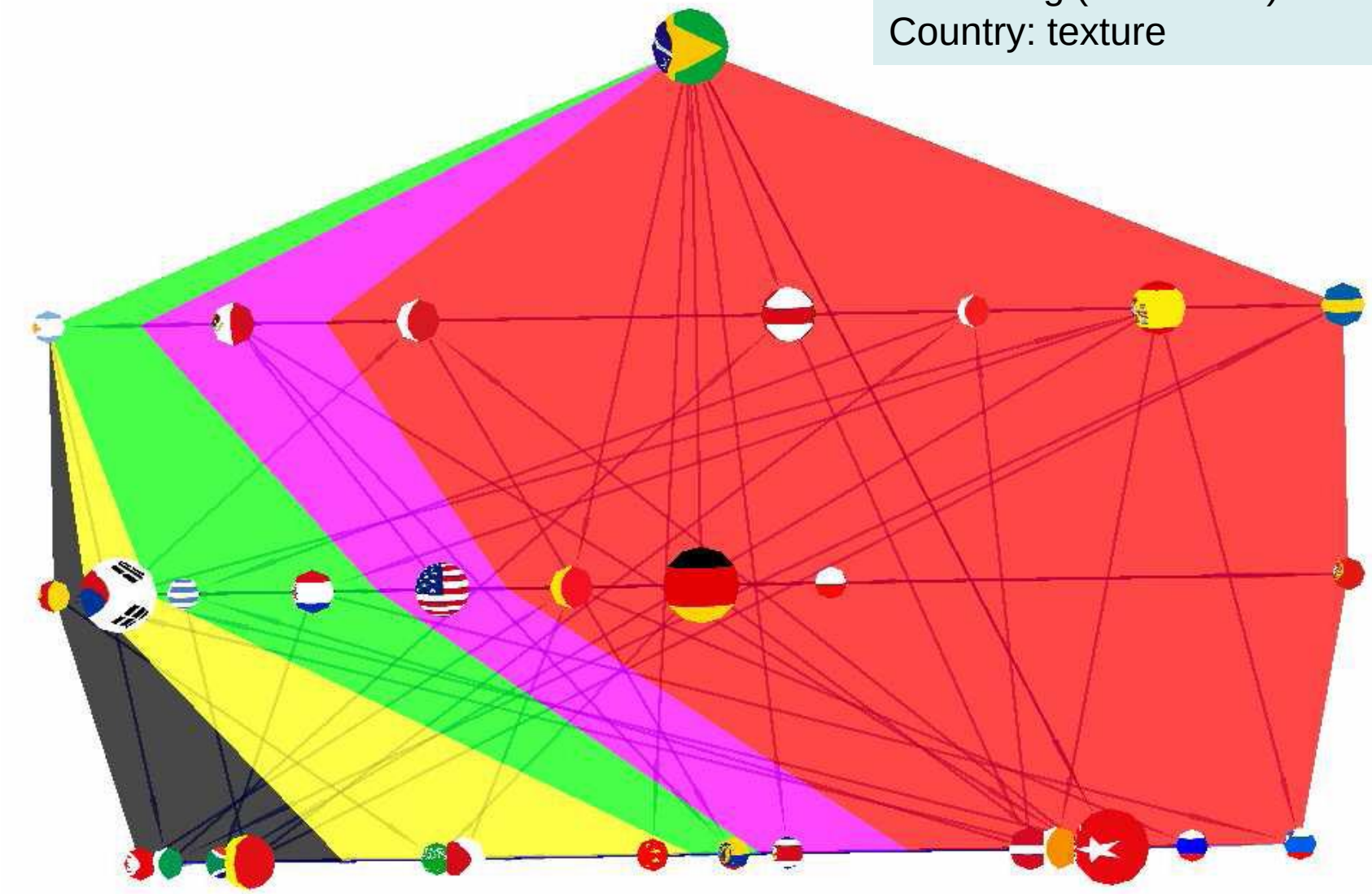
History of World Cup: animation



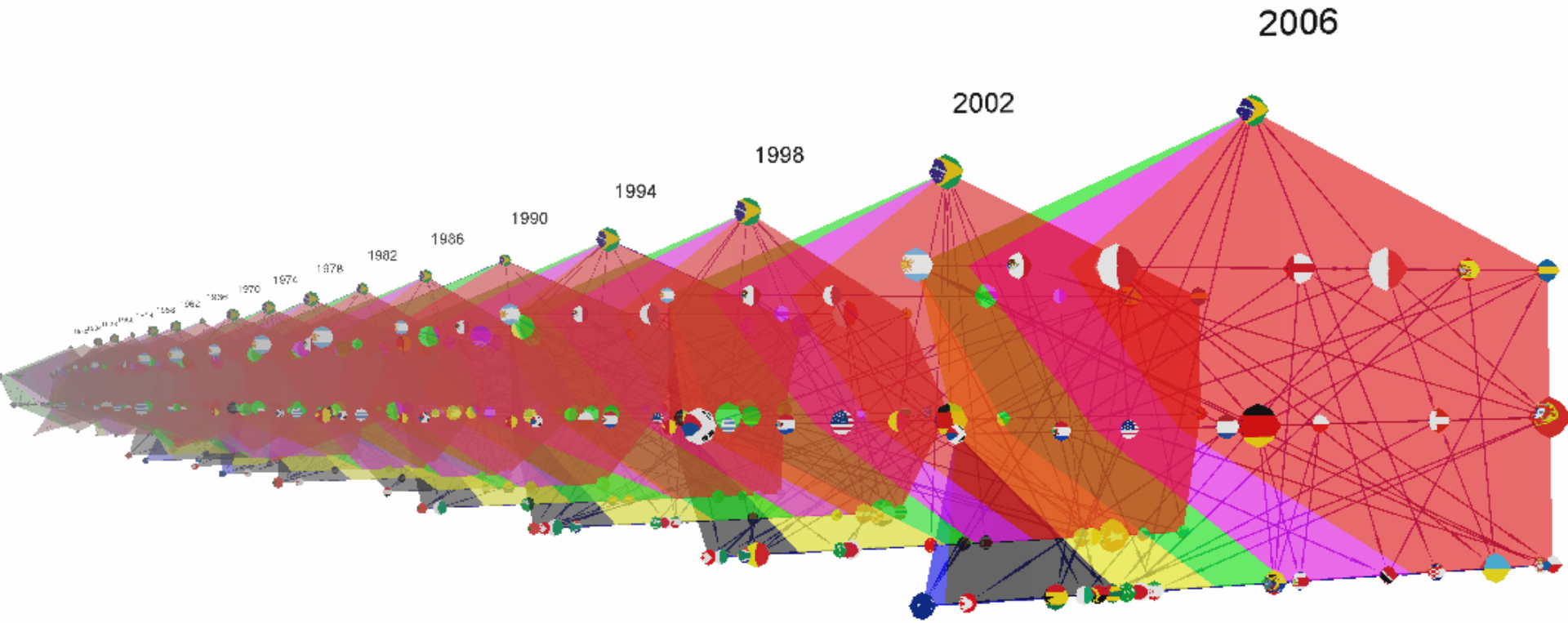
Hierarchical Layout

2002

Union graph approach
Degree centrality: hierarchy (level)
Overall performance: position
Performance of specific year: size
Clustering (continents): color
Country: texture



History of World Cup



Movie: GD 2006 contest

4. Big/Complex data sets, Reduce/Filter the data sets (size, dimension, volume)

GD Contest 2005
Evolution of IMDB 1st Place
Sunbelt Viszard Session

Task: Evolution (temporal dynamics)

1. Actor collaboration network

2. Actor-Company Clustered Network

1. Actor Collaboration Network (Evolution of Kevin Bacon Network)

1. Evolution of Kevin Bacon Network

- Original data set:
  - Graph Drawing Contest 2005

- Filtering:
  - Only actors with a ***Kevin Bacon number of 1***
  - These actors and their movies were broken into ***decade****s*
  - ***actor collaboration (co-starring) network***
  - Edge weights: number of movies in which two actors co-starred, within each decade

