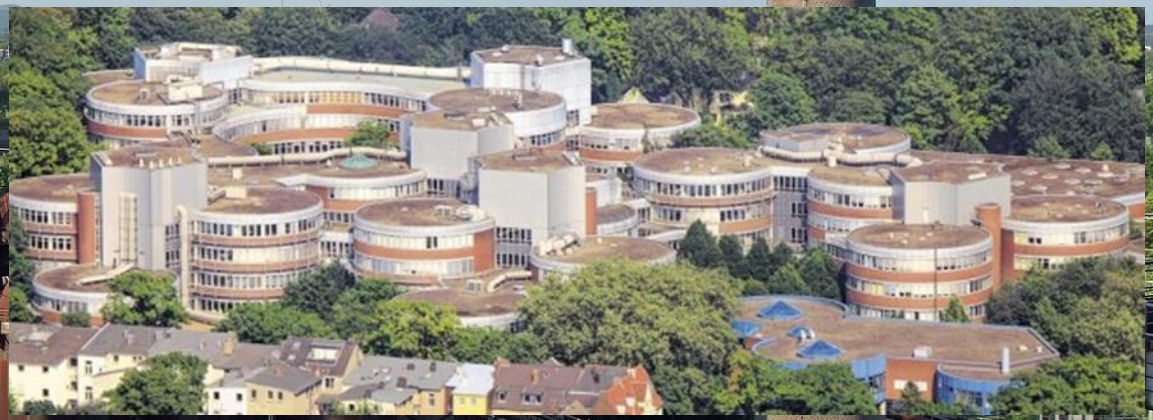
Network Analysis of Online Communities - Applications and Tools

H. Ulrich Hoppe

Universität Duisburg-Essen / COLLIDE

London, April 2015

University of Duisburg-Essen
created 2003 by a merger
40,000 students



Our Department

… of Computer Science and Applied Cognitive Science in the Engineering Faculty (Duisburg Campus)

17 professorships / groups (13 in CompSci)

Focus on Interactive Media and "Human-oriented Computing"

3 B.Sc./M.Sc. Study Programmes with approx. 1500 students

COLLIDE Research Group

"Collaborative Learning in Intelligent Distributed Environments" founded in 1995

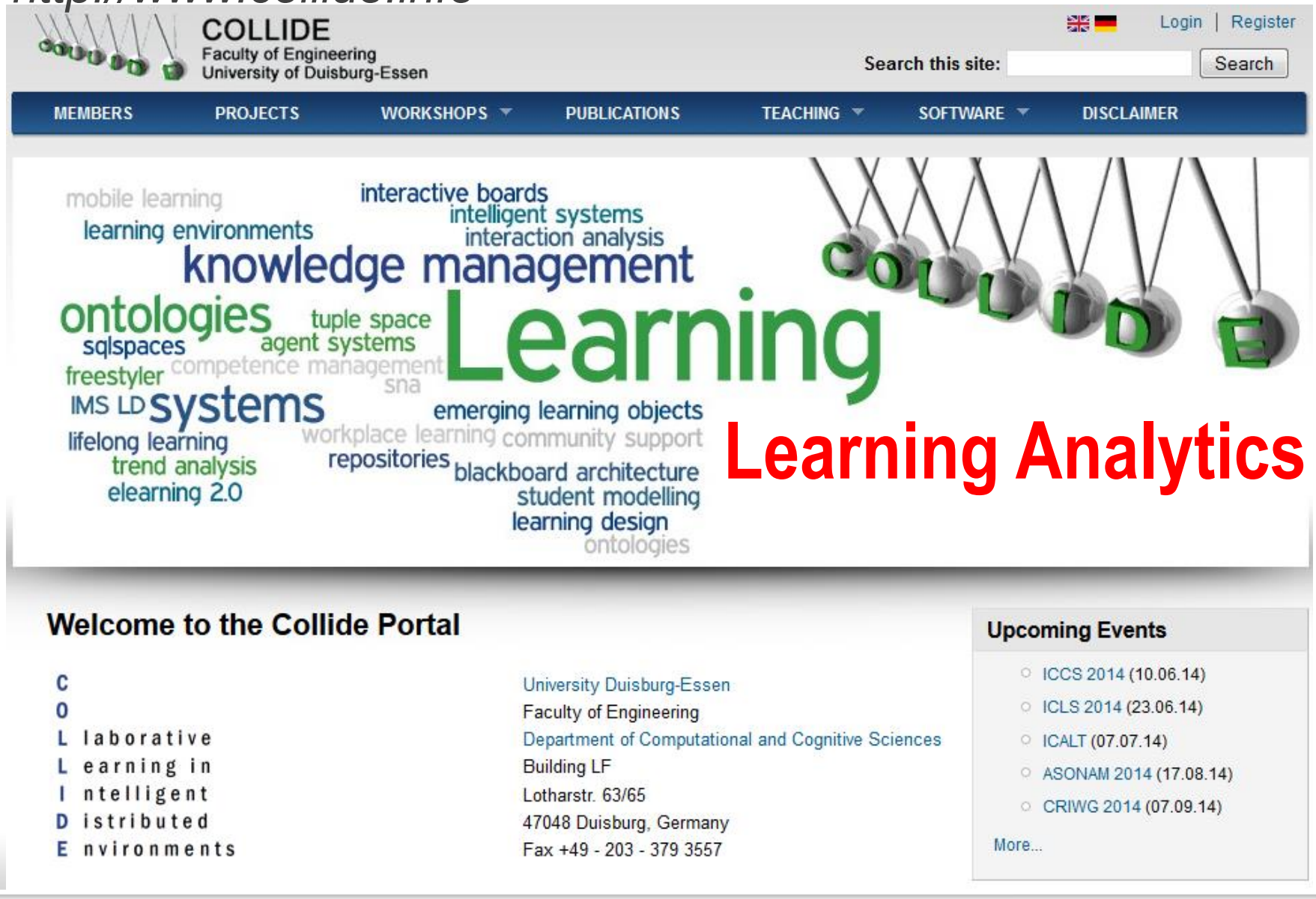
Applied Computer Science perspective

Lines of research:

- *Distributed computing for CSCL and knowledge building communities (systems/architectures)*
- *"Learning Analytics"*
- *Social network analysis and community support*

COLLIDE

http://www.collide.info



Case Study 1:
„Productive Online Communities“

(with Sam Zeini, Tilman Göhnert)

„Productive Online Communities“

… as opposed to

- communities of interest
  (professional networks, photo communities,
  peer help communities, …)

- “socialising communities”

Open Source Projects as communities of practice with open data sources

- Contributors/students engaged in OSPs form communities of practice
- Most data (log files from SVN/CVS repositories as well as mailing lists) freely available
- Different roles and skills of actors have to be considered (hierarchies, commit rights)
- Data has to be cleaned (e.g. merging name duplicates)

Example: OpenSimulator (OpenSim)

OpenSimulator:

- Open source server platform for hosting virtual worlds
- Compatible with the client for Second Life
- Can host alternative worlds

OpenSimulator

used within KoPIWA research project (German project on Open Innovation and competence management in IT)

Source Data

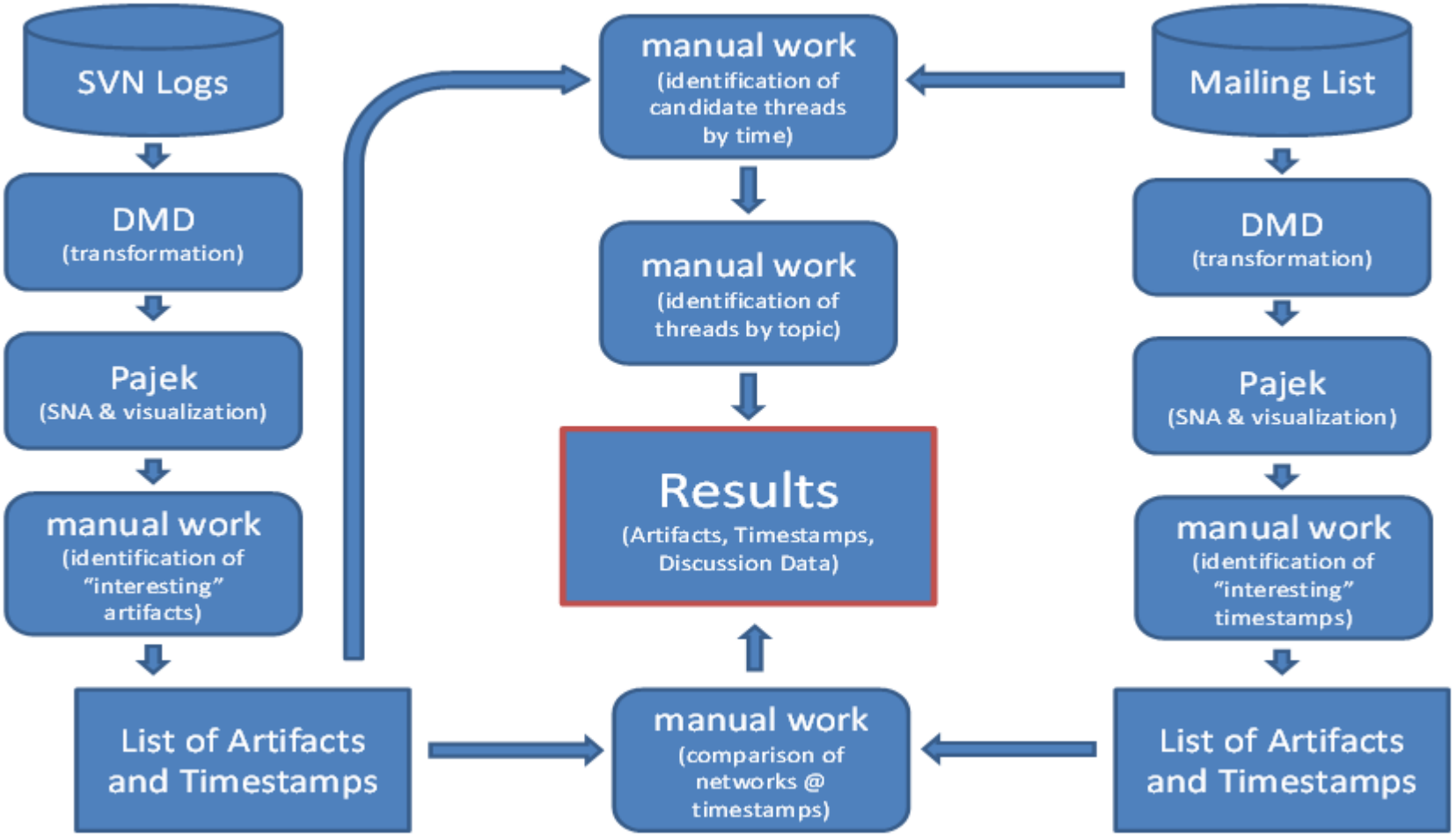
OpenSim Developer mailing list and SVN source code repository between Sept. 2007 and Feb. 2009

- Cleaned by merging duplicate names in mailing lists
- Mapping real names from mailing lists to SVN nicknames
- SVN contains 26 users and 6011 objects (classes) based on 32867 objects including revisions
- Developer mailing list contains 197 users and 1184 topics based on 5505 emails.

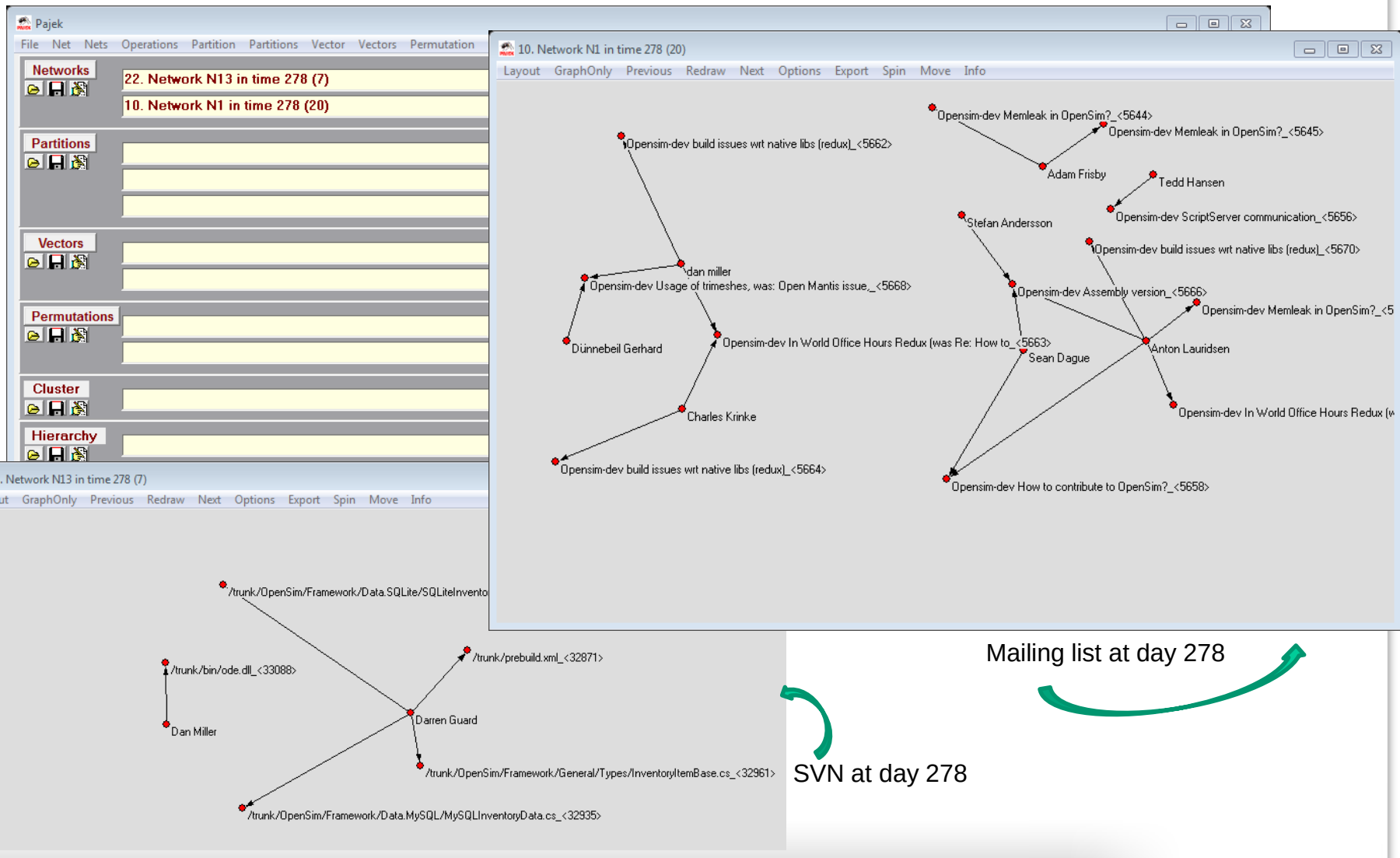
Dataset from the SVN repository + discussion board of the student subproject from Oct. 2009 to March 2010

- SVN data from the students subproject contains 4 users and 235 objects based on 326 revisions.
- Discussion board hosted in a liferay based portal contain 12 students discuss 78 threads based on 243 messages (in German)

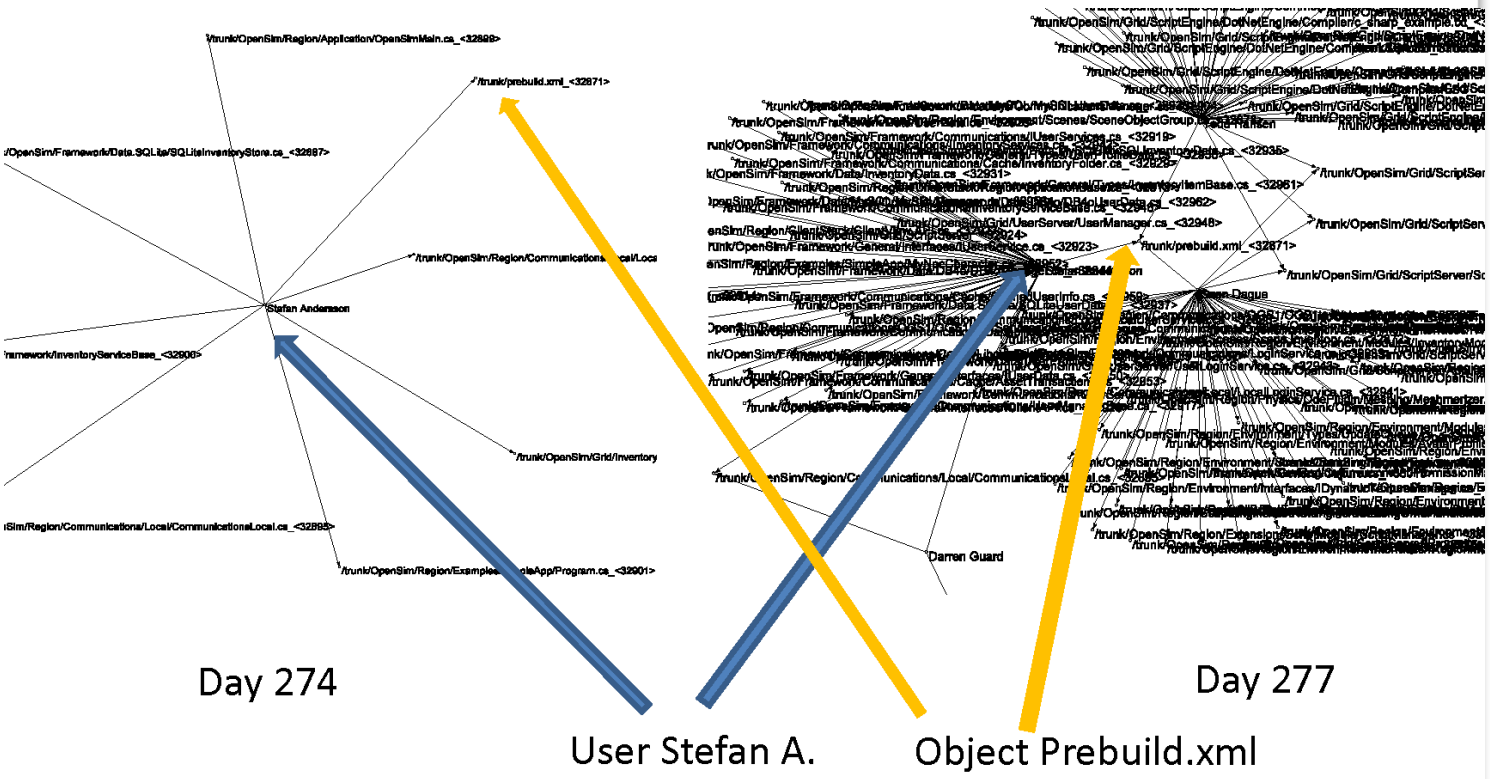
Semi-automatic Analysis Workflow



Analysis using Pajek



Example Observation



- Burst of activity in SVN
- Restructuring of the network (bridging between components)