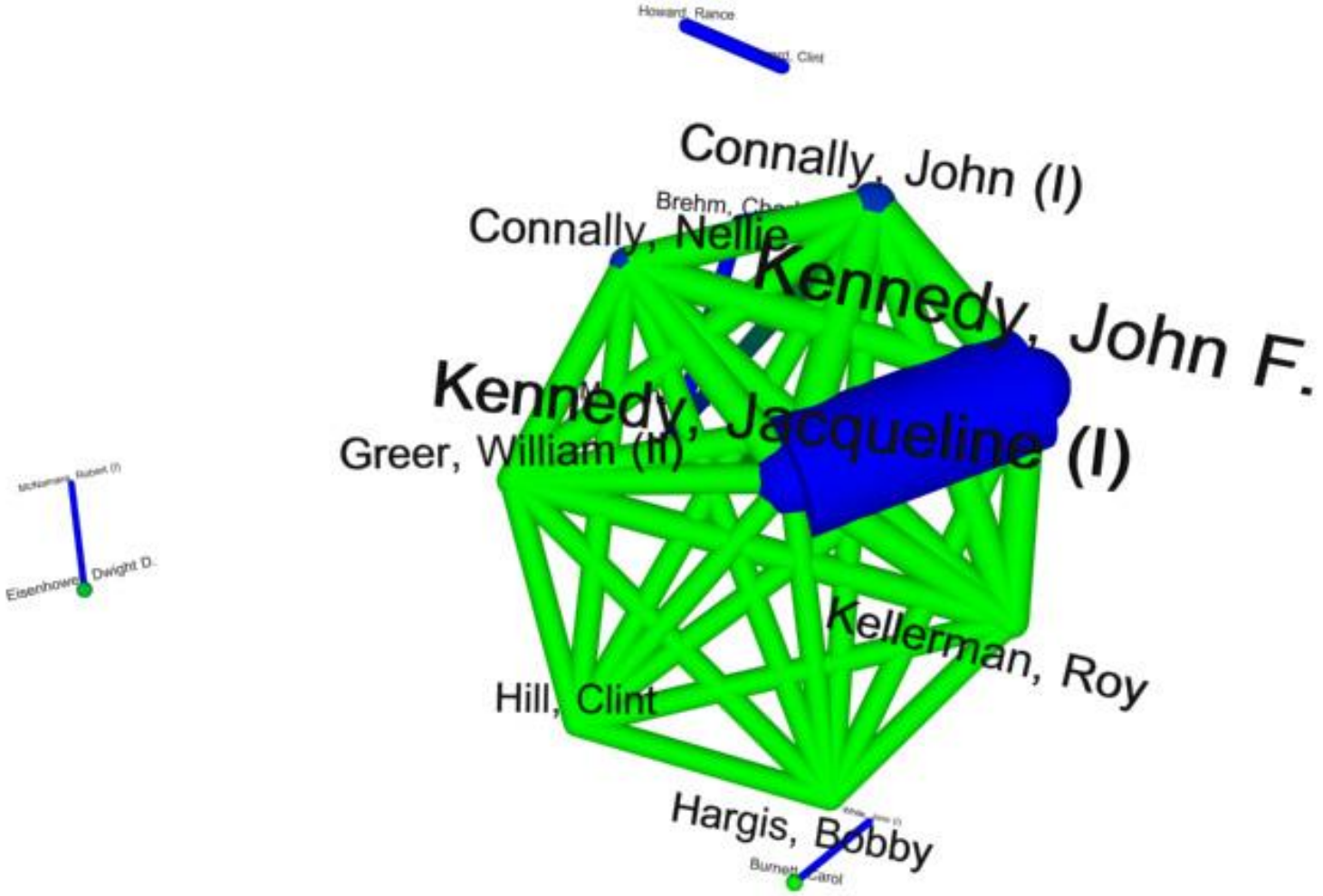
The Data Set

Graph	# Vertices	# Edges
Original	1 324 748	3 792 390
KB #1 (all decades, no edge weight filtering	2 742	336 060
KB #1 (1910s, filtered)	16	18
KB #1 (1920s, filtered)	4	2
KB #1 (1930s, filtered)	25	53
KB #1 (1940s, filtered)	17	17
KB #1 (1950s, filtered)	19	18
KB #1 (1960s, filtered)	16	35
KB #1 (1970s, filtered)	79	411
KB #1 (1980s, filtered)	59	73
KB #1 (1990s, filtered)	207	425
KB #1 (2000s, filtered)	124	208

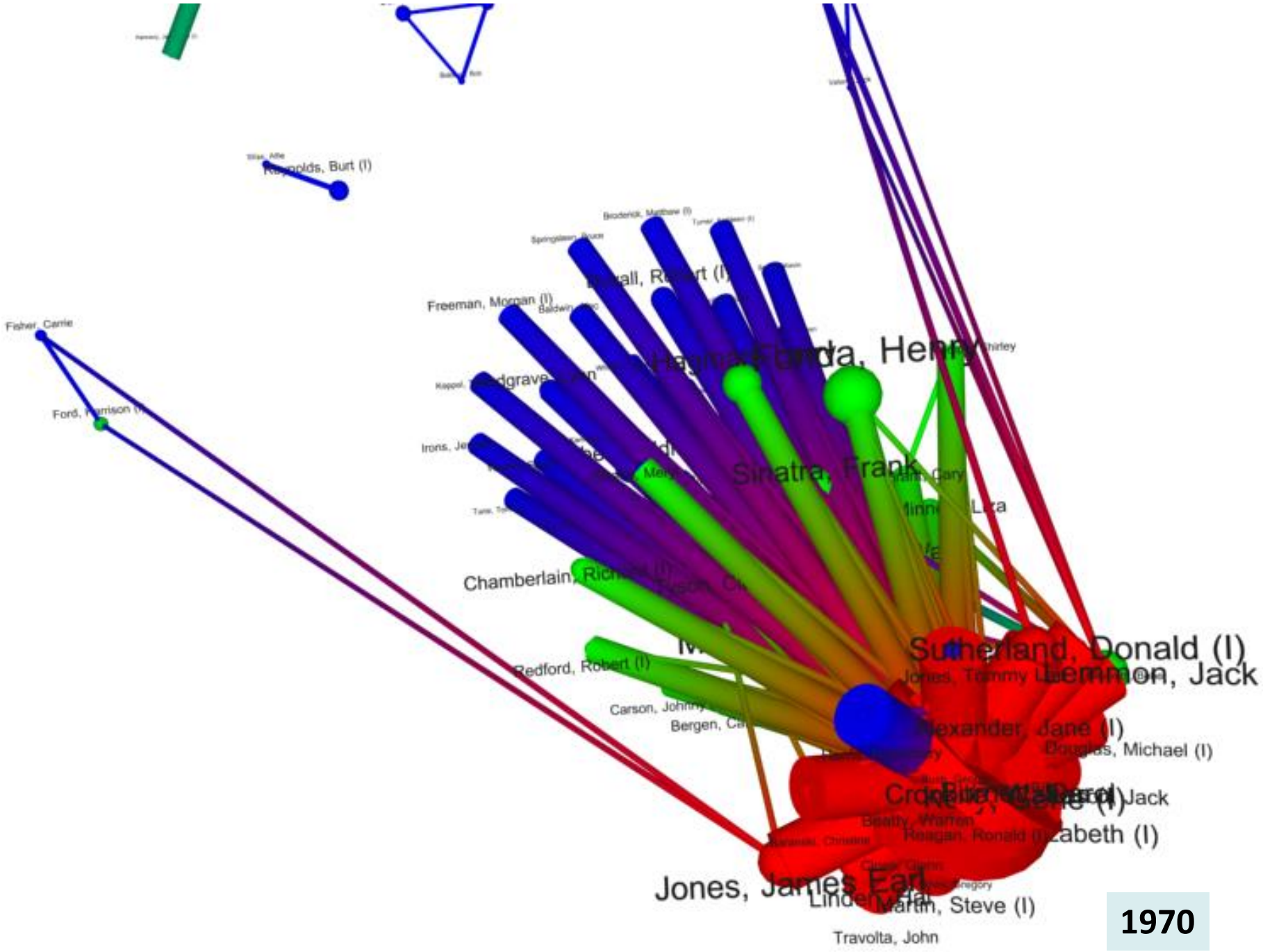
Visualisation and Animation

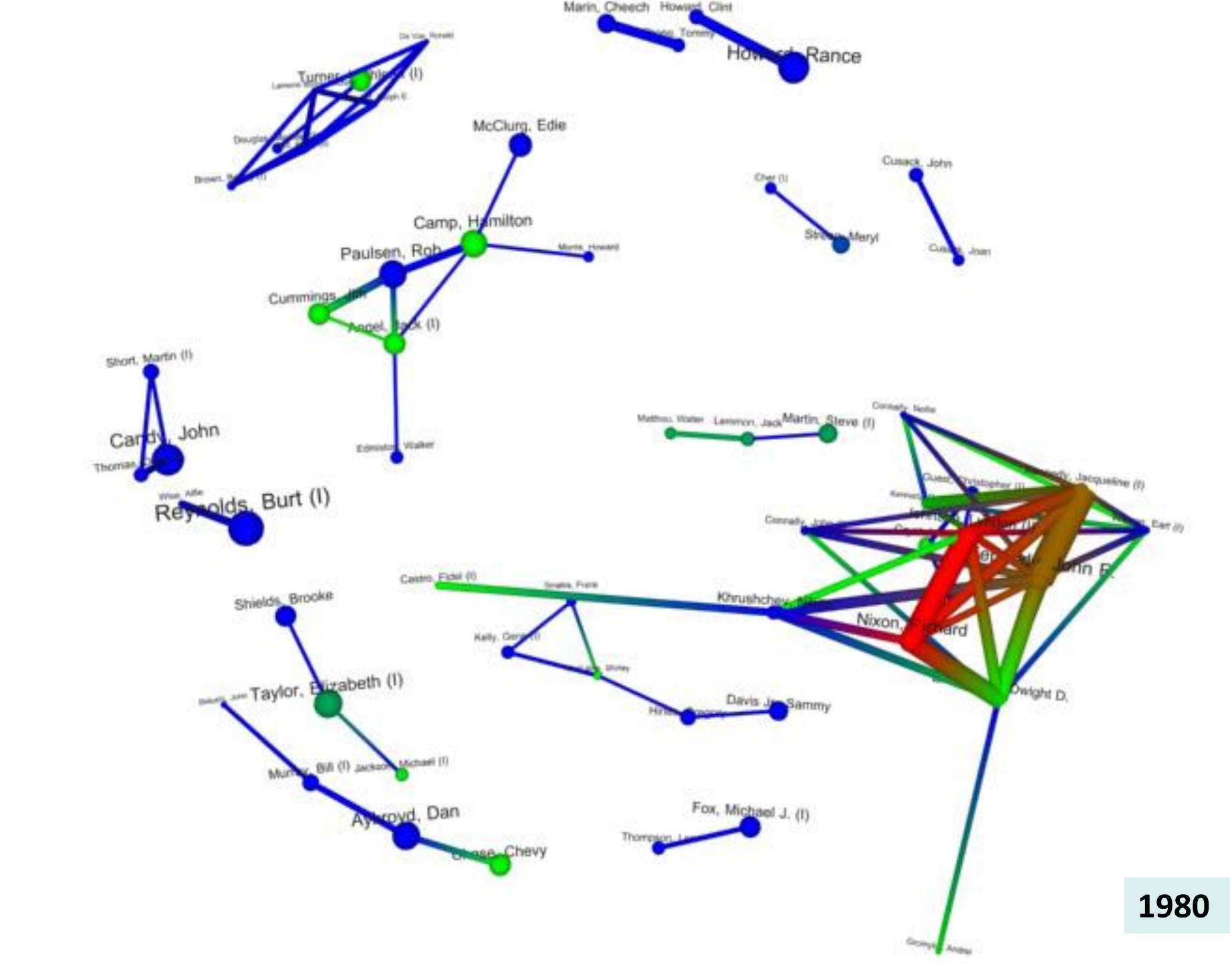
- **2.5D scale-free network layout** for each decade
  - Nodes represent actors (size = **degree centrality**)
  - Edges represent collaborations (**thickness** = # shared movies)
  - Nodes layered on 3 concentric spheres according to degree
    - Coloring: blue (innermost) – green – red (outermost: high degree)

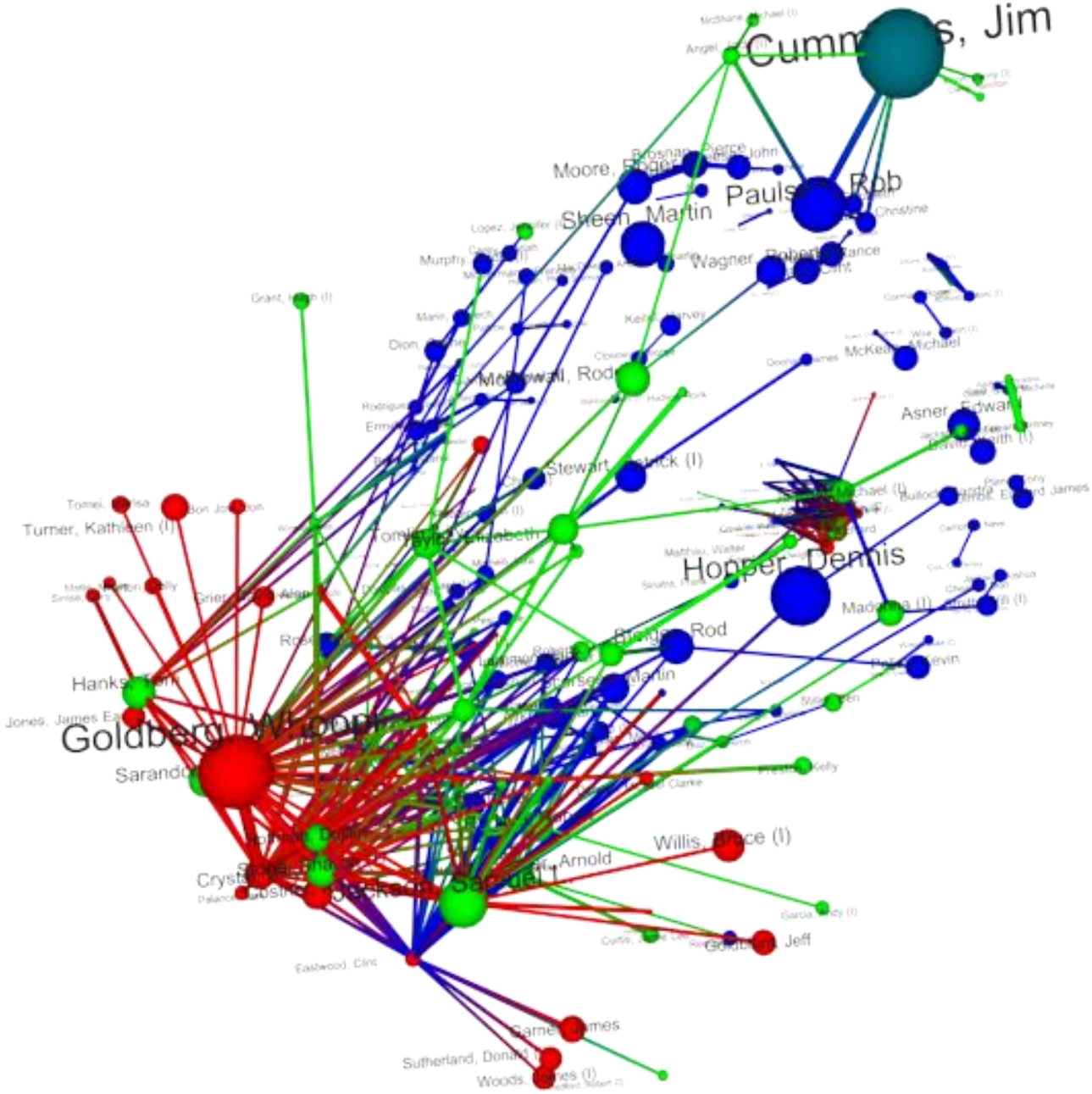
- **Smooth animation, between each decade, to preserve the mental map**
  1. Nodes and edges not continuing to the next decade ***fade out***
  2. Continuing nodes move to their new positions
  3. New nodes and edges ***fade in***


Howard, Rance
Connally, John (I)
Connally, Nellie
Kennedy, John F.
Kennedy, Jacqueline (I)
Greer, William (II)
Kellerman, Roy
Hill, Clint
Hargis, Bobby
Eisenhower, Dwight D.