Comparison of Centrality Dynamics

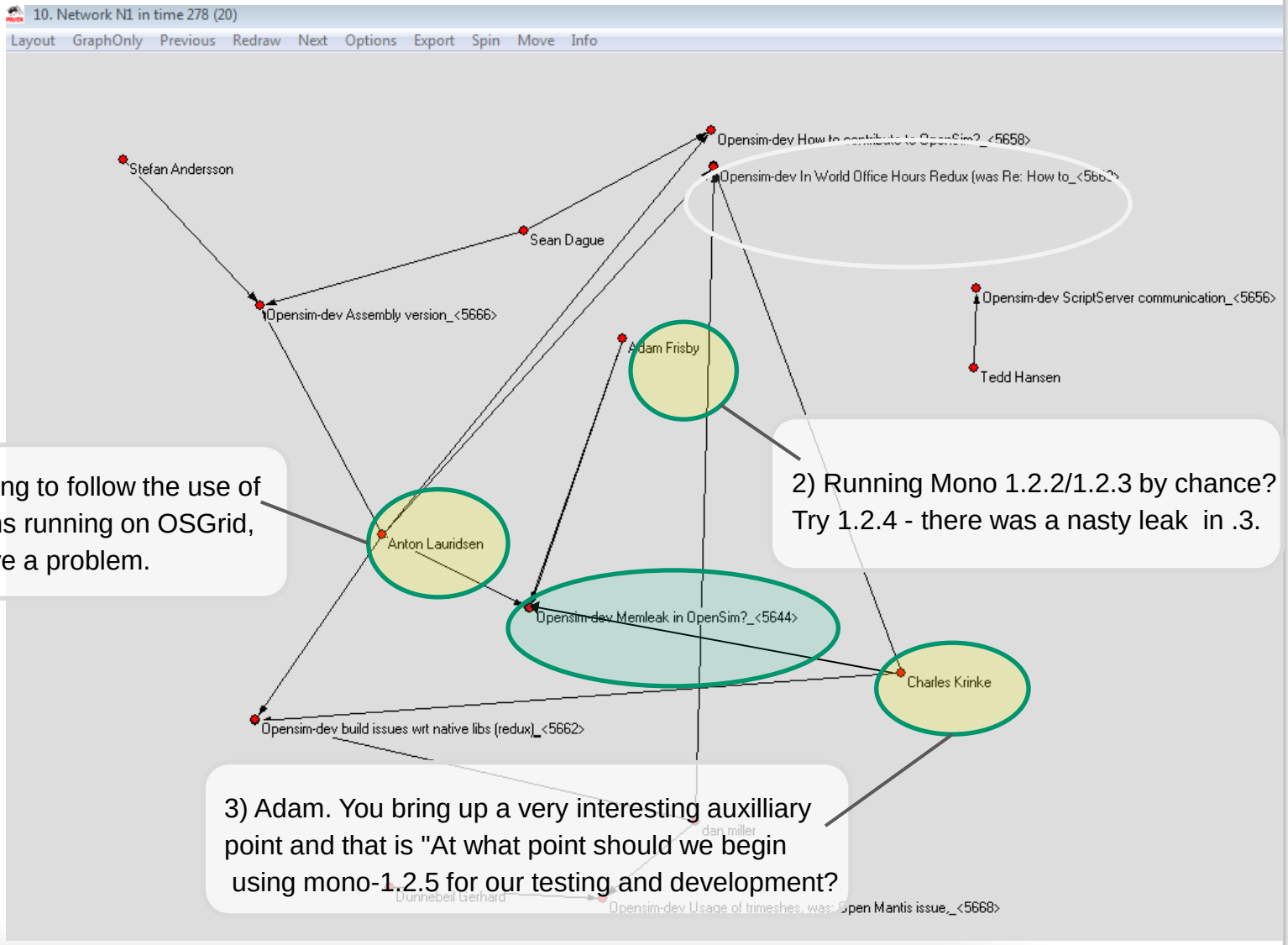
Name	Delta T1/T2	Delta T2/T3	Delta T3/T4	Delta T1/T4
Adam F.	0	0,125	-0,125	0
Aldon H.	0	0,083333	-0,083333	0
Anton L.	0,1875	0,104167	0,333333	0,625
Brian M.	-0,125	0	0	-0,125
Charles K.	0,0625	0,020833	-0,083333	0
Chris D.	0	0	0	0
Dalien T.	0,1875	-0,354167	-0,020833	-0,1875
Dan M.	0,4375	-0,291667	0,166667	0,3125
Darok K.	0	0	0	0
Gerhard D.	0	0,083333	-0,083333	0
Izumi N.	0	0	0	0
Jeff A.	0,125	0	-0,125	0
Mic B.	0	0,125	-0,125	0
Michael W.	0,125	-0,145833	0,395833	0,375
Sean D.	0,375	-0,416667	0,229167	0,1875
Stefan A.	0,625	-0,479167	0,291667	0,4375
Tedd H.	0,125	-0,1875	0,1875	0,125
	Rank 4	Rank 3	Rank 2	Rank 1

T= Time slices with length of 10 days

T1= Day 250-259
T2= Day 260-269
T3= Day 270-279
T3= Day 280-289

- Interesting time point according to interaction increase
- Actor “Tleiades” (Anton L.) identified in communication network at same time point by his increase of centrality (shooting star)

Communication (mailing list)



Example analysis (OpenSim)

- First indicator: high activity followed by network restructuring
- Second indicator: degree gradient points to specific actor “Tleiades” and his entry to the core of the network
  (in the end, degree centrality moved to 0.0505 in the developer mailing list – rank 47 out of 199)
- Semantic background: T. initiated relevant discussions on XML-RPC as well as on group standards for build system

Abstraction:
“Centrality gradient pattern”

Input:
a time sequence of 1-mode networks (e.g. of actors)

Workflow:

- *calculate centrality measure (degree, betweenness, …) for all nodes in all time slices*
- *calculate differences*
- *find maximum differences*

Indicator for …

role changes of actors (periphery -> centre)

Cohesive Subgroups ("community detection")

Available basic methods:

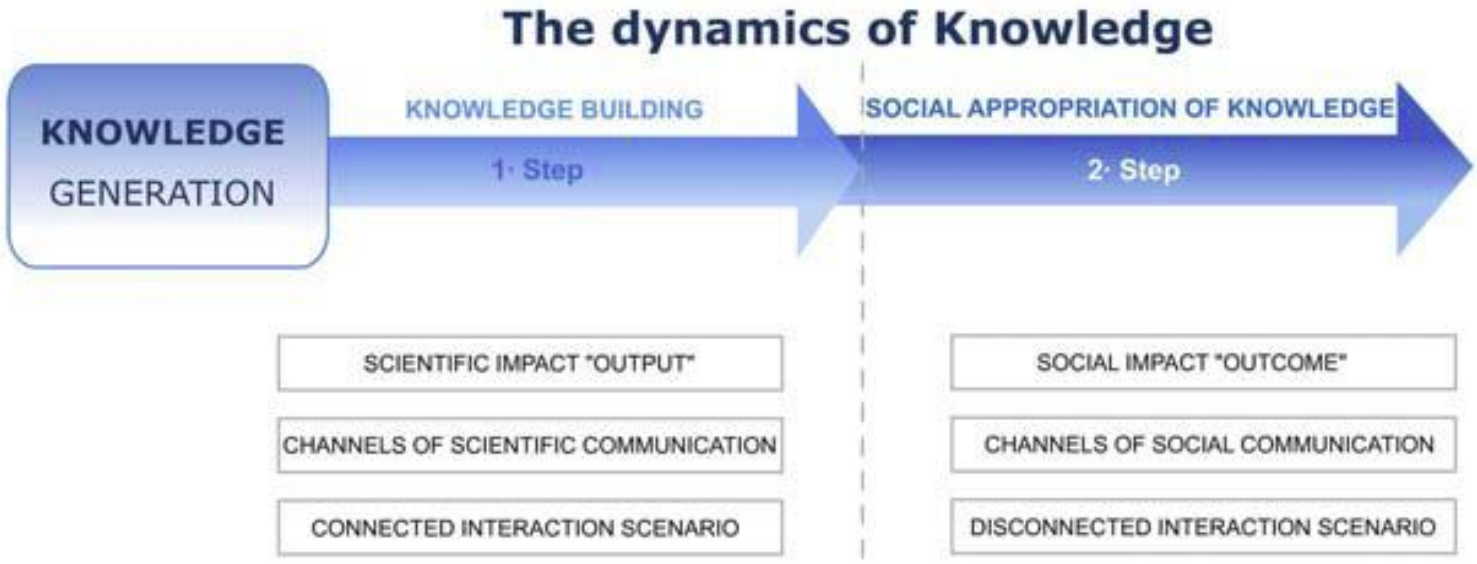
- k-cores
  (basic method to "thin out" a network,
  loss of connectivity => subgroups)
- Modularity-based algorithms
- Clique percolation method
  (allows for detecting overlapping subgroups)

Often desired: *detection of overlaps!*

The SISOB Analytics Workbench

SISOB Project (2011-13)

- Goal: Develop tools for measuring and predicting the social impact of science (beyond bibliometrics)



- Basic technology: computer supported *Social Network Analysis (SNA)*

SISOB Partners

- Universidad de Malaga (UMA)
- Innovation, Science and Enterprise Regional Ministry of Andalucía (CICE)
- Universität Duisburg-Essen (UDE)
- Institute for Research Organization, Hungarian Academy of Sciences (IRO-HAS)
- Frontiers Research Foundation / Scientific Editors (Frontiers)
- Fondazione Rosselli (FR)
- Red de Indicadores de Ciencia y Tecnología (RICYT)

Motivation

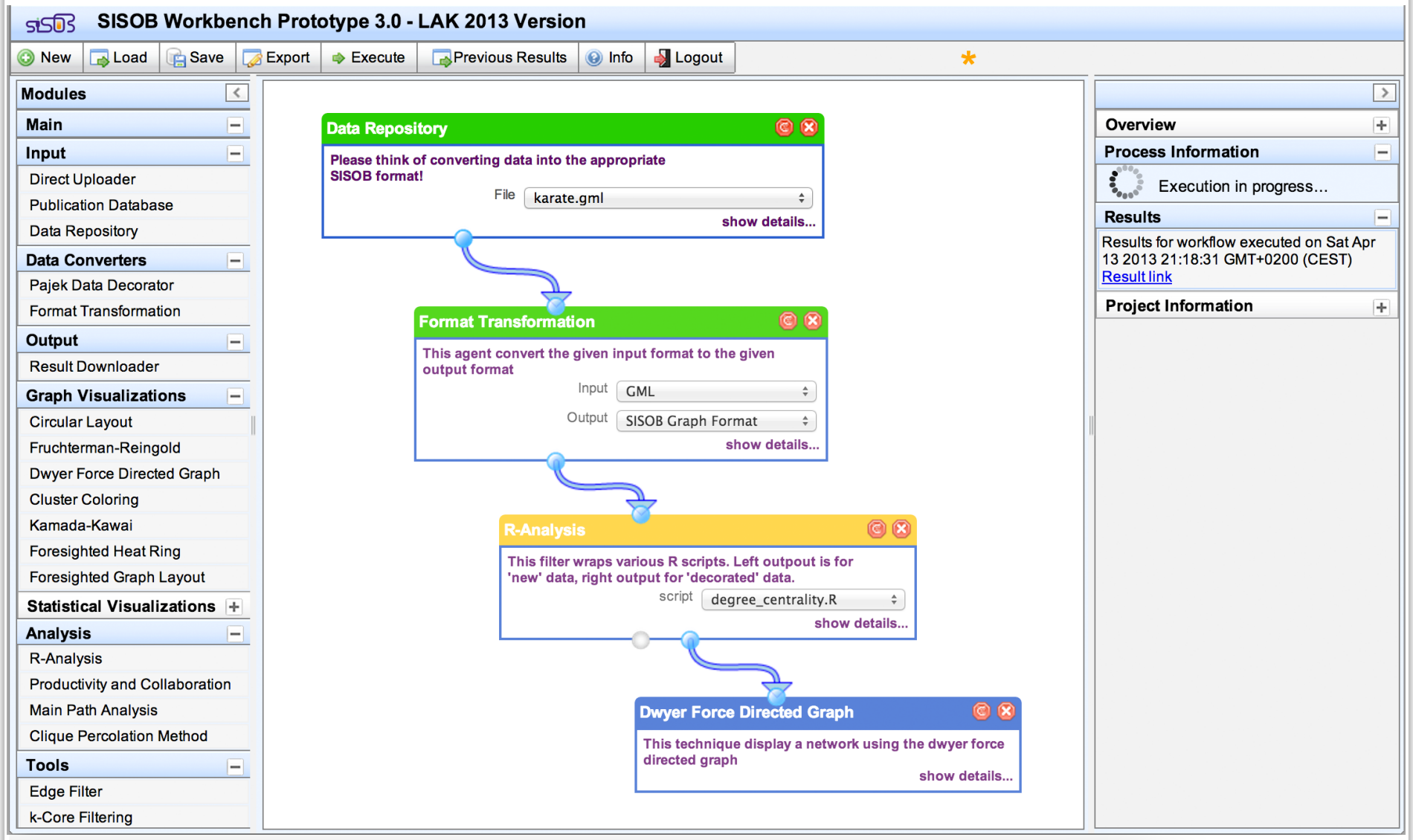
Various methods of network analysis and network modeling are increasingly used to analyze online communities. Based on network visualization

"The ***scientometer's workbench*** should provide readily available processing chains for known use cases but it should also allow for setting up new ones. The user interface should allow handling the complexity of the underlying system for users who are not computer experts. One possible approach could be a pipes and filters metaphor for processing chains."

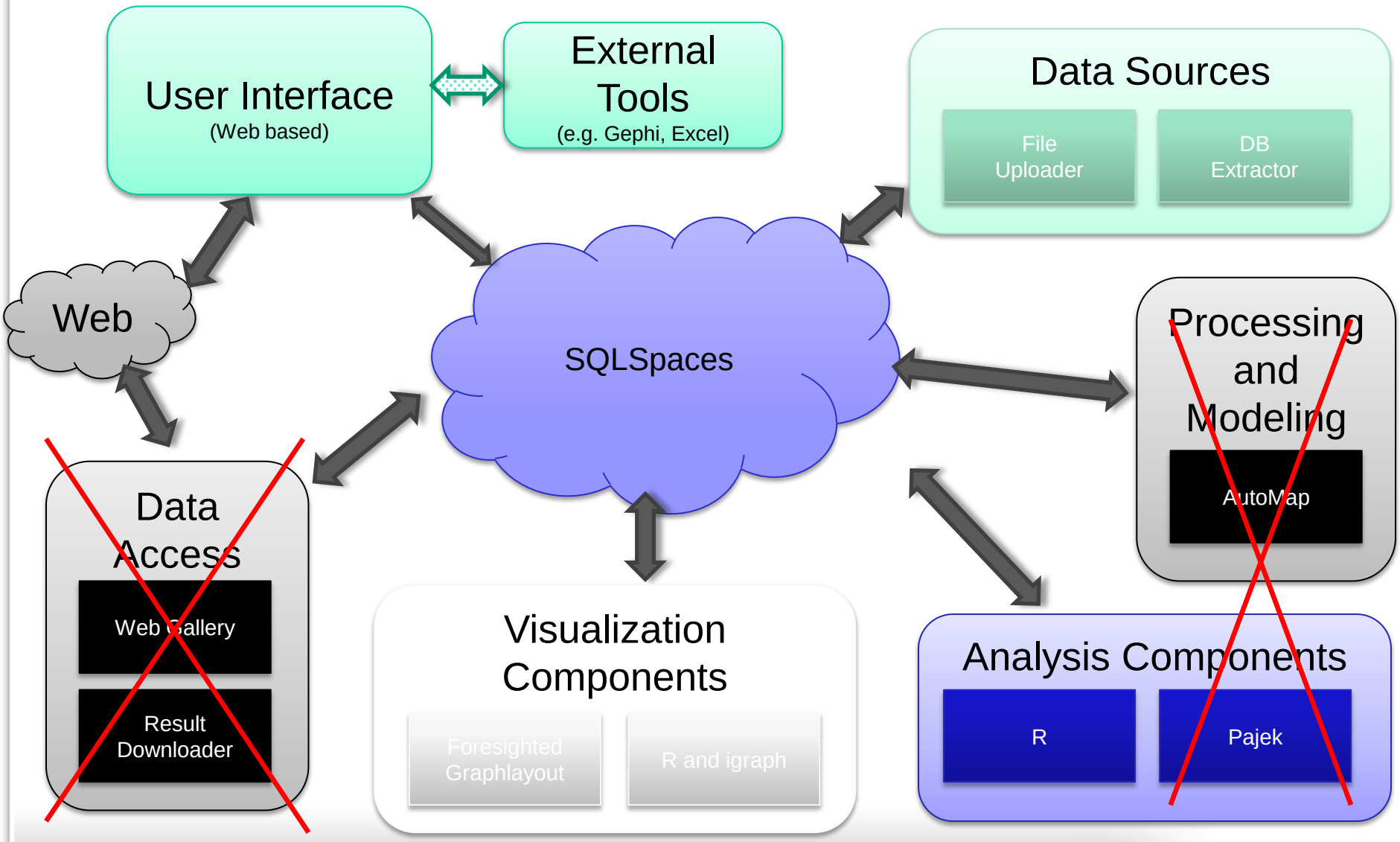
Features

- Explicit representation of analysis workflows based on a „pipes & filters“ metaphore
- Saving/loading of workflows + access to previous results
- Information about state of analysis process
- Web-based interface (runs in browser)