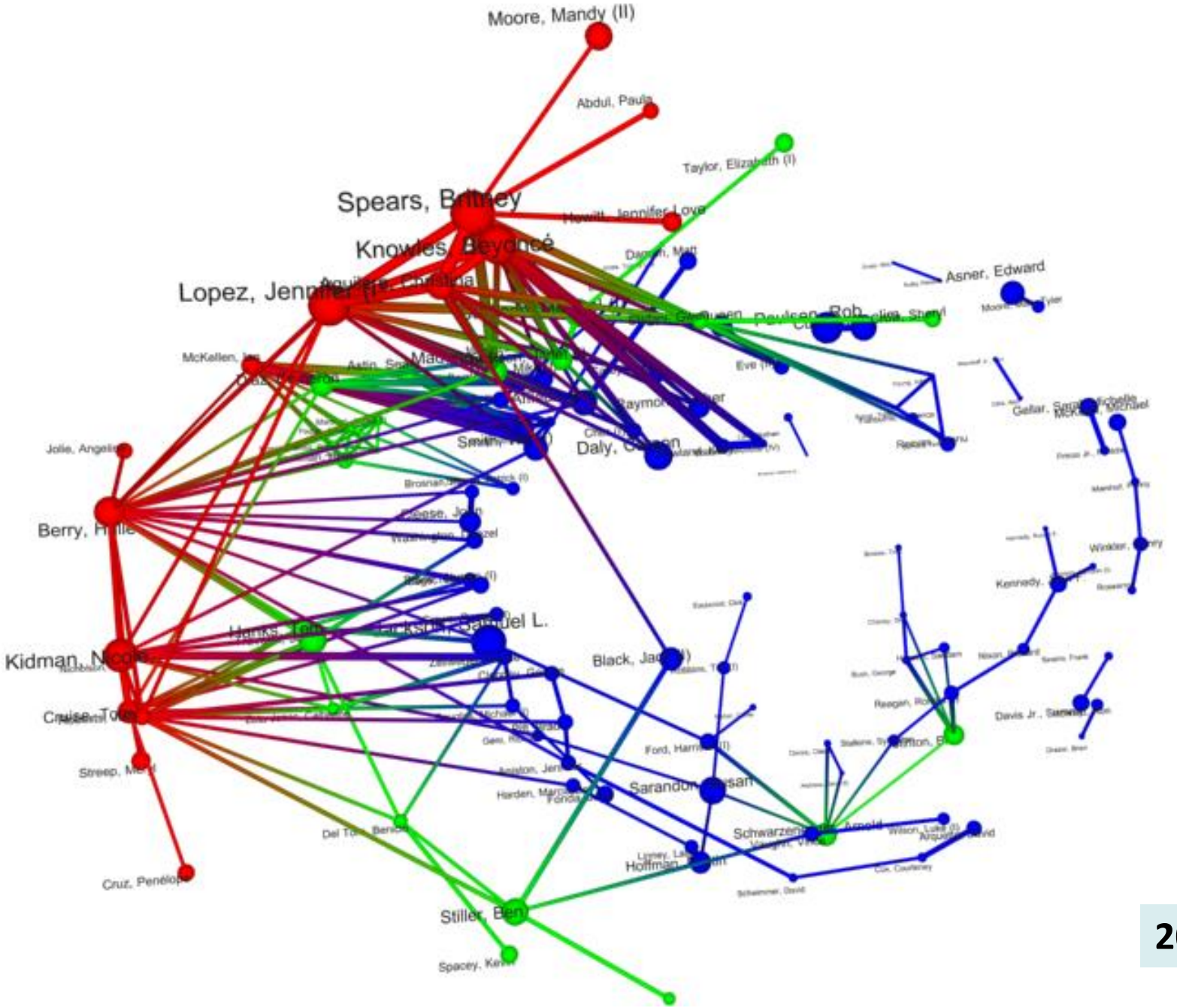
1960


Reynolds, Burt (I)
Freeman, Morgan (I)
Fisher, Carrie
Ford, Harrison (I)
Irons, Jeremy
Chamberlain, Richard (I)
Redford, Robert (I)
Carson, Johnny
Bergen, Candice
Sinatra, Frank
Sutherland, Donald (I)
Lemmon, Jack
Douglas, Michael (I)
Reagan, Ronald (I)
Jones, James Earl
Martin, Steve (I)
Travolta, John
1970


Marin, Cheech
Howard, Clint
Howard, Rance
McClurg, Edie
Camp, Hamilton
Paulsen, Rob
Cummings, Jim
Angel, Jack (I)
Edmiston, Walker
Cusack, John
Cusack, Joan
Short, Martin (I)
Candy, John
Reynolds, Burt (I)
Matthau, Walter
Lemmon, Jack
Martin, Steve (I)
Connelly, Nellie
Castro, Fidel (I)
Khrushchev
Nixon, Richard
Davis Jr., Sammy
Shields, Brooke
Taylor, Elizabeth (I)
Murray, Bill (I)
Aykroyd, Dan
Fox, Michael J. (I)
Dwight D.
1980


Jim
Moore,
Sheen, Martin
Wagner
Hopper, Dennis
Rod
Turner, Kathleen (I)
Willis,
Arnold
Jeff
Sutherland, Donald
Woods,

1990


Moore, Mandy (II)
Abdul, Paula
Taylor, Elizabeth (I)
Spears, Britney
Howitt, Jennifer Love
Knowles, Beyoncé
Lopez, Jennifer
Asner, Edward
Jolie, Angelina
Berry, Halle
Kidman, Nicole
Cruise, Tom
Streep, Meryl
Cruz, Penélope
Black, Jack (I)
Ford, Harrison (I)
Sarandon, Susan
Schwarzenegger, Arnold
Hoffman, Dustin
Stiller, Ben
Spacey, Kevin
Reagan, Ronald
Kennedy, John F.
Davis Jr., Sammy
Winkler, Henry
2000

Movie: GD 2005 contest

5. *Storytelling: narrative visualisation*

Storytelling: narrative visualisation

2. Clustered Actor Network

- Build a hypothesis
- Look for visual evidence to confirm/reject the hypothesis

“The Star System” Conjecture

- In the “golden era” of movie making, a hand-full of production companies dominated
- They instigated the “*star system*”…
  1. Pick an actor to be an up-and-coming “star”
  2. Make a succession of movies designed to further that star’s career
  3. The actor remains (or forced to remain) loyal to the production company
- *company-driven* movie making

- In modern times…
  1. A company develops a script for a movie
  2. They offer the script to the biggest “star” they can afford
  3. The actor picks which movies they want to act in for which companies
- *star-driven* movie making

VA Approach to "The Star System" Conjecture

Aim:

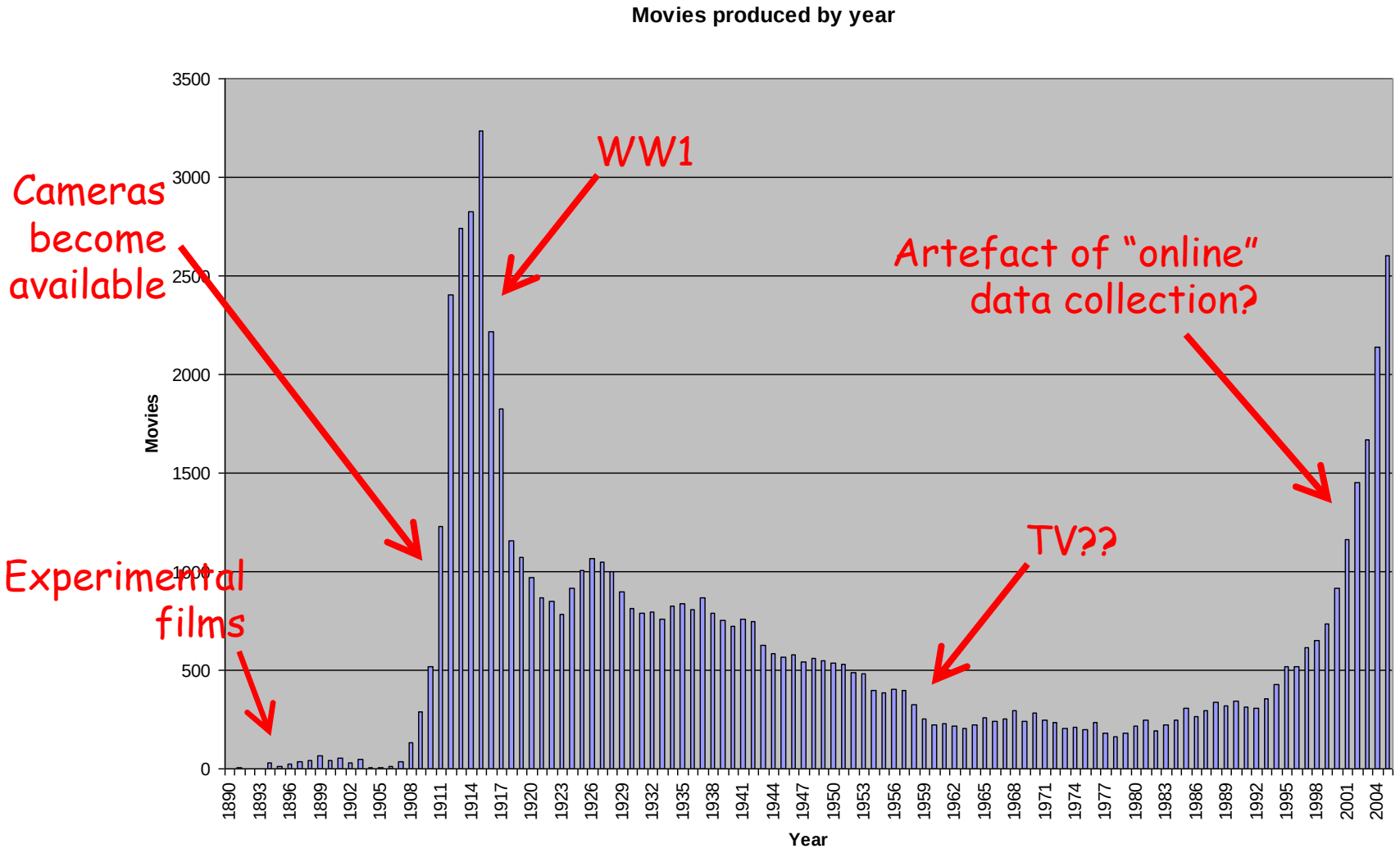
To use visual analytic approach to

- *characterise the differences between company-driven and star-driven movie making*
- *identify when the transition from one to the other occurred*
- *characterise what makes an actor a star*

Data set

- 1890-2005
- Only "movies" (no TV shows, straight-to-video, etc.)
- Only production companies inside the U.S.A.
- 69903 movies
- 277962 actors and actresses
- 14045 production companies