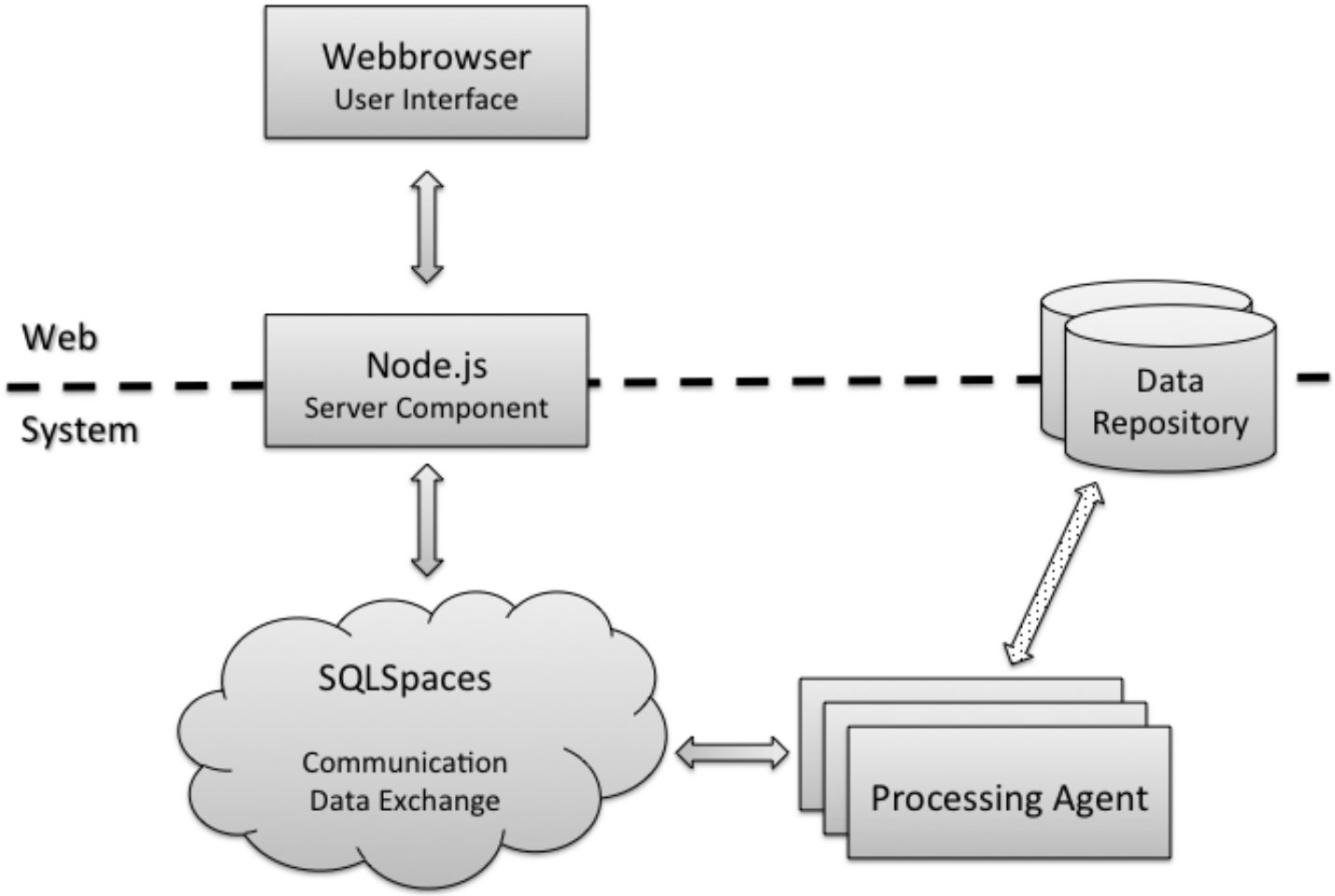
User Interface



Original Architecture (2011/12)



Current Architecture



SQLSpaces as Communication Platform

- Loose coupling of components based on tuple protocol
  - *Command Tuples*
    for initiating and monitoring process
  - *Data Tuples*
    for transporting data between agents

- Language heterogeneity
  - Java
  - Python
  - JavaScript
  - Prolog

component implementation

Data Formats

Internal formats, based on JSON

- Graph format (similar to Graph ML)
- Table format (based similar to graph format)

Format transformation

- Graphs: Pajek .net, UCINET DL, GML

Analysis Agents

Dedicated Agents

- Main Path Analysis
- Clique Percolation Method
- Productivity and Collaboration

„Allround“ Agent: R-Analysis

- Wrapper for R statistical computing language
- Allows inclusion of R scripts
- Also used as basis for more complex modules,
  e.g. k-core filter or main path analysis

Visualization Agents

Statistical Visualization

- Bar Chart
- Donut Chart
- Line Chart
- Scatter Plot
- Stacked Area Chart
- Stacked Bar Chart

Graph Visualization

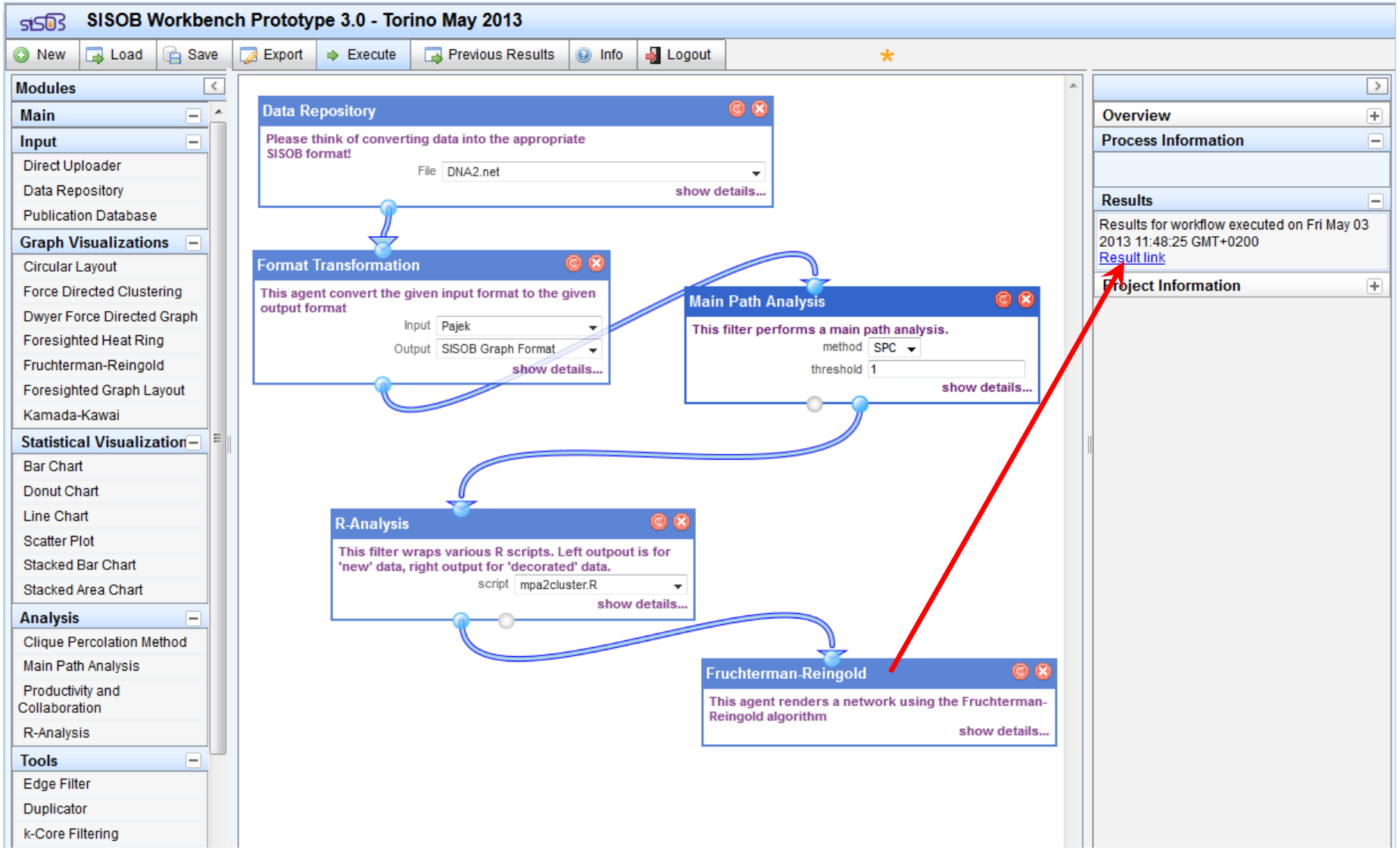
- Dwyer Force Directed Graph
- Force Directed Clustering
- Circular Layout
- Fruchterman-Reingold
- Kamada-Kawai
- Foresighted Graph Layout
- Foresighted Heat Ring

Visualization Architecture

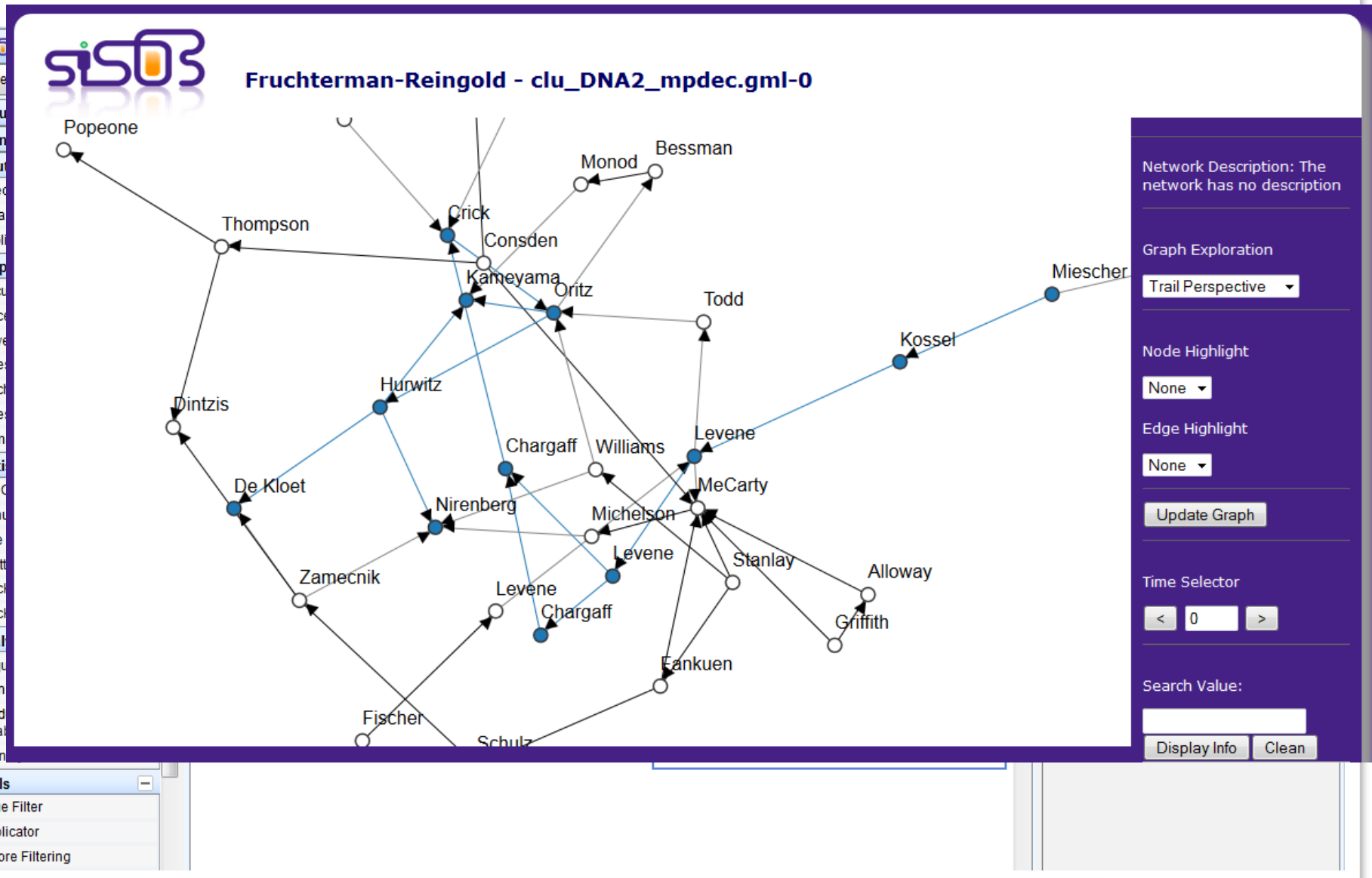
… supports both server-side and client-side processing:

- server-side: based on JUNG (*Java* Universal Network/Graph Framework)
- client-side: based on D3.js (*JavaScript* library to display digital data in dynamic graphical forms)

Example


SISOB Workbench Prototype 3.0 - Torino May 2013
New
Load
Save
Export
Execute
Previous Results
Info
Logout
Modules
Main
Input
Direct Uploader
Data Repository
Publication Database
Graph Visualizations
Circular Layout
Force Directed Clustering
Dwyer Force Directed Graph
Foresighted Heat Ring
Fruchterman-Reingold
Foresighted Graph Layout
Kamada-Kawai
Statistical Visualization
Bar Chart
Donut Chart
Line Chart
Scatter Plot
Stacked Bar Chart
Stacked Area Chart
Analysis
Clique Percolation Method
Main Path Analysis
Productivity and Collaboration
R-Analysis
Tools
Edge Filter
Duplicator
k-Core Filtering
Data Repository
Please think of converting data into the appropriate SISOB format!
File DNA2.net
show details...
Format Transformation
This agent convert the given input format to the given output format
Input Pajek
Output SISOB Graph Format
show details...
Main Path Analysis
This filter performs a main path analysis.
method SPC
threshold 1
show details...
R-Analysis
This filter wraps various R scripts. Left outpout is for 'new' data, right output for 'decorated' data.
script mpa2cluster.R
show details...
Fruchterman-Reingold
This agent renders a network using the Fruchterman-Reingold algorithm
show details...
Overview
Process Information
Results
Results for workflow executed on Fri May 03 2013 11:48:25 GMT+0200
Result link
Project Information

Example

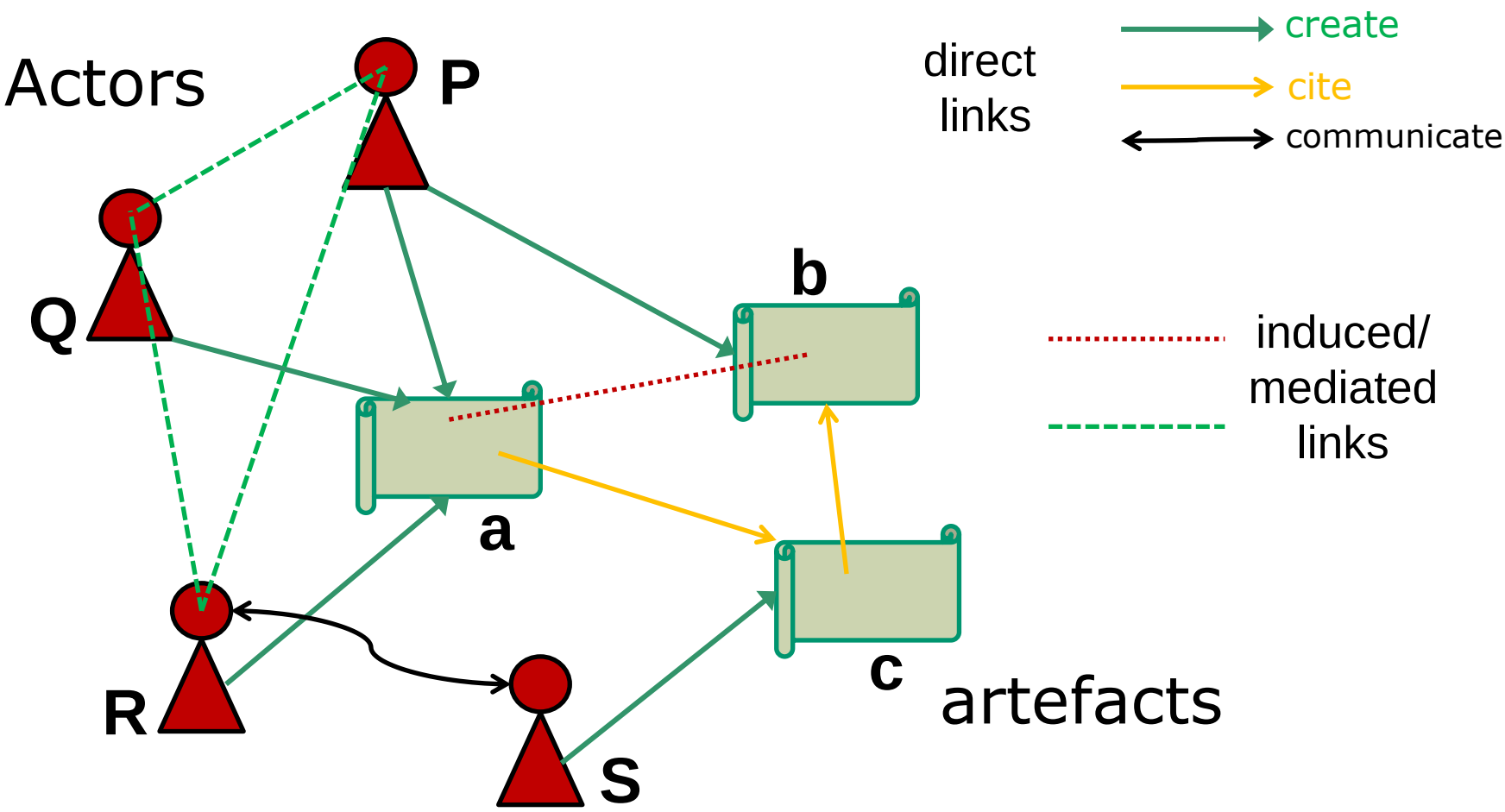

Fruchterman-Reingold - clu_DNA2_mpdec.gml-0
Popeone
Bessman
Monod
Crick
Thompson
Consden
Kameyama
Oritz
Todd
Miescher
Kossel
Hurwitz
Dintzis
Chargaff
Williams
Levene
De Kloet
MeCarty
Nirenberg
Michelson
Levene
Stanlay
Alloway
Zamecnik
Levene
Chargaff
Griffith
Fankuen
Fischer
Schulz
Network Description: The network has no description
Graph Exploration
Trail Perspective
Node Highlight
None
Edge Highlight
None
Update Graph
Time Selector
<
0
>
Search Value:
Display Info
Clean

Case Study 2:

Analysing the evolution of ideas

with Iassen Halatchliyski, Tobias Hecking, Tilman Göhnert

Relations around epistemic artefacts



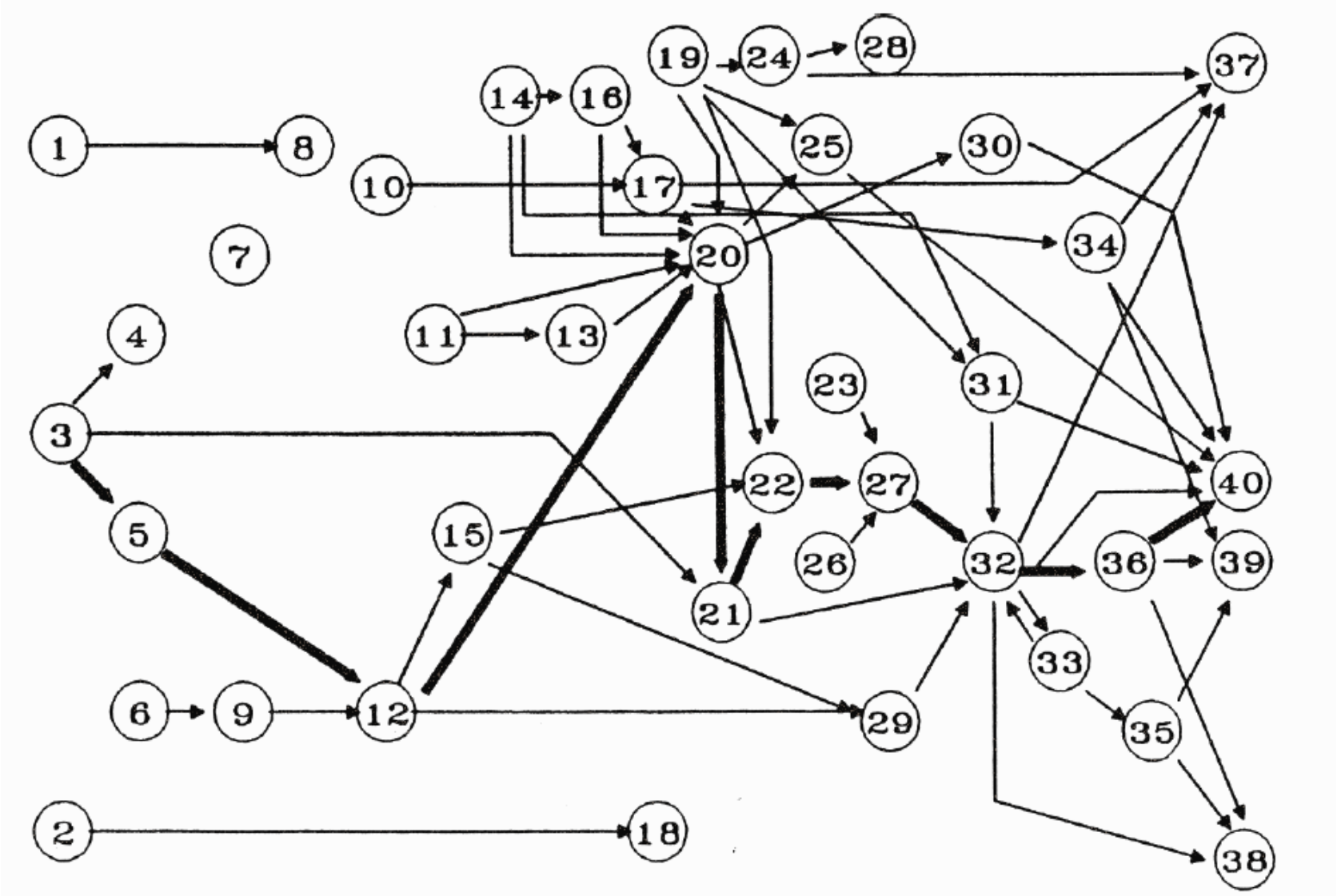
Main Path Analysis - guiding question

Given a graph of dependencies between knowledge items, how to identify the main pathways of the evolution of knowledge?

Concretely - use *references* (citation links) *relations* between publications

=> acyclic graph (no cross references!)

DNA Example (Hummon & Doreian, 1989)



Hummon, N. P. and Doreian, P. 1989. Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks*, 11, 39-63.

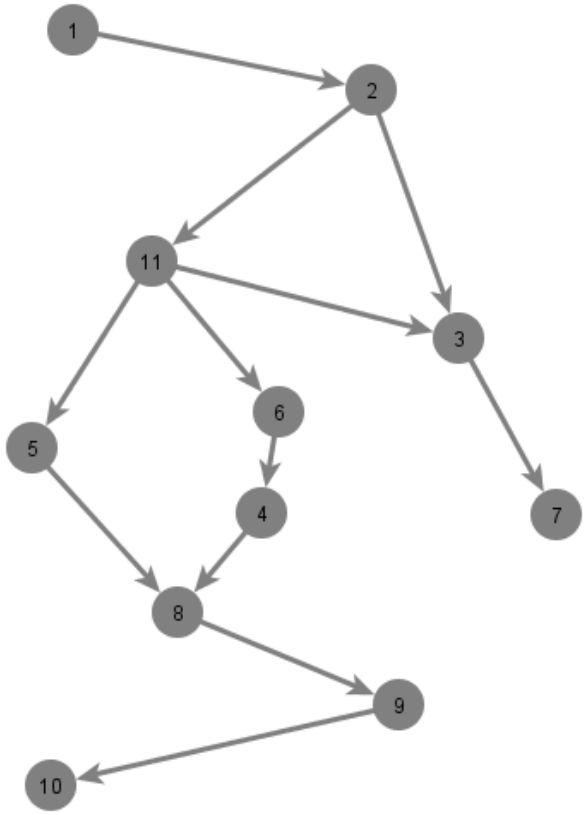
Main Path Analysis - formal properties

- Citation networks are DAGs (Directed Acyclic Graphs)

- Implicit notion of time:
  - sources are the oldest (or not cited) publications
  - sinks are the newest ones

  (inverse citation relation)

- DAGs have always at least one source and one sink node:
  - add one virtual source node “before” all original sources
  - add one sink node “after” all original sinks

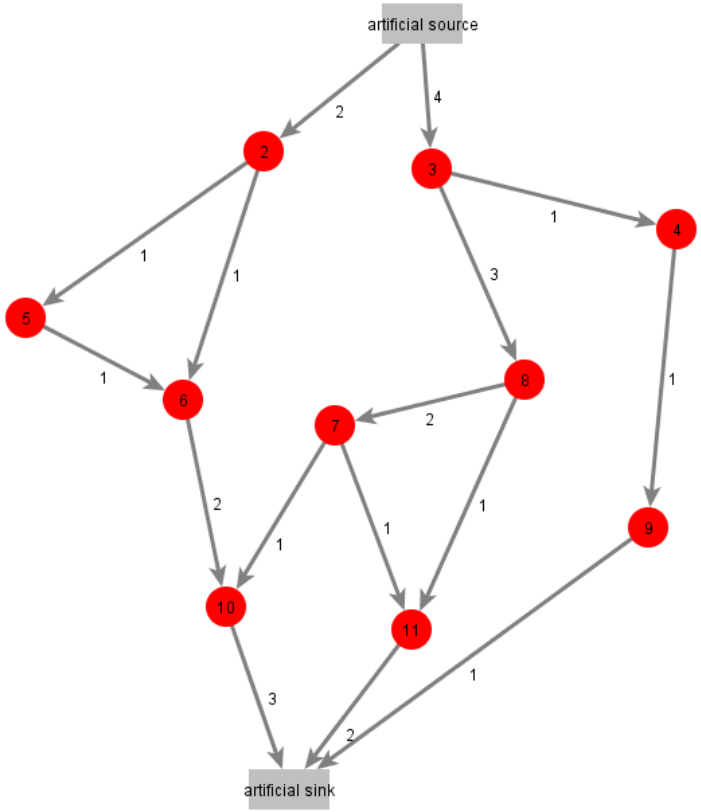


Main Path Analysis - procedure

1st step Calculate edge weight according to a weighting scheme.

Example: Search Path Count
(Batagelj, 2003)

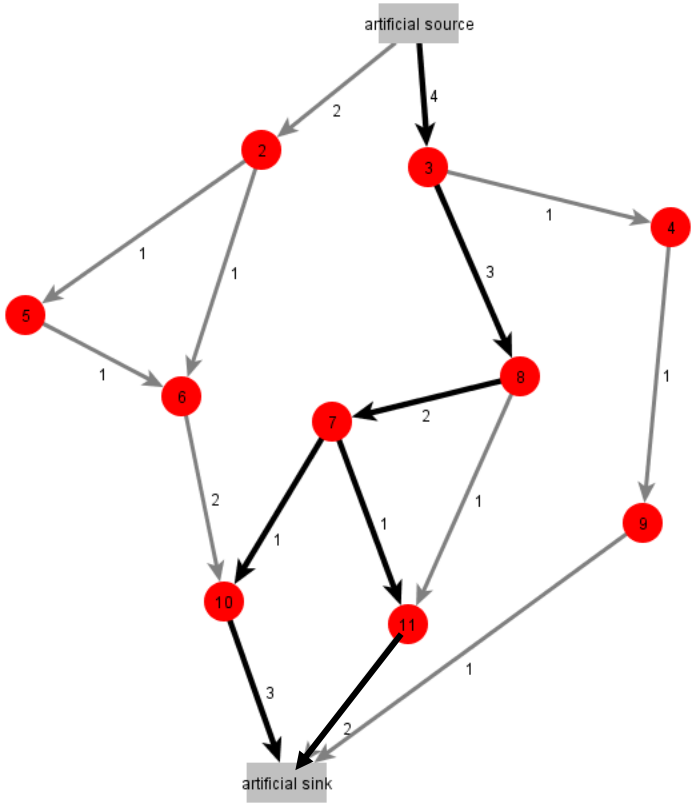
Edge weight corresponds to the number of paths from a single source to a single sink node an edge occurs.



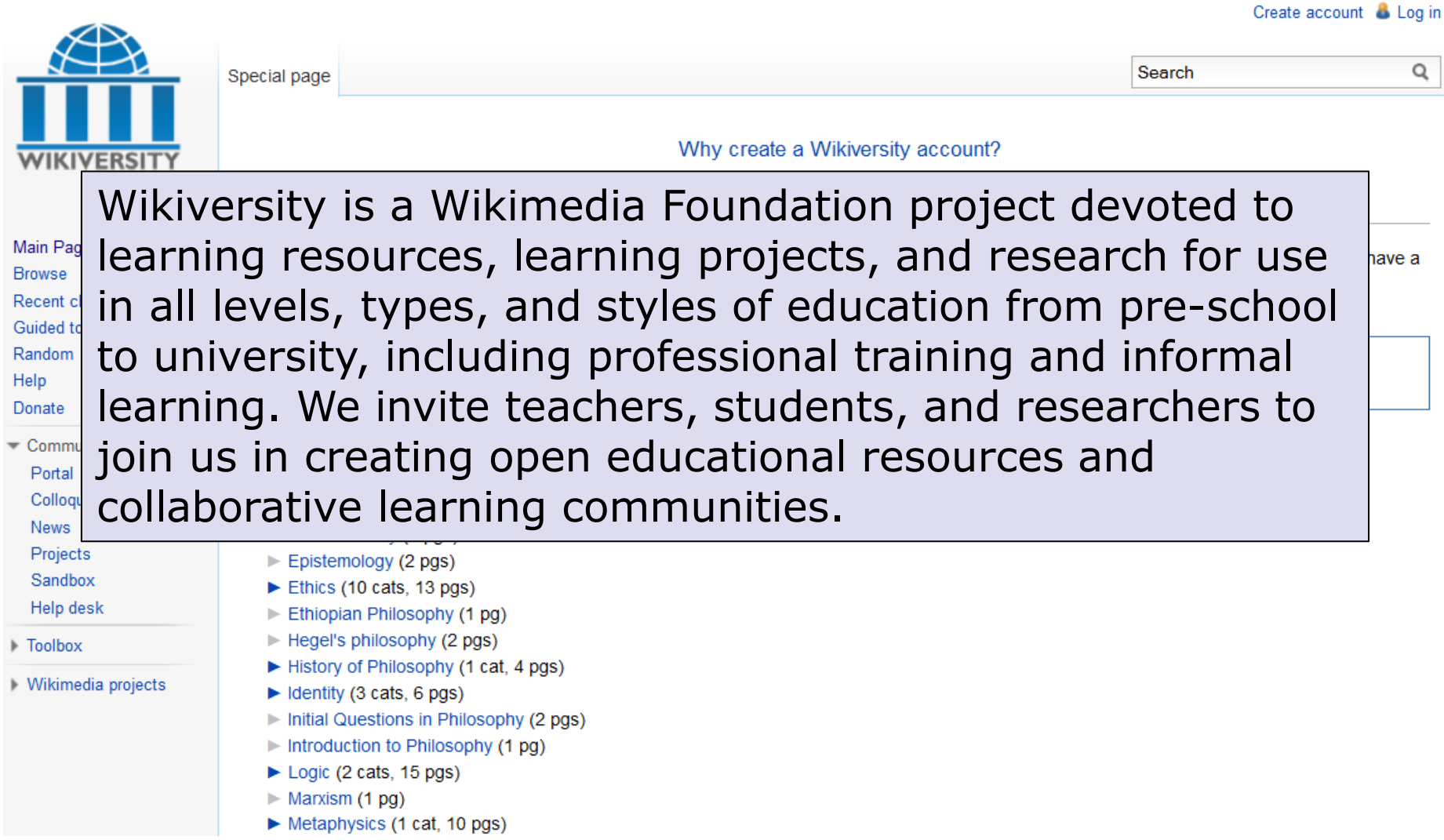
Main Path Analysis - procedure

2nd step:

Traverse graph from source to sink by taking the edges with the highest weights.



Learning Resources in Wikiversity



Learning Resources in Wikiversity

Why create a Wikiversity account?

Category tree

Enter a category name to see its contents as a tree structure. Note that this requires advanced JavaScript functionality known as AJAX. If you have a very old browser, or have JavaScript disabled, it will not work.

Show category tree

Category: Philosophy categories only Show tree

Parents: *Humanities*

- ▼ Philosophy (28 cats, 128 pgs)
  - ▶ Continental philosophy (5 pgs)
  - ▶ Critical theory (3 pgs)
  - ▶ Epistemology (2 pgs)
  - ▶ Ethics (10 cats, 13 pgs)
  - ▶ Ethiopian Philosophy (1 pg)
  - ▶ Hegel's philosophy (2 pgs)
  - ▶ History of Philosophy (1 cat, 4 pgs)
  - ▶ Identity (3 cats, 6 pgs)
  - ▶ Initial Questions in Philosophy (2 pgs)
  - ▶ Introduction to Philosophy (1 pg)
  - ▶ Logic (2 cats, 15 pgs)
  - ▶ Marxism (1 pg)
  - ▶ Metaphysics (1 cat, 10 pgs)

Main Page
Browse
Recent changes
Guided tours
Random
Help
Donate

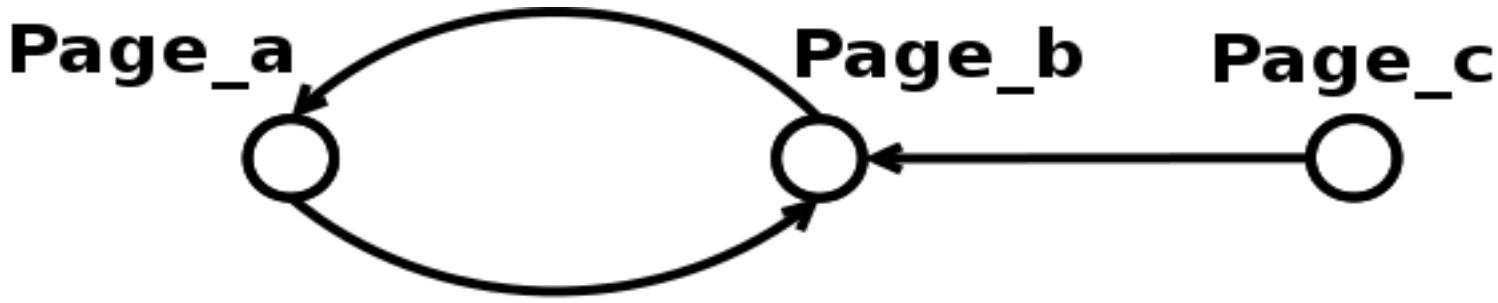
Community
Portal
Colloquium
News
Projects
Sandbox
Help desk

Toolbox

Wikimedia projects

Structure of (collaborative) hypermedia

- Web pages are connected by hyperlinks, often bi-directional
- No implicit notion of time
- The resulting graph is *not* a DAG!