Data Analysis



Clustered Actor Network Visualisation

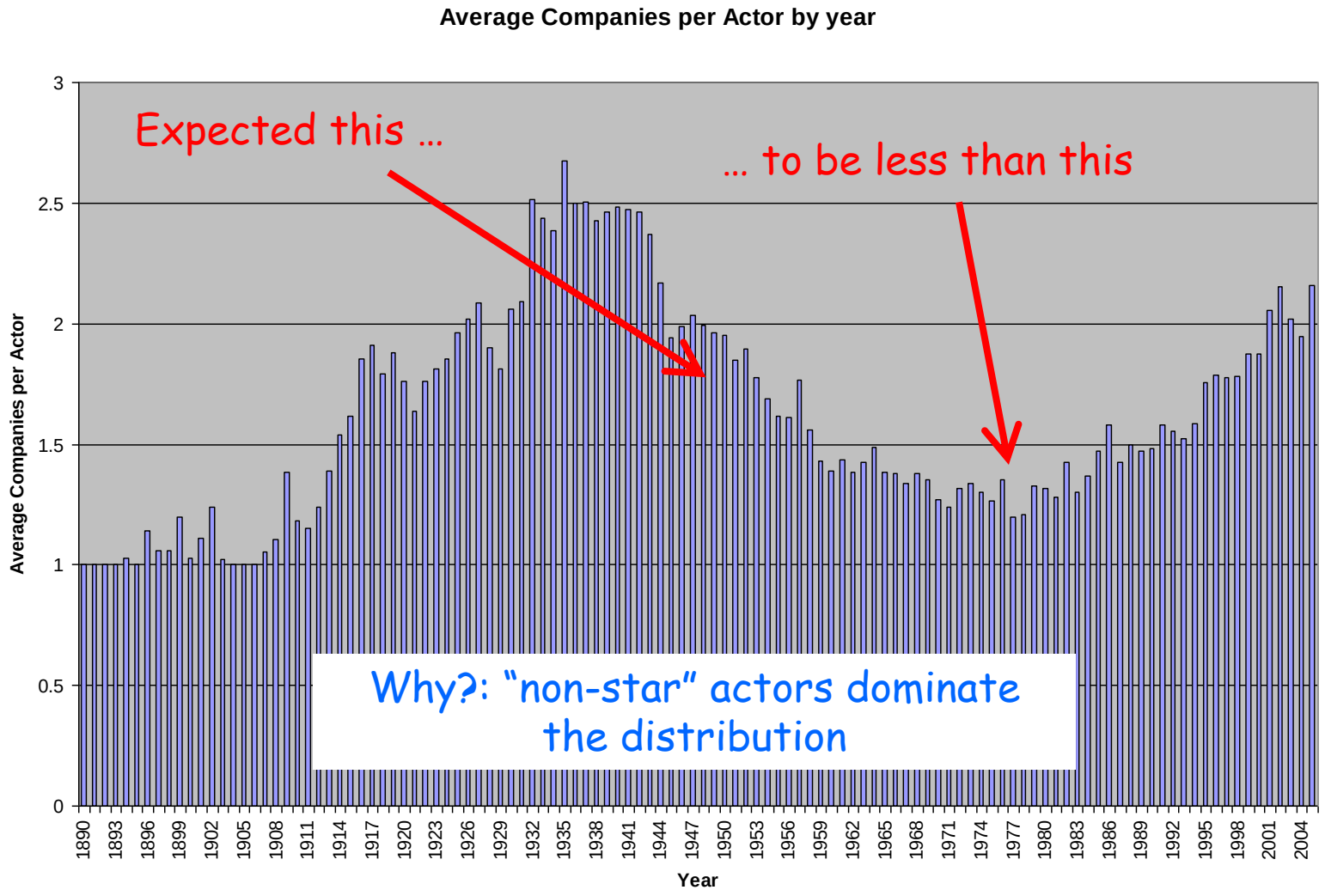
Use force-directed layout for clustered graphs

- Production company: cluster (circle)
- Actor working for one company: vertex in cluster
- Vertices (in different clusters) for same actor are joined by edges

Animation

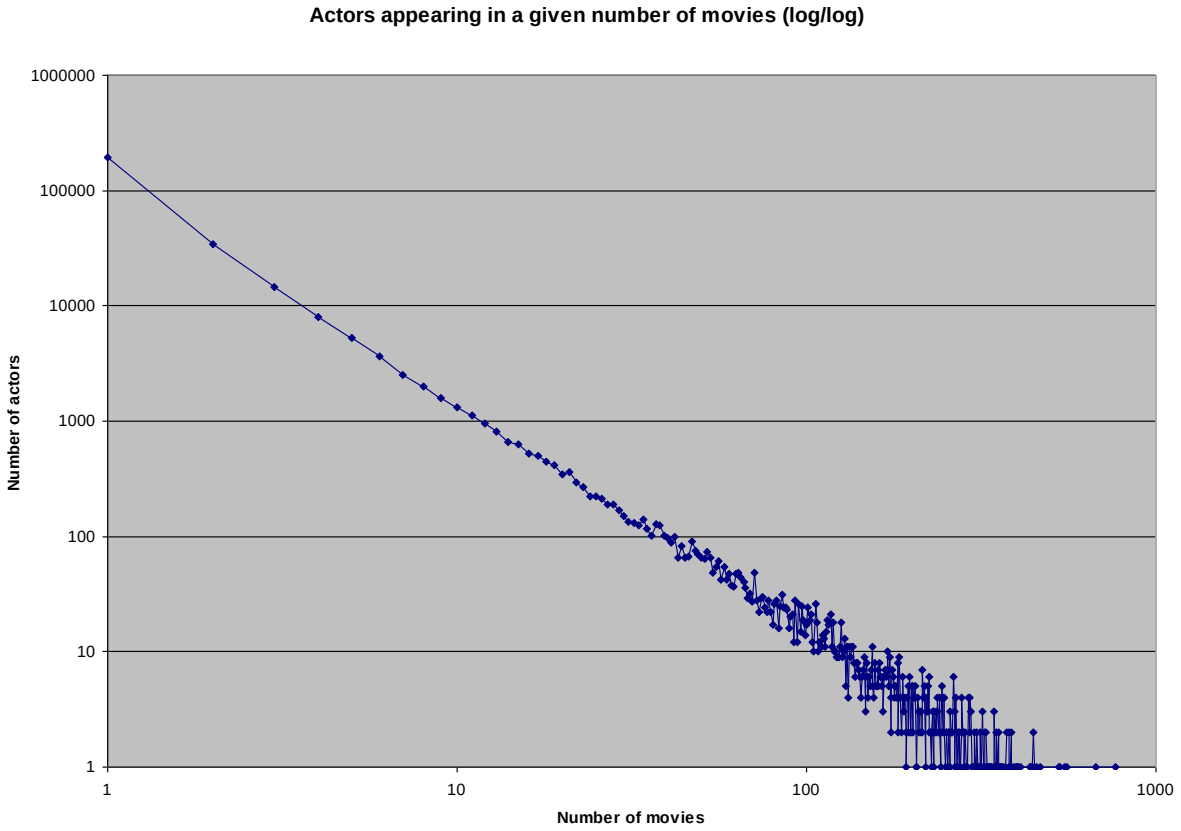
- Use animation to show movement of actors between companies
  1. New actors and companies fade in
  2. Old actors and companies fade out
  3. An actor simultaneously working for two companies is linked by a *static edge*
  4. An actor moving from one company to another is linked by an *animated edge with flow*
- ***Movie: early years***

Initial Analysis



Stars

- How do we distinguish a “star” from other actors?
- Many actors only appear in a few movies – probably not stars
  - Exceptions: James Dean only appeared in 6 movies (3 uncredited)
- The actor-movie network exhibits scale-free characteristics



Stars

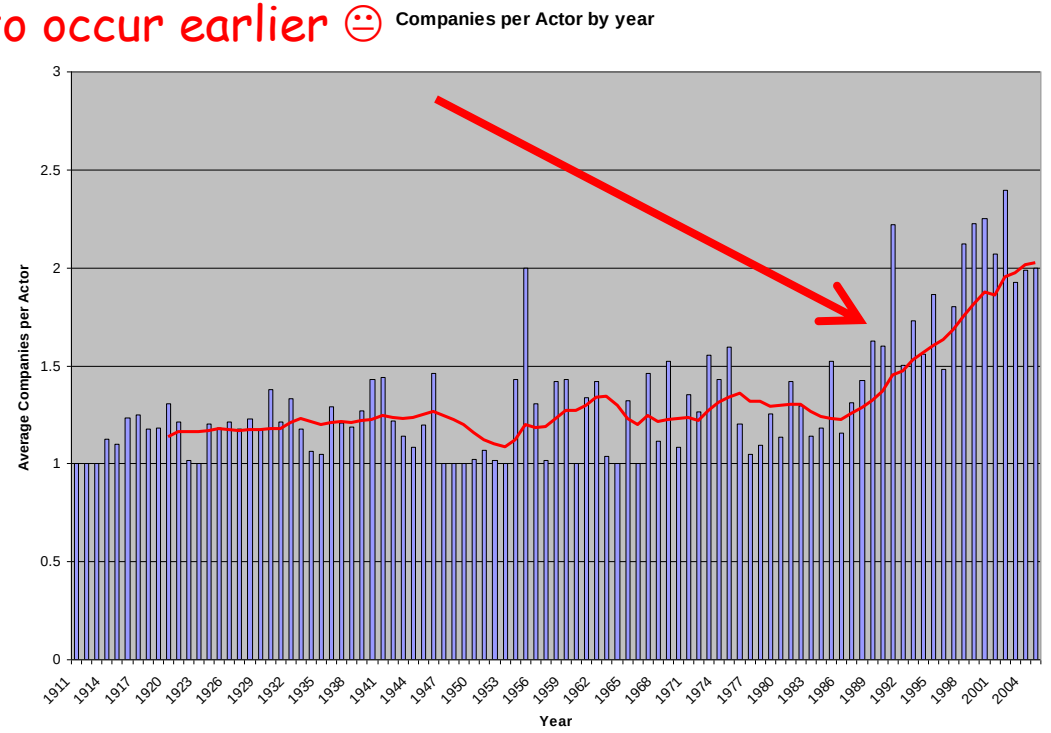
- Stars should appear in many movies, but in the most movies?
- Most prolific = Mel Blanc
  - The voice of Porky Pig, Daffy Duck and Bugs Bunny
- Second most prolific = Bess Flowers
  - Almost always minor roles
- ***Movies: Mel and Bess***

Importance

- Stars should appear in several <u>“important” movies</u>
  - Important ⇒ popular, commercial success, cult status or critical acclaim
- Filter out actors who do not appear in at least 3 movies rated in the top 20% of their year

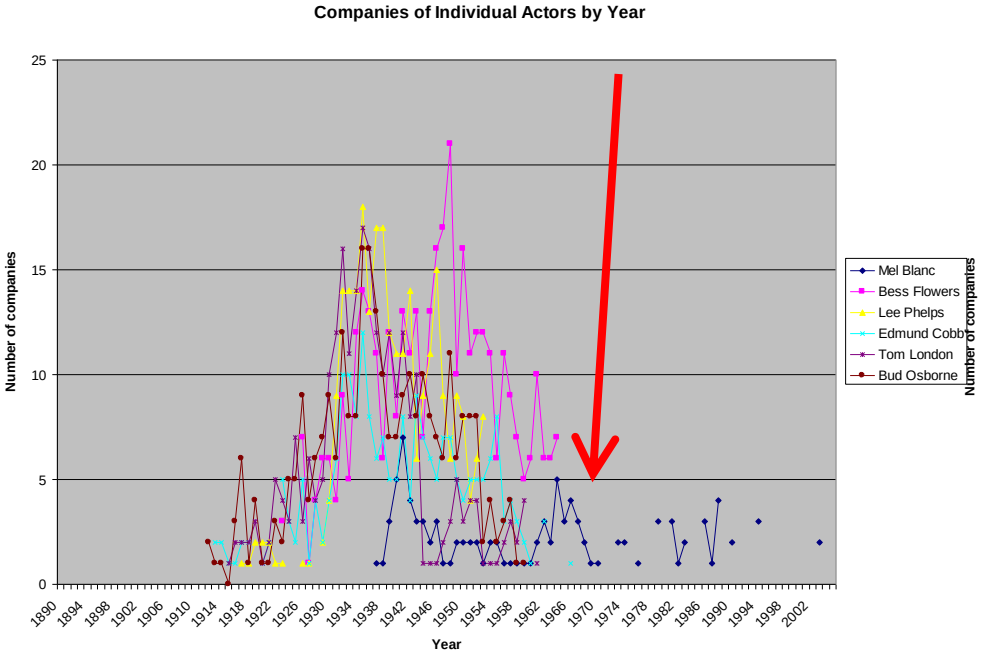
This is the expected trend ☺

but it was expected to occur earlier 😐

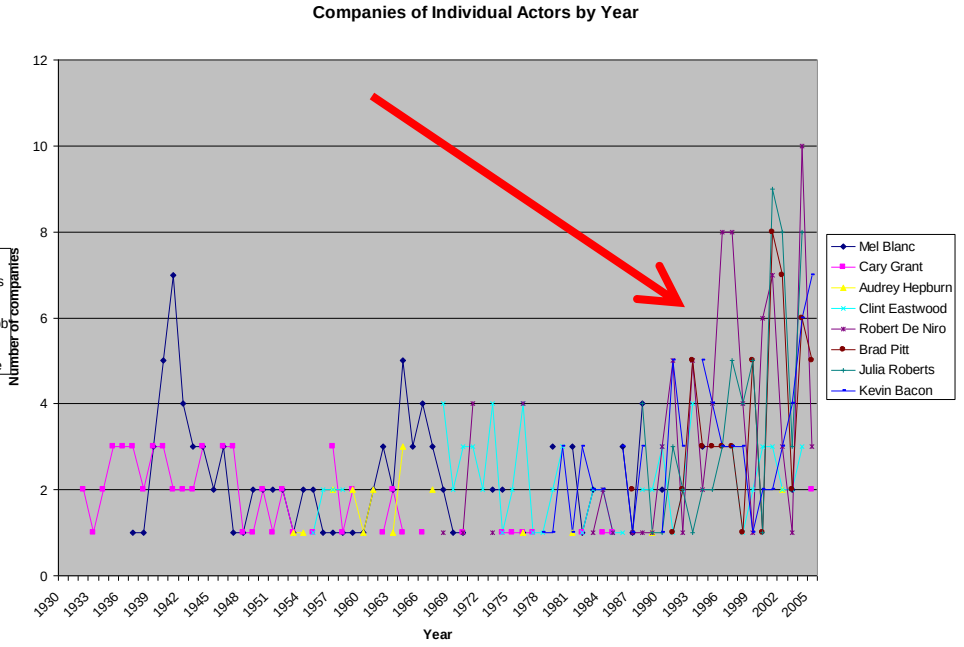


Comparison: Stars

Mel Blanc worked for significantly fewer companies than other prolific actors

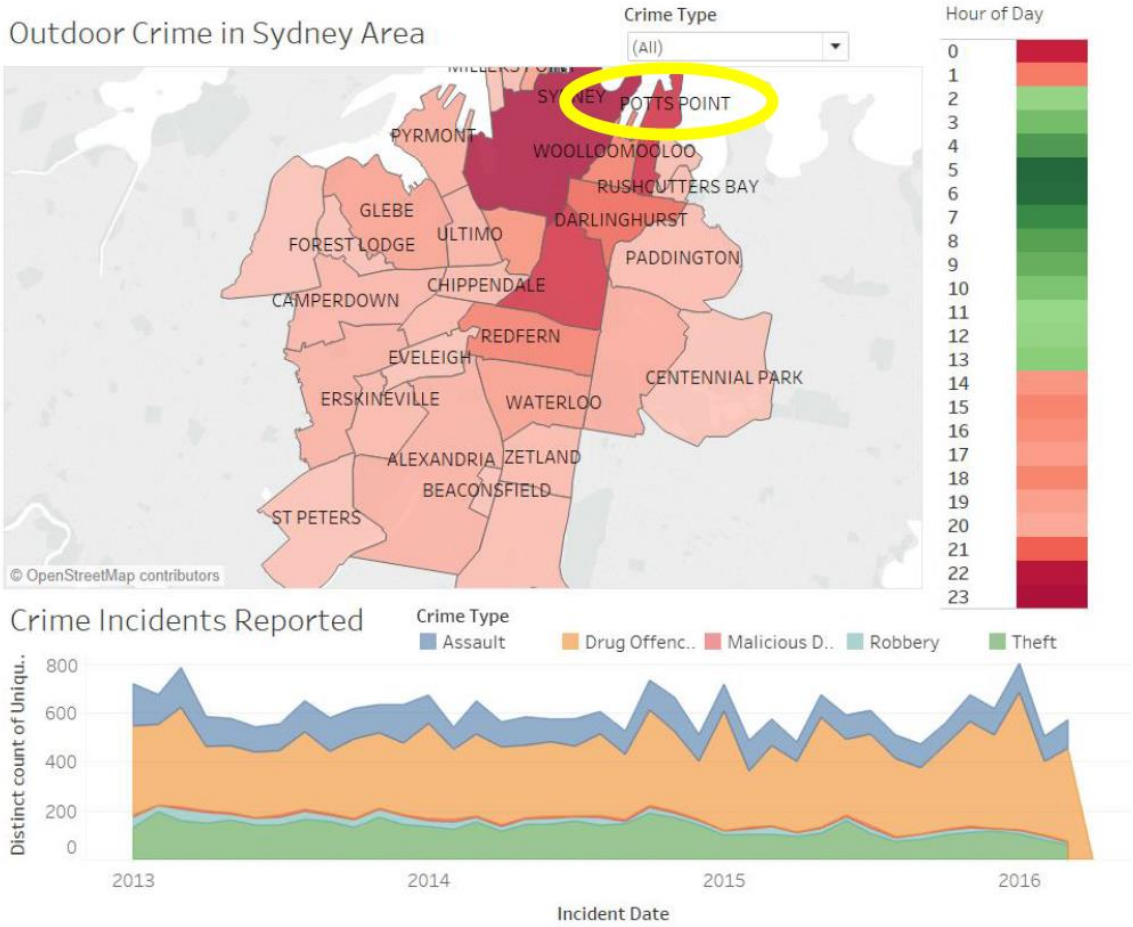


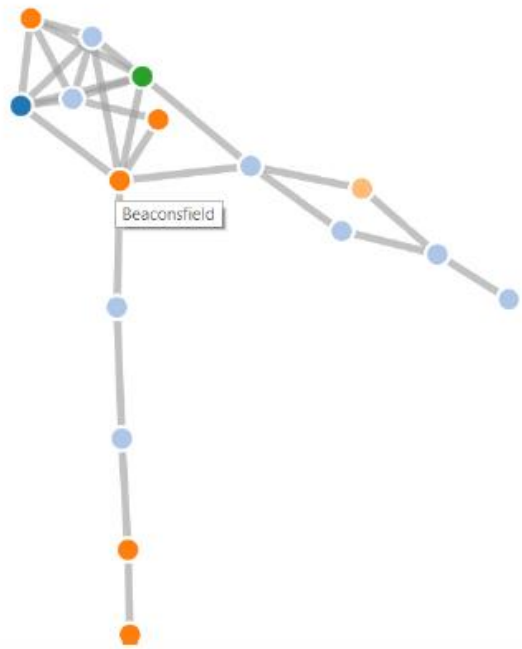
Modern stars work for more companies than older stars – but only significantly from 1990 onward