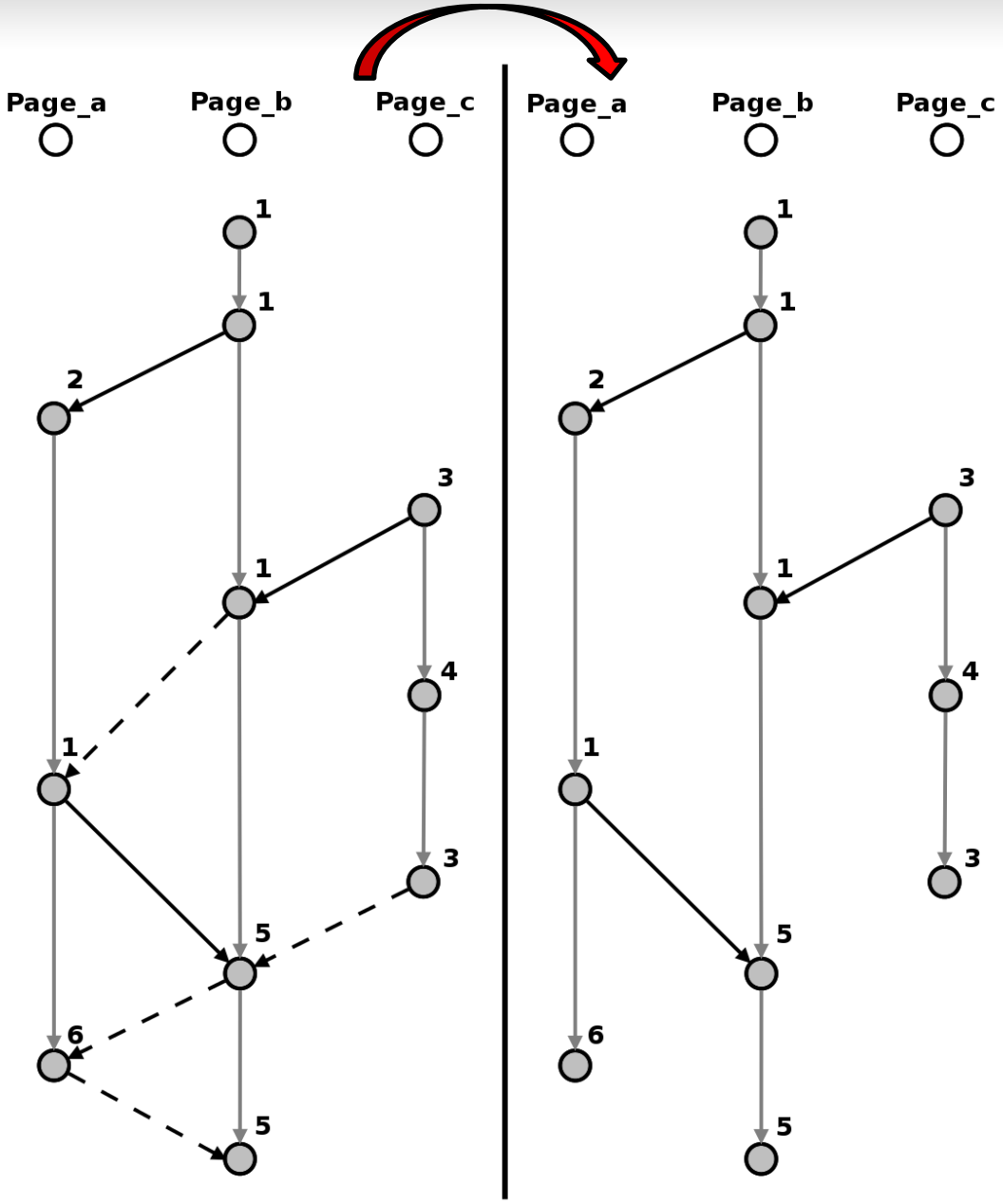
Solution / modification

- introduce version relations (updates)

  => DAG property

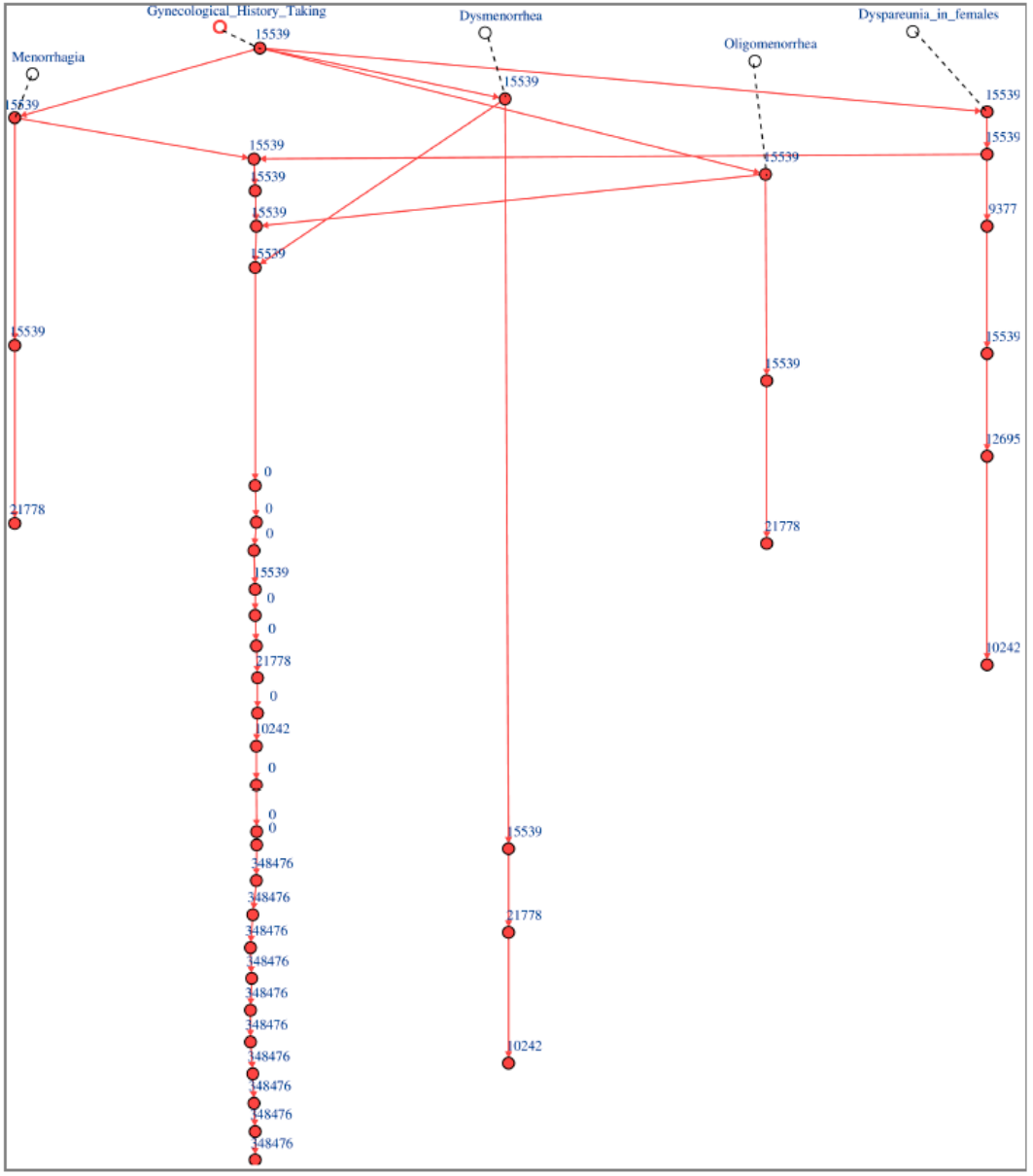
- use swim lane diagrams

- filter out of redundant edges

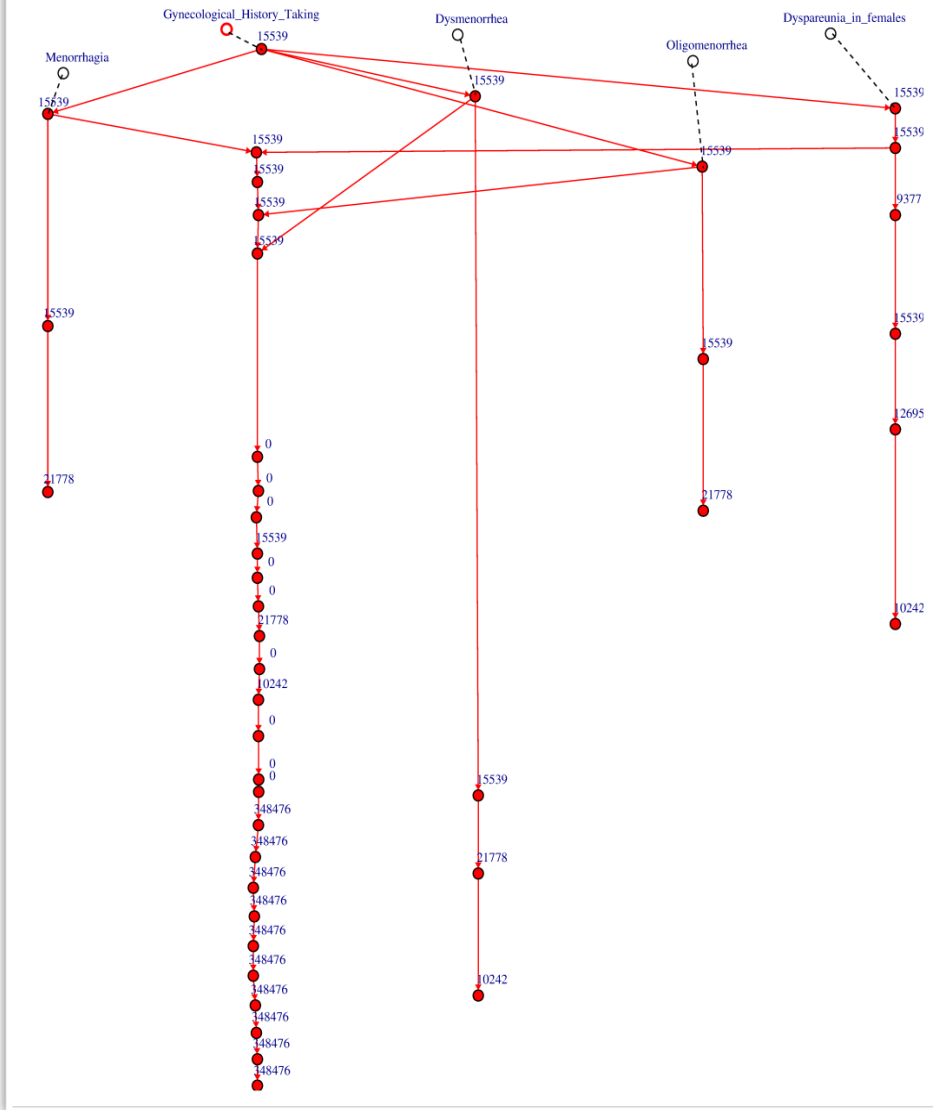


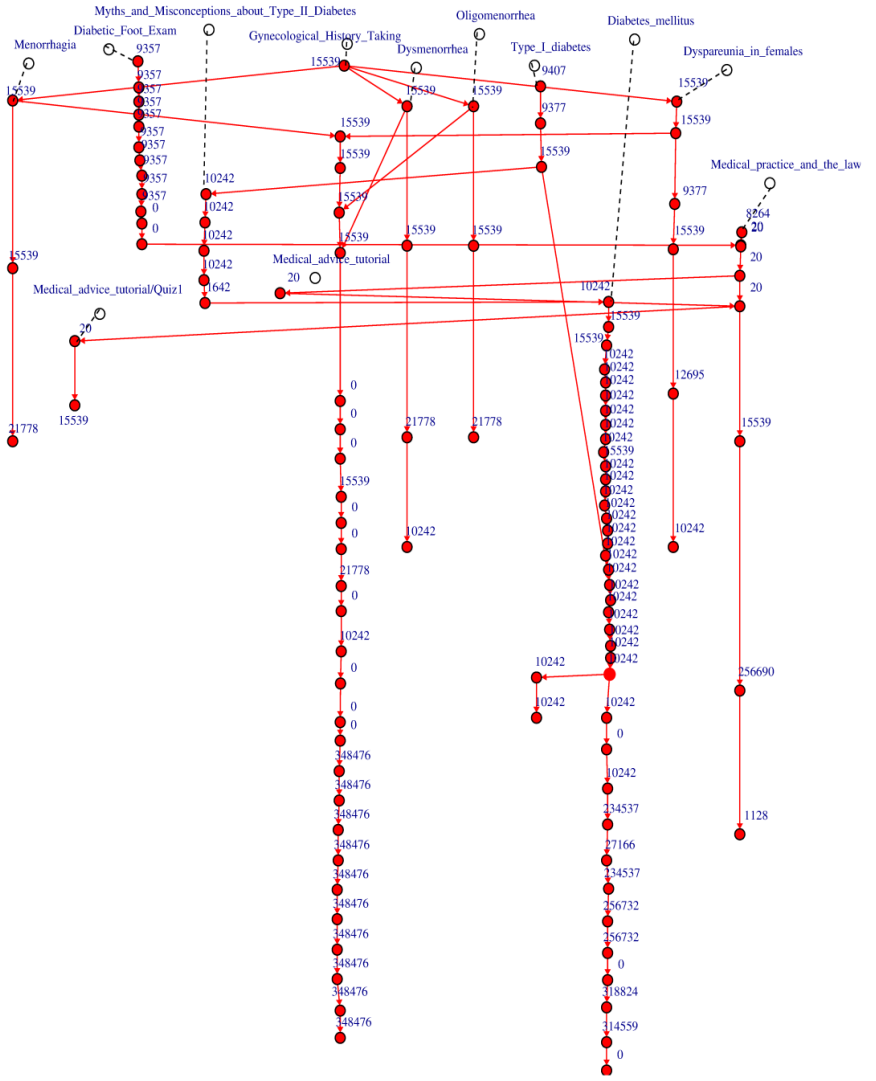
Wikiversity Example

Main path of Wikiversity category Medicine

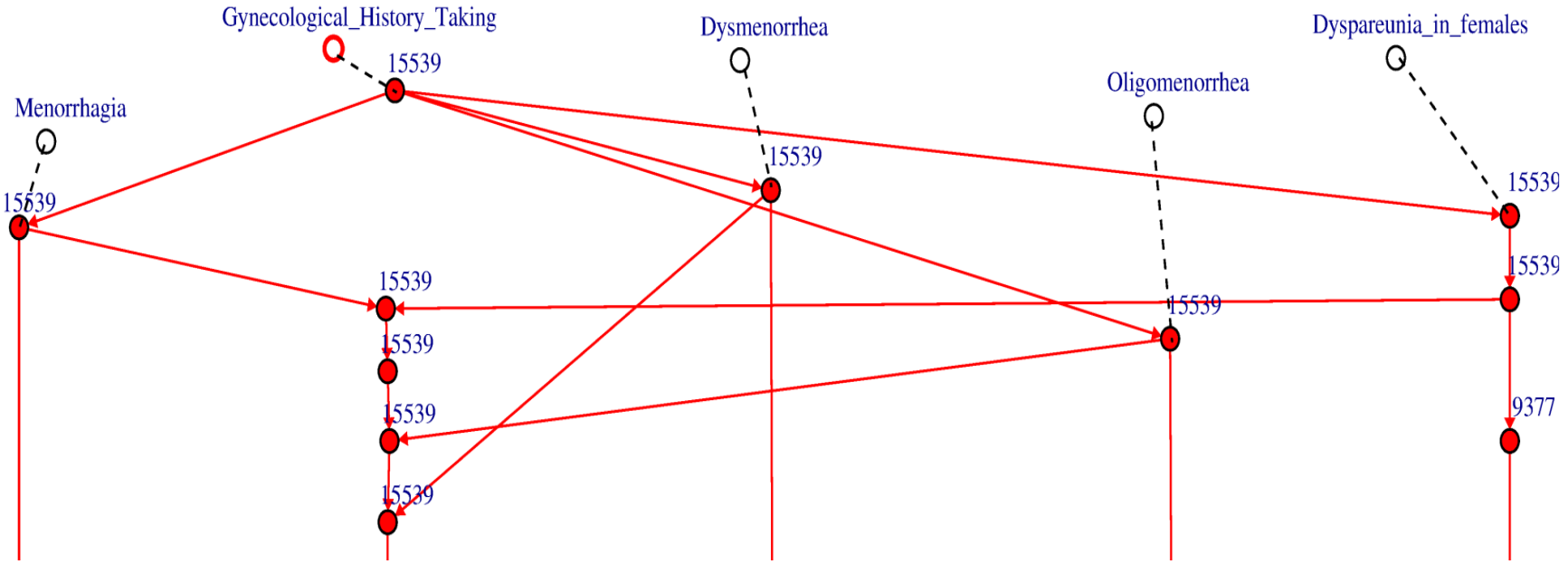


Variation of branching (simple vs. multiple)