Examples of
Capstone Project 2017

Visual Analytics of Crime Data [Nardi]

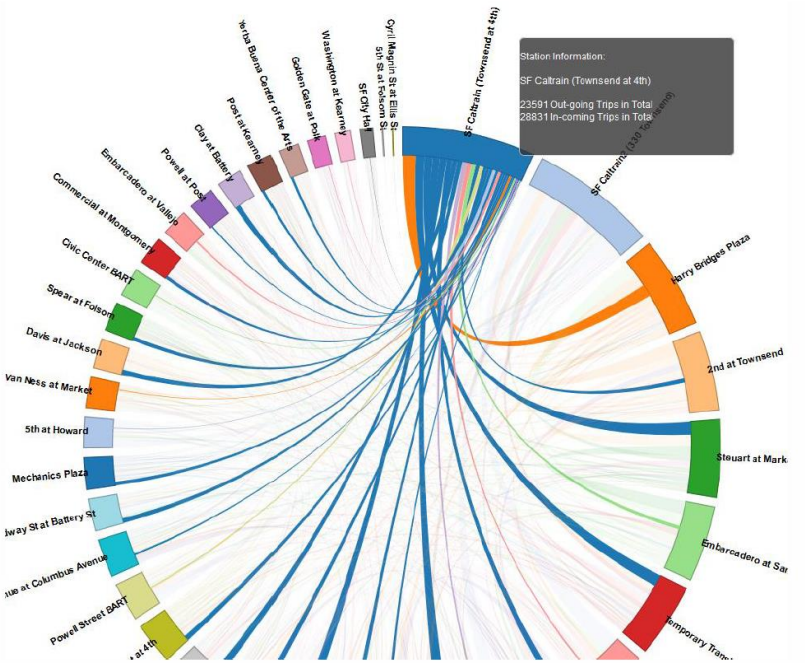




Tasks

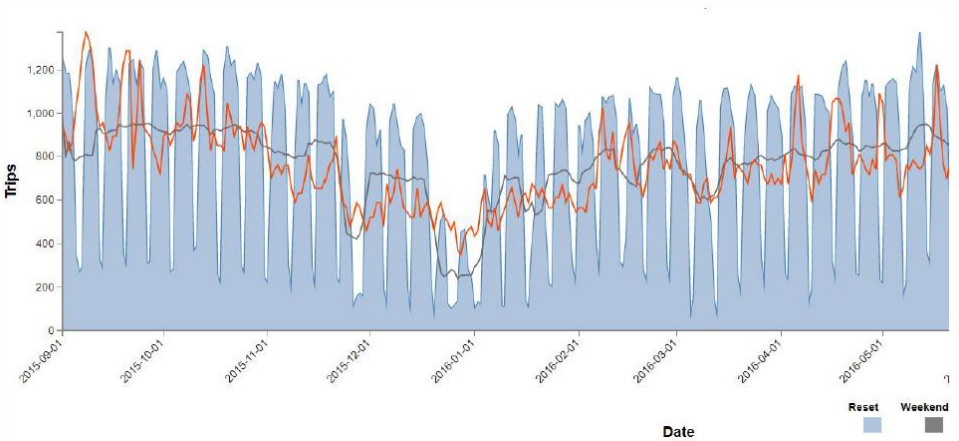
- Identify hot spots, time
- Crime trend analysis
- Similarity between suburbs

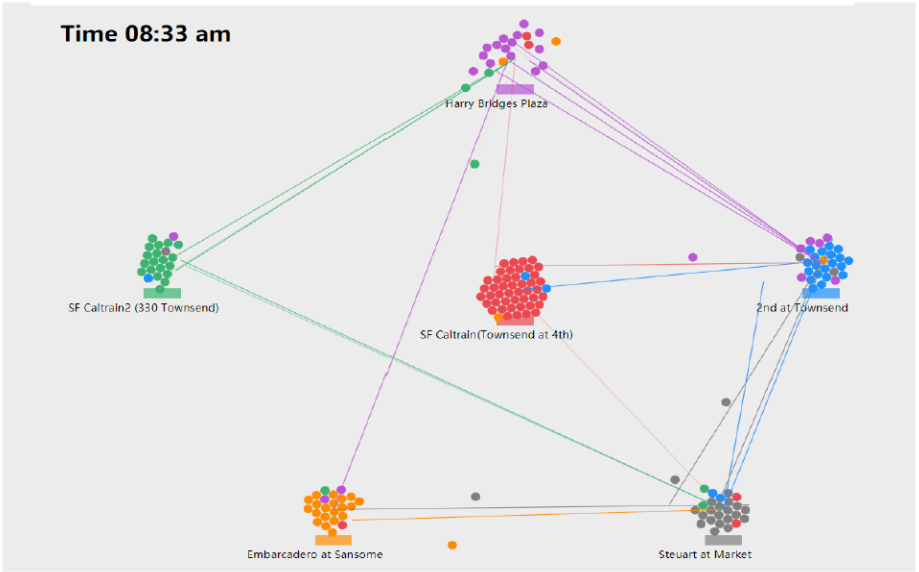
Visual Analytics of Bike Sharing Data [Chen]

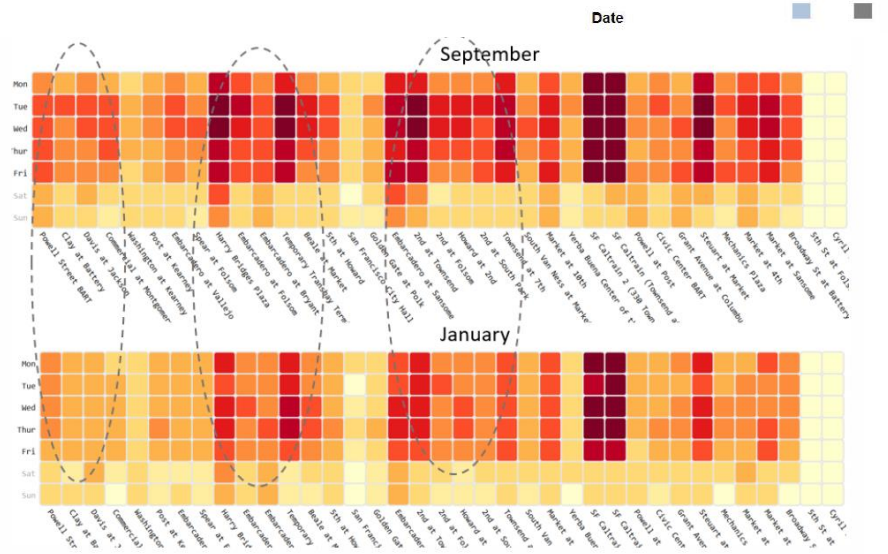


Tasks

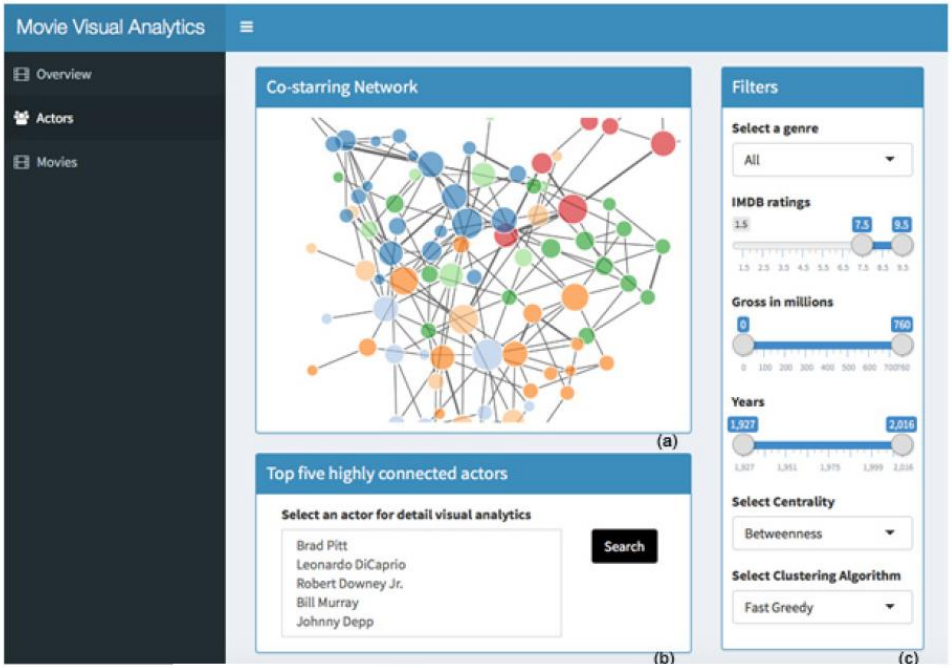
- Identify hot locations, time, season
- Routing network analysis
- Trajectory analysis