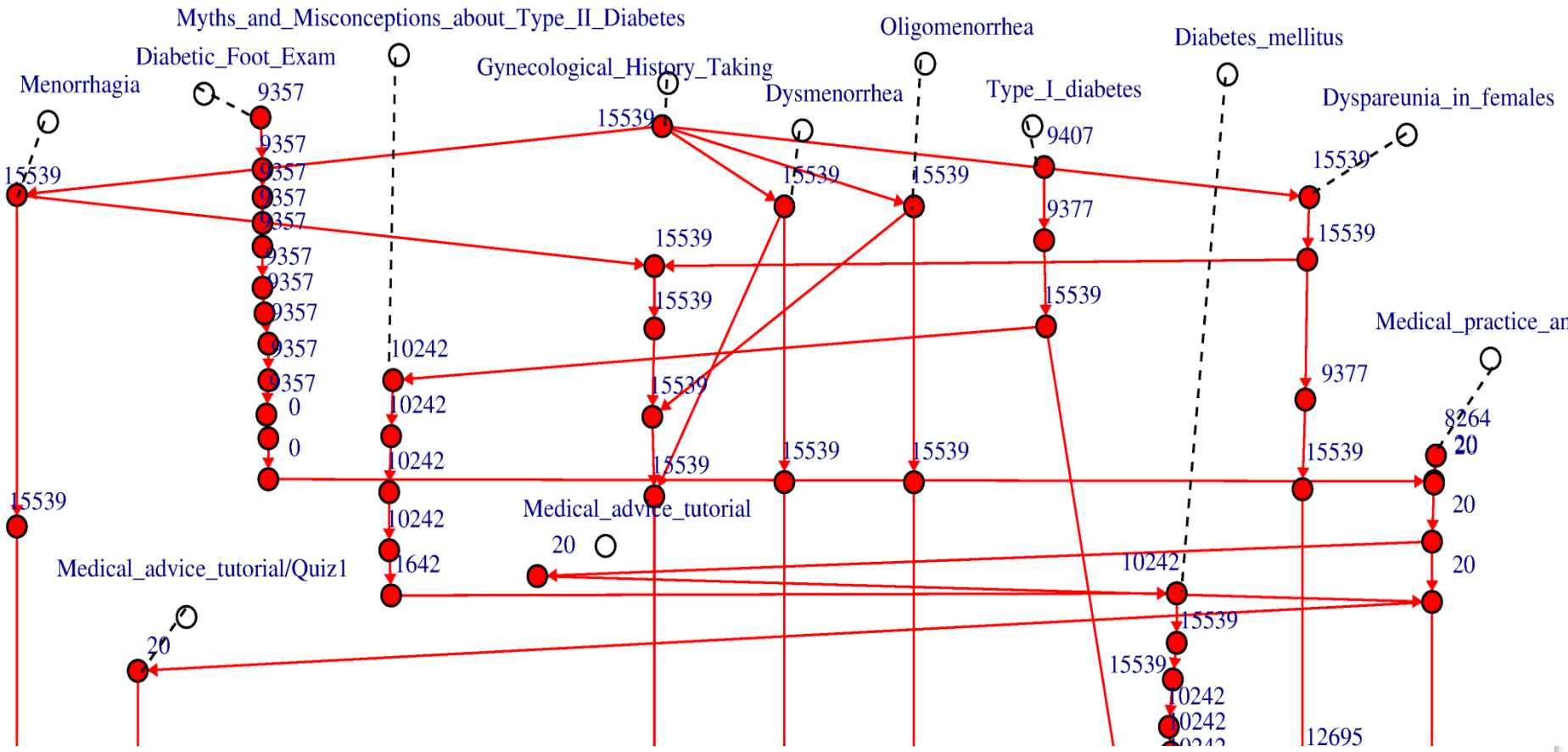




Main Path - low branching (equal weights)
Author 15539 interlinks several concepts



Multiple main path - higher branching (threshold)



Reflection on added value …

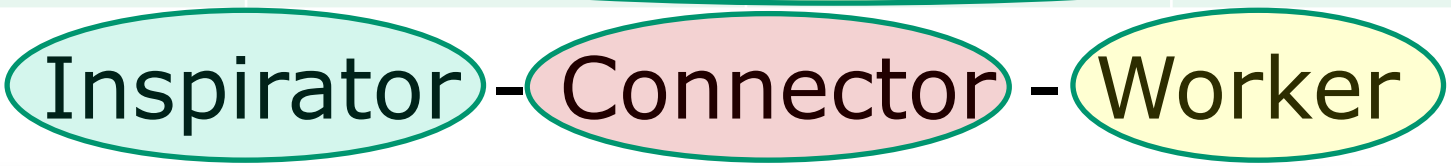
What ***insights*** do we gain from this knowledge flow analysis?

⇒ Better understanding of contributions in context!

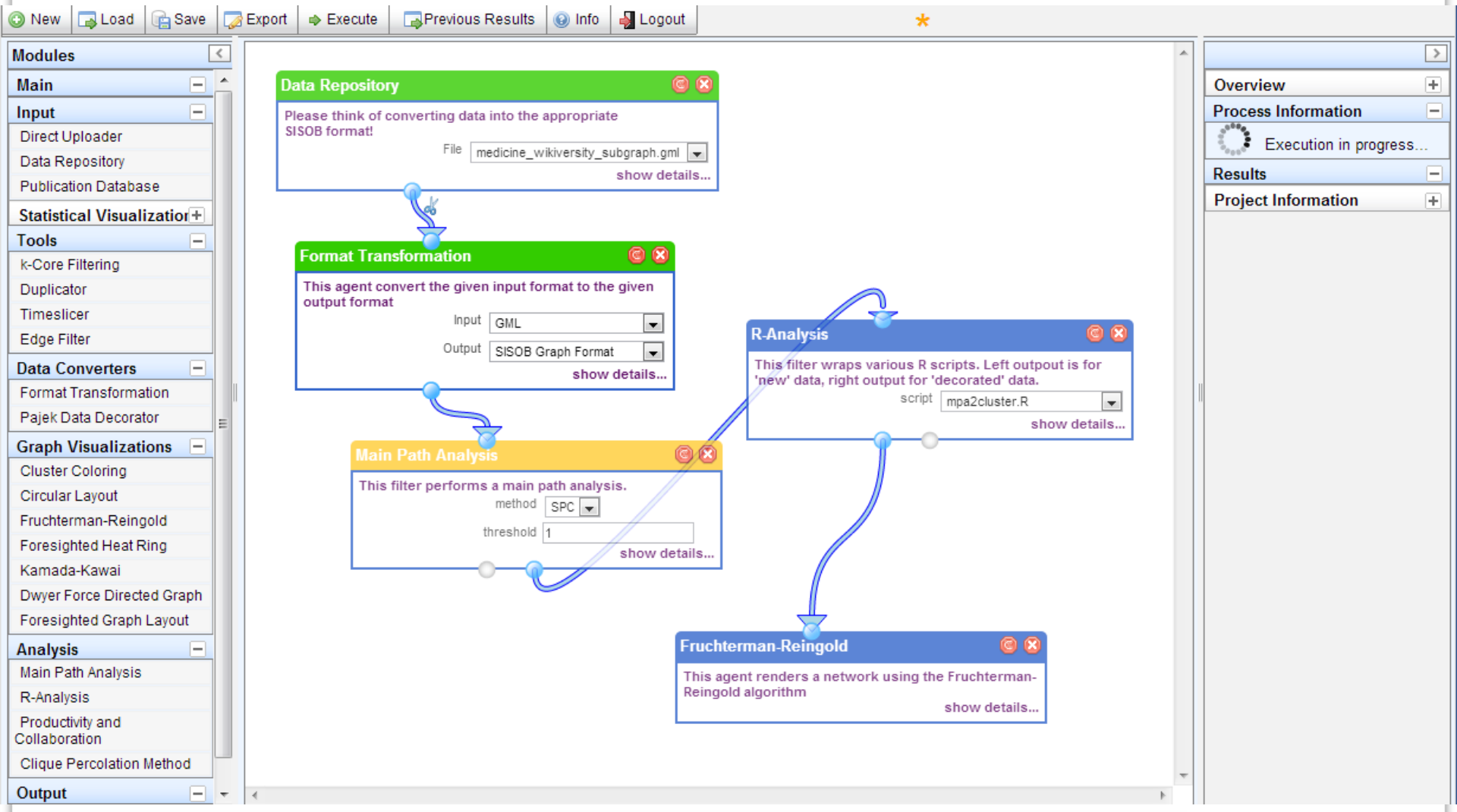
⇒ Application: user/learner profiling

User characteristics - roles

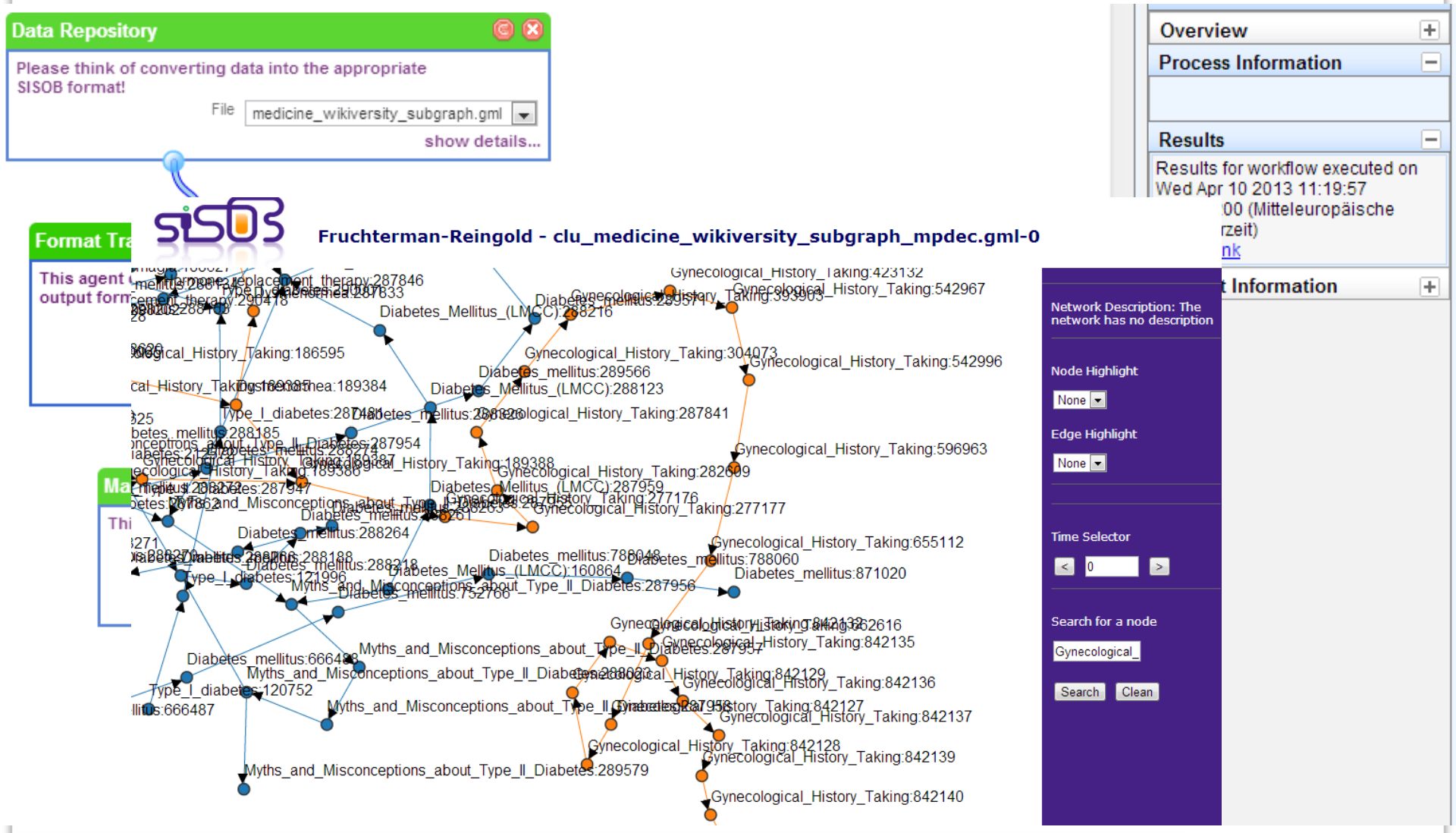
Author Id	# contributions / ... on main path	# references received / on main path	# references established / on main path
10242	272 / 34	10 / 1	14 / 4
15539	253 / 21	20 / 10	15 / 8
12385	98 / 0	1 / 0	0 / 0
9357	349 / 10	2 / 0	1 / 0



Mapping MPA to the Workbench



Mapping MPA to the Workbench



Case Study 3:

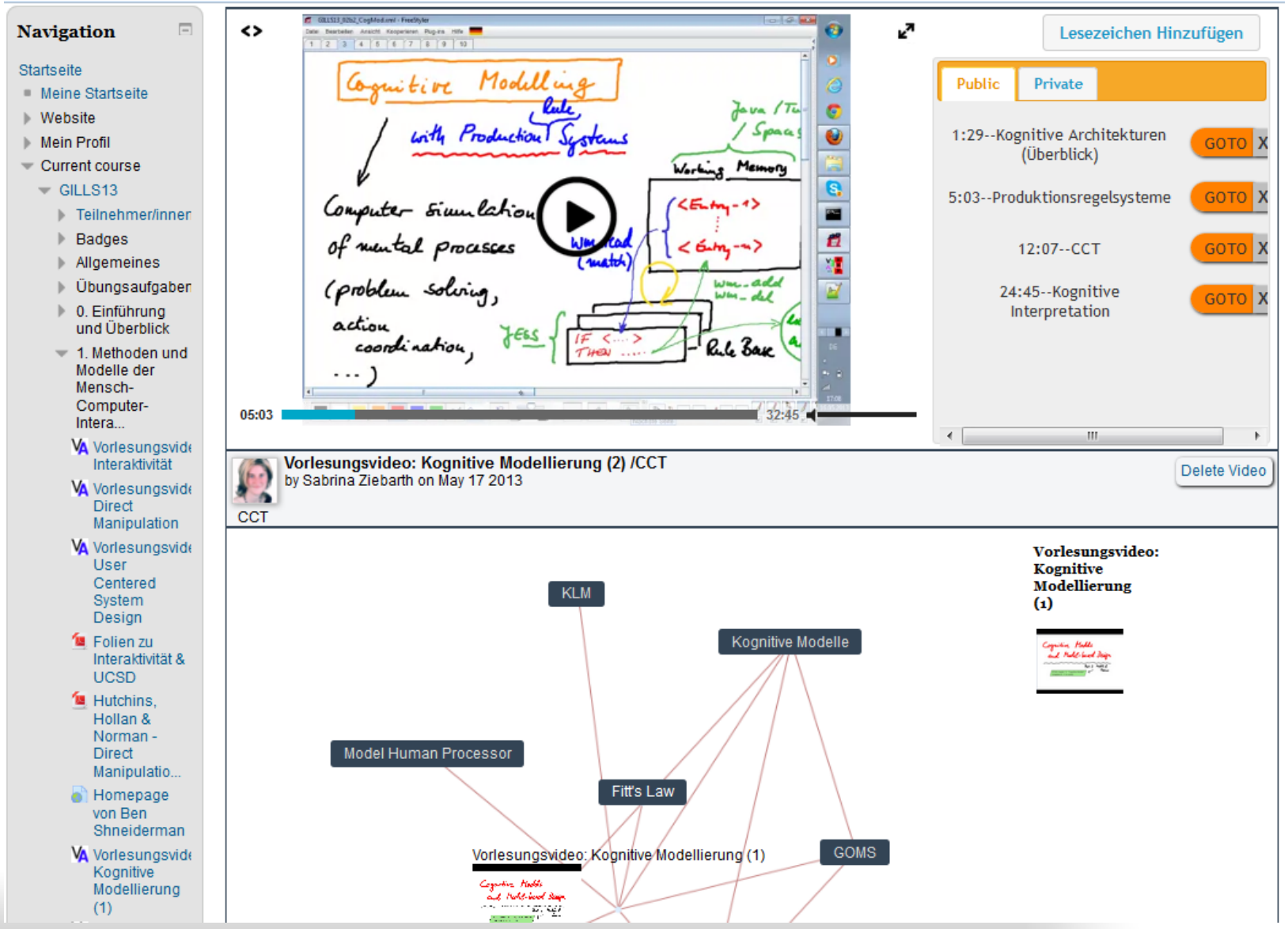
Ressource Access Patterns on a Learning Platform