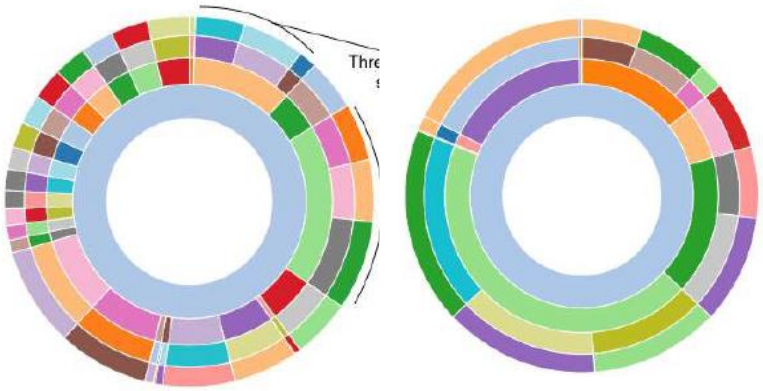


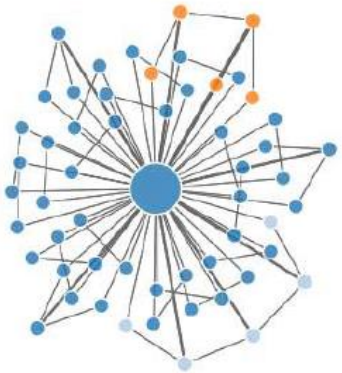


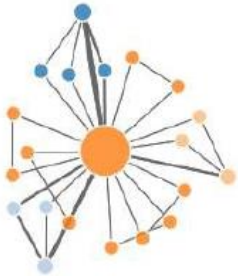


Visual Analytics of IMDB [Torkel]









Tasks
- Movie: temporal trends of genre
- Actor: comparison of collaboration network, genre, director

Assignment 2

1. Each Group: 6 people
- choose a group leader: record weekly meeting minute (week 6-12)
(Everyone should contribute)

2. Choose a data set
- define tasks
- data processing
- design: analysis/visualisation
- implementation: tools
- evaluation: tasks

3. Implementation
- you can use any existing tools (for analysis, visualisation)
- you can build a system using open source tools (D3, Gephi, GEOMI)

4. Write up reports
- initial report (5 mark): Week 7
- presentation (10 mark): Week 10 (week 10-12)
- final report (15 mark): Week 13

Produce good *visualisation* to support *analytic tasks* of a data:

1. VAST Challenge 2018: Mini Challenge 1
2. VAST Challenge 2018: Mini Challenge 2
3. VAST Challenge 2018: Mini Challenge 3
4. Olympic Athletes
5. World Bank Funding Decision
6. Mobile App Store Data
7. Kickstarter project
8. Global Terrorism database

More specifically:

1. *Design* (i.e., determine what *analysis* and *visualisation* to use)
2. *Implement* (i.e., determine what tools to use)
3. *Evaluate* with a given data sets and associated *analytic tasks*.

You can use whatever software you like or implement your system using existing tools, but you must acknowledge all your sources.

Marking Criteria:

1. Quality of your solution (esp. visualisation) to support tasks
2. Quality of Implementation
3. Quality of Report, Presentation/Demo

Initial report **should be in the following format *(min 5 pages)*:**

1. Introduction
1.1 Data set
1.2 Tasks

2. Design and Approaches
2.1 Analysis
2.2 Visualisation

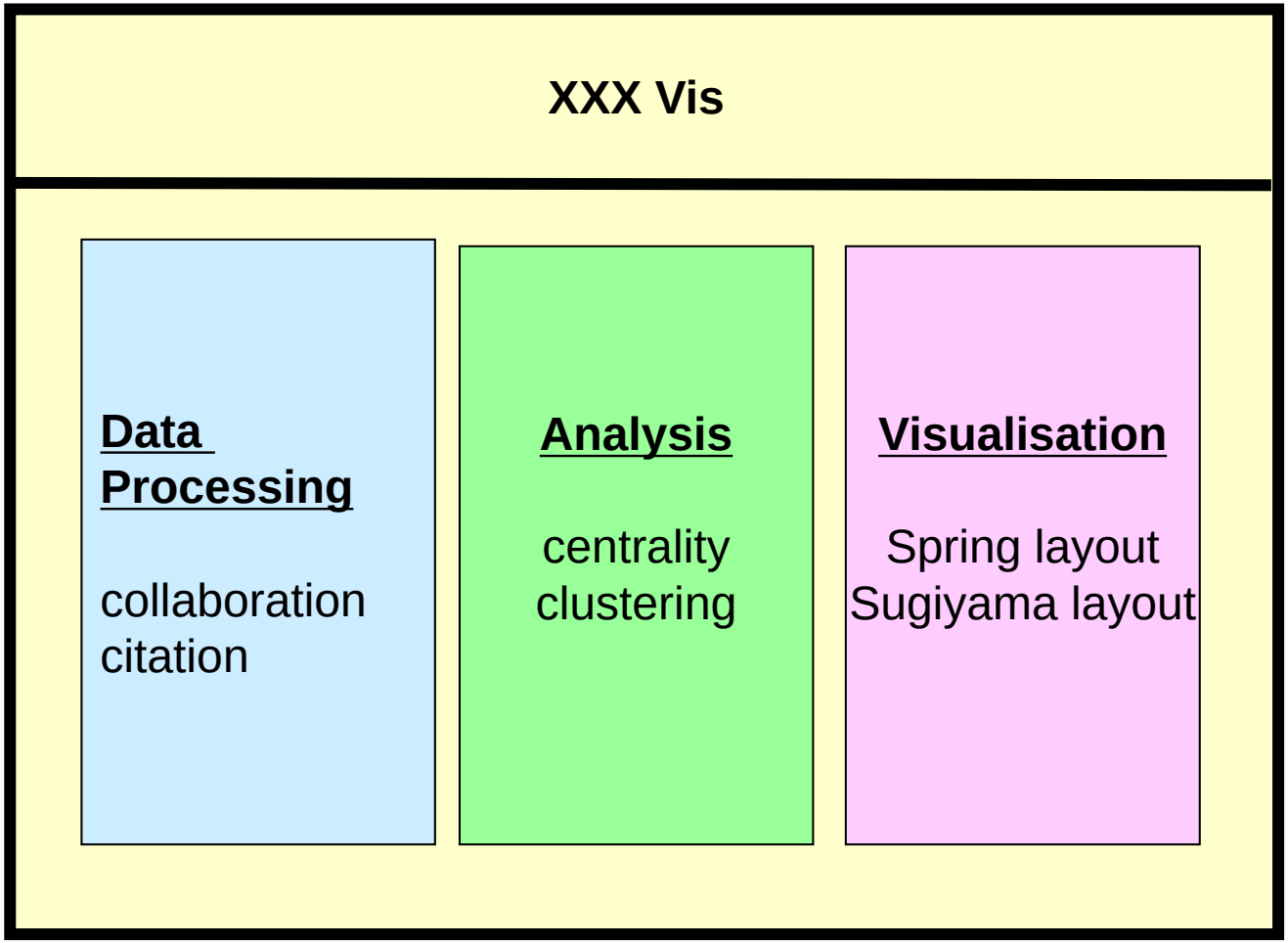
3. Implementation

4. Evaluation

5. Planning (weekly, role assignment)

6. References

Framework of VA System



Presentation **should be in the following format *(max 10 slides)*: 10 min**

1. Introduction
1.1 Data set
1.2 Tasks

2. Design and Approaches
2.1 Analysis
2.2 Visualisation

3. Implementation

4. Evaluation

5. Progress

6. Planning

Final report **should be in the following format *(10-15 pages)*:**

1. Introduction
1.1 Data sets and Tasks
1.2 Aims and Contribution

2. Design
2.1 Analysis
2.2 Visualisation

3. Implementation

4. Evaluation
4.1 Results
4.2 Discussion

5. Conclusion
6. References

7. Appendix: *Group meeting minutes (0.5-1 page per week: week 6-12)*
8. Appendix: Code

Questions?