with Tobias Hecking, Sabrina Ziebarth

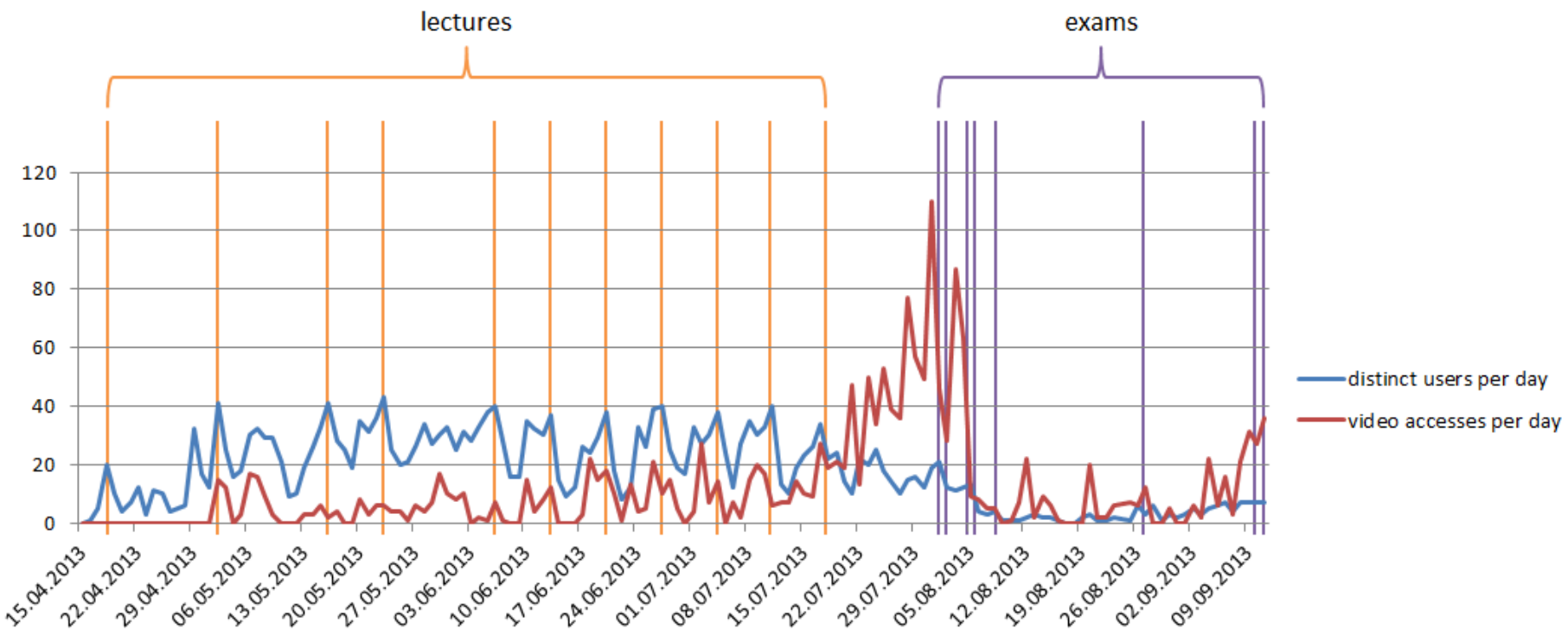
Example & Data Source: A Resource-intensive Blended Learning Course

- Regular master-level course on *„Modeling & designing interactive systems and educational environments“*
- 44 participants (about 80% presence in lecture)
- Regular exercises replaced by online activities
- Moodle with extensions (e.g. for video) as a learning platform
- Oral exams (36 participants)

Platform View



Platform and Video Access



Analysis of Resource Access Patterns

Presence of various media, materials and activities in resource-intensive online courses enable more autonomous, self-regulated learning.

?

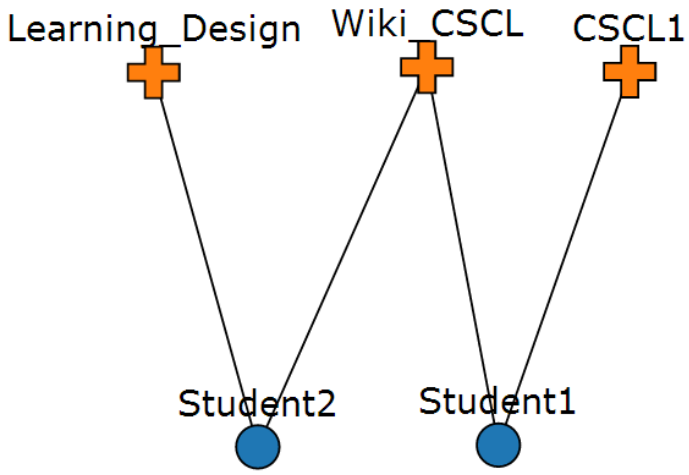
Patterns of resource access over time can lead to valuable insights in these learning processes.

→ Network analysis methods to investigate relations between groups of students and learning resources.

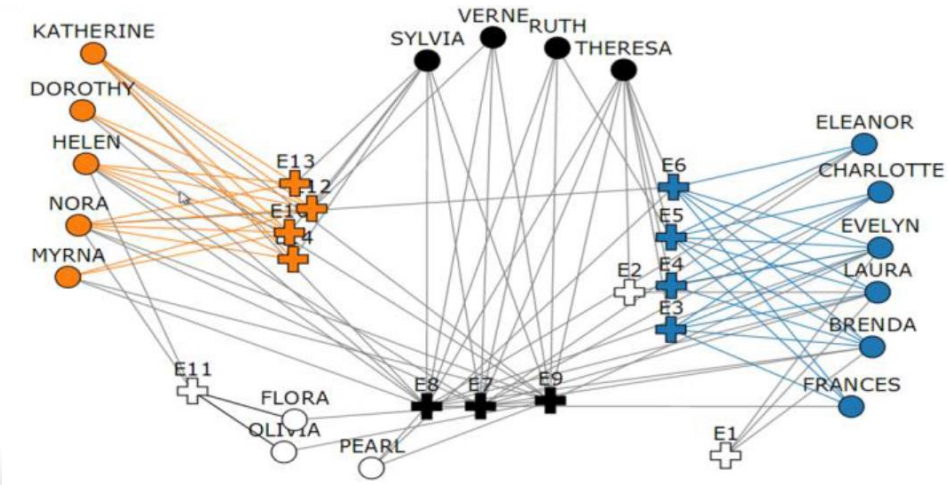
From log files to student resource networks

Timestamp	Name	Action	Info
07112013-1400	Student 1	Video view	CSCL 1
07112013-1430	Student 2	Video view	Learning Design
07112013-1435	Student 2	Resource view	Wiki CSCL
07122013-1500	Student 1	Resource view	Wiki CSCL

Two-mode network: actors - artefacts



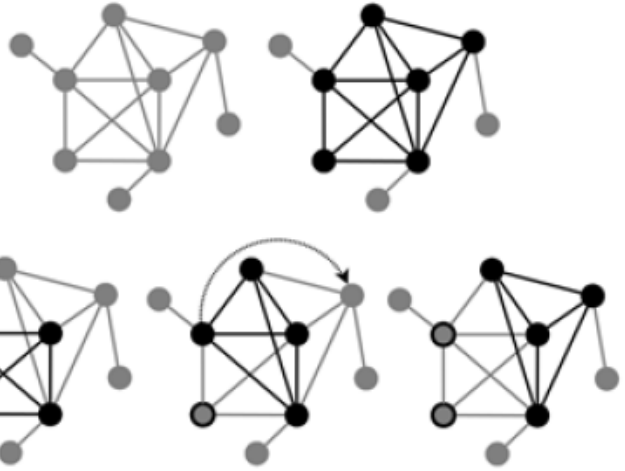
Two-mode sub-communities ?

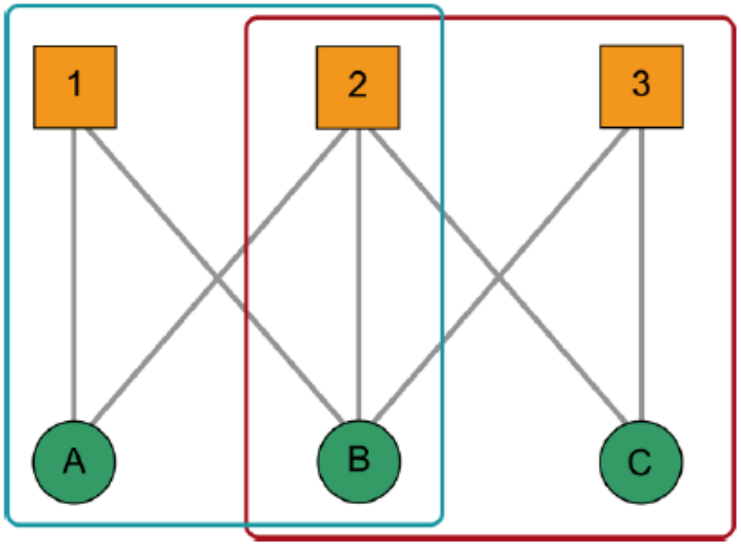


Method: Bi-Clique Percolation

Idea of CPM: a k-clique (complete subgraph of size k) "percolates" through the graph

k is input parameter

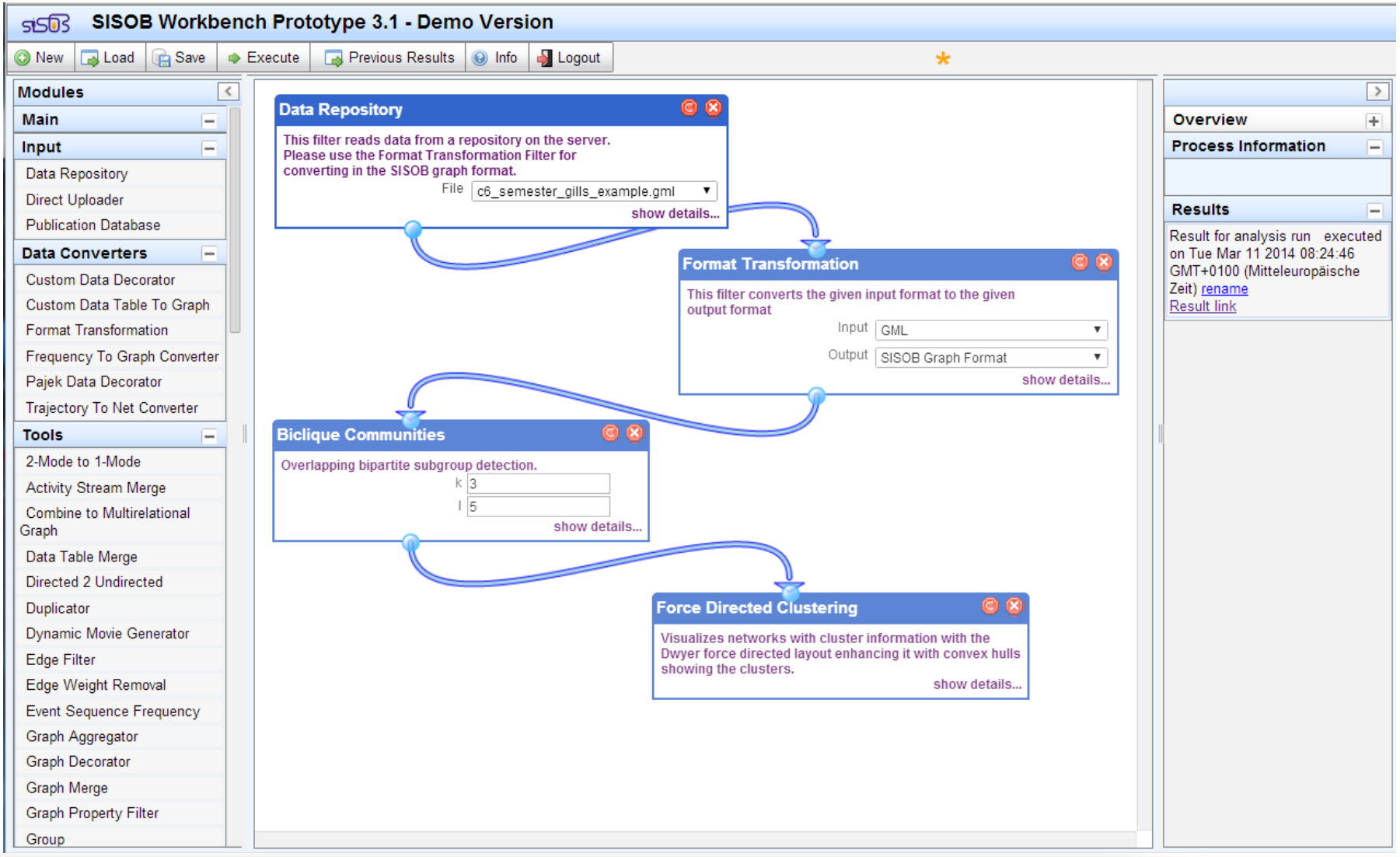




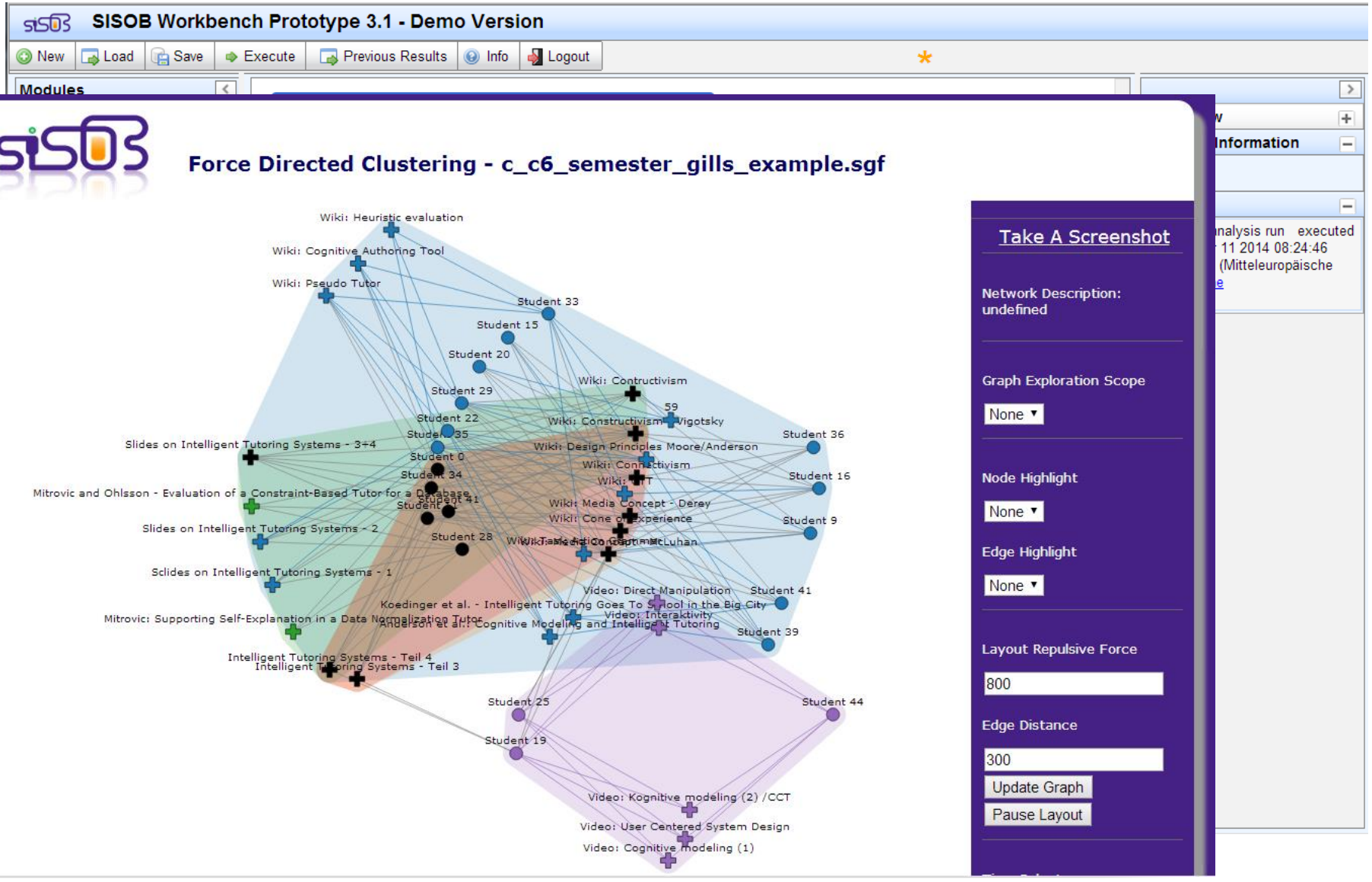
Extension for bi-cliques

(completely connected bi-partite graphs)

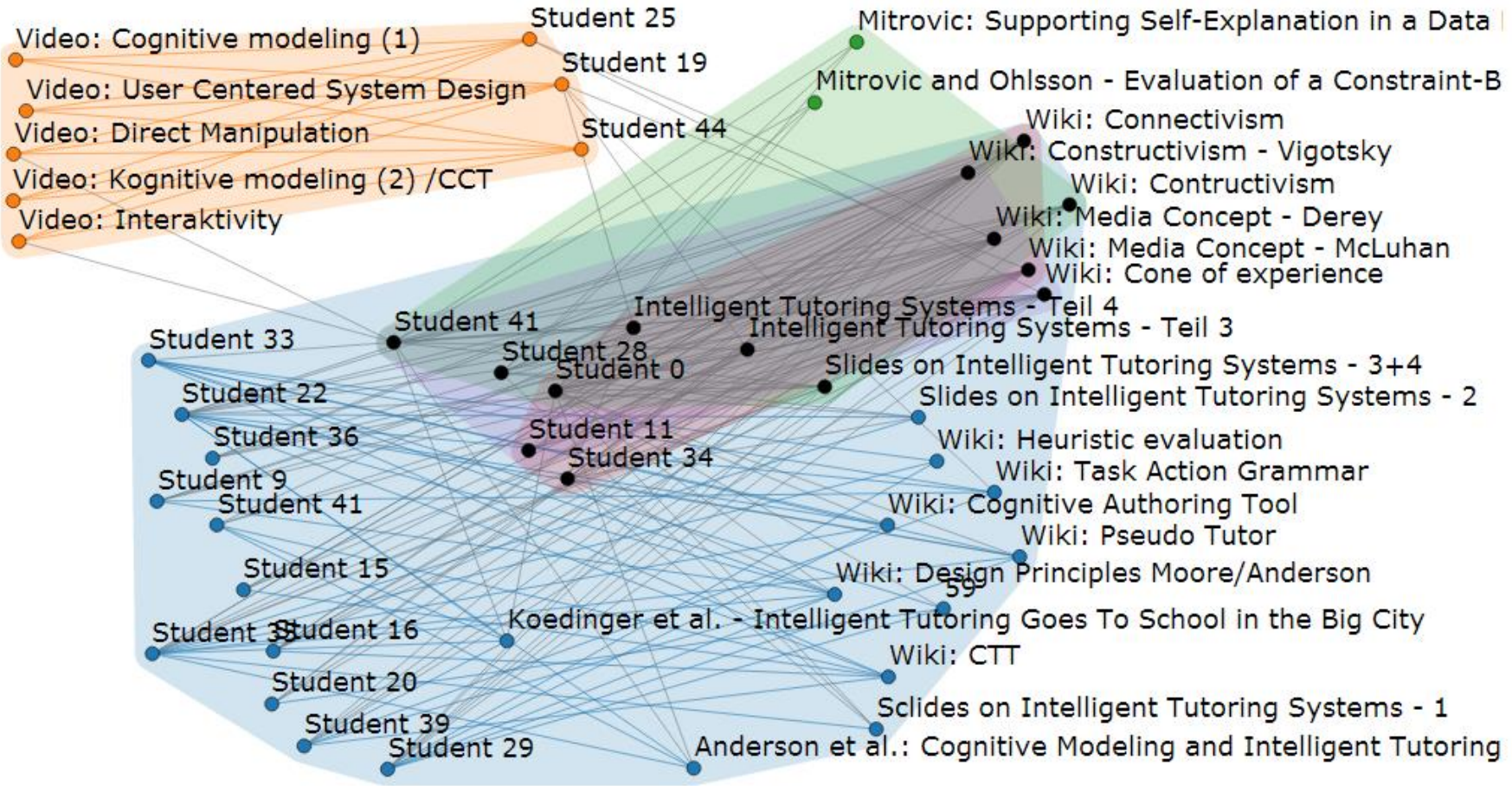
Workbench – Biclique Analysis



Workbench – Biclique Analysis



Resource Access Patterns (example)

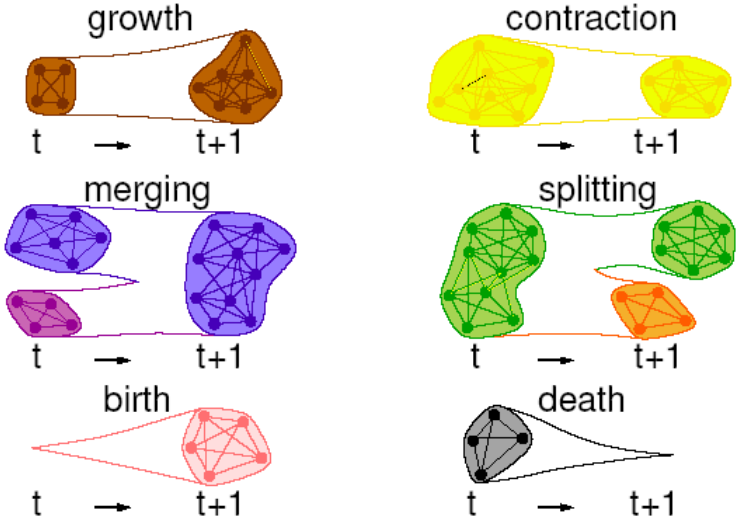


Research Challenges

- How to embed analytics results into the platform?
  → beyond hand-wired dashboards
  →“natural” interfaces for teachers and students
  → generation based on an “analytics backend”

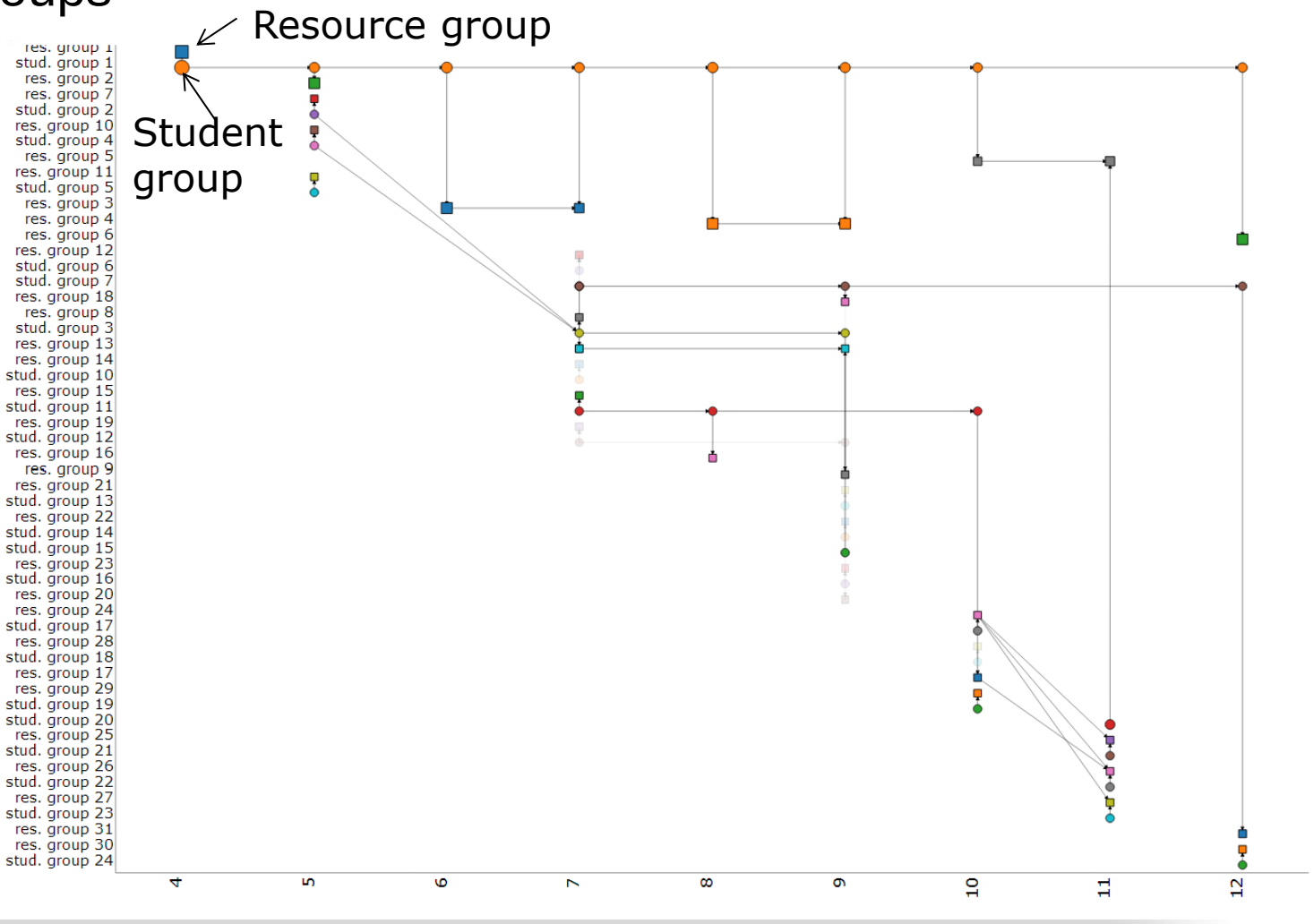
- How to track groups dynamically (considering group identity)

- How to use analytics results for “intelligent” group formation?



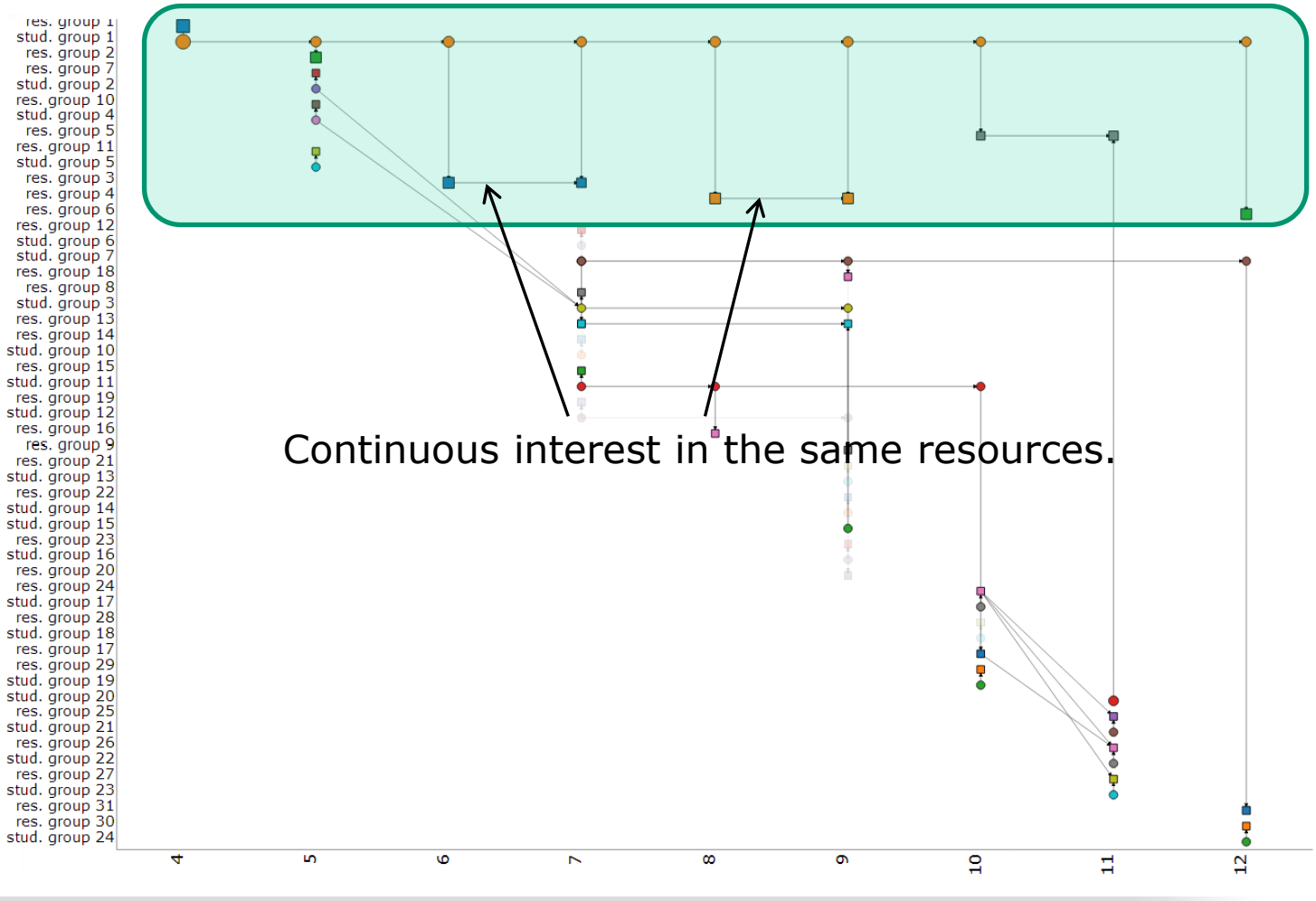
Student resource clusters over time during the lecture period

- Circles represent student groups / squares represent resource groups



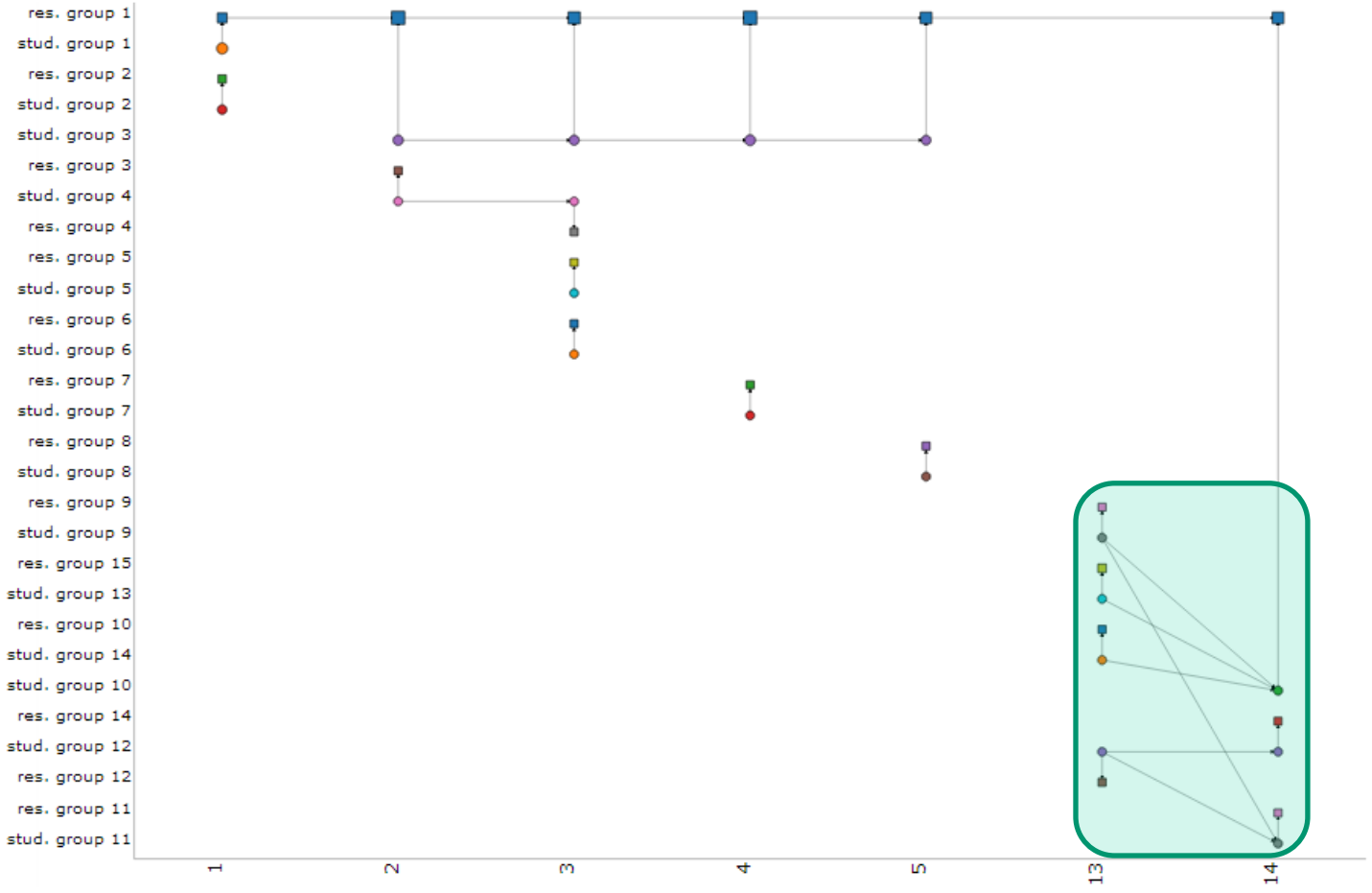
Student resource clusters over time during the lecture period

- A majority group uses the resources of the corresponding lecture week.
- Only a few small groups split.



Student resource clusters during exam preparation

- Student in the second block of oral exams behave differently in the beginning of their exam preparation phase.



Empirical Findings

- Similar behavioural pattern for a majority of students during the semester.
- Smaller groups split often and are less stable.
  - Evolution of access patterns over time reflects task assignments (as expected).
  - Unexpected grouping of students around videos.
- Continuous interest in wiki articles created by other students => peer reviewing approach works appropriately.
- Different resource usage patterns for students from different study programmes during exam phase.

Methodological Research Topics

- Dynamics of cohesive clusters in bipartite graphs
- Application of signed graphs to social media (e.g. using Wikipedia admin election votes)
- Multi-relational blockmodeling
- Detection of motifs (small subgraphs of a specific type) in complex networks

Literature

Batagelj, V. (2003). Efficient algorithms for citation network analysis. arXiv preprint cs/0309023.

Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3), 75-174.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, *11*(1), 39-63.

Lehmann, S., Schwartz, M., & Hansen, L. K. (2008). Biclique communities. Phys.Rev.E, 78(1).

Newman, M. E. J., Barabási, A.-L., & Watts, D. J. (2006). The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press.

Nguyen, N. P., Dinh, T. N., Ying Xuan, & Thai, M. T. (2011). Adaptive algorithms for detecting community structure in dynamic social networks. INFOCOM, 2011 Proceedings IEEE, 2282-2290. doi:10.1109/INFCOM.2011.5935045

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043), 814-818.

Wasserman, S., Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.

Literature

Göhnert, T., Harrer, A., Hecking, T., & Hoppe, H. U. (2013, August). A workbench to construct and re-use network analysis workflows: concept, implementation, and example case. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1464-1466). ACM.

Halatchliyski, I., Hecking, T., Göhnert, T., & Hoppe, H. U. (2014). Analyzing the path of ideas and activity of contributors in an open learning community. Journal of Learning Analytics, JLA, 1(2), 72-93.

Hecking, T., Göhnert, T., Zeini, S., & Hoppe, U. (2013). Task and time aware community detection in dynamically evolving social networks. *Procedia Computer Science*, *18*, 2066-2075.

Hecking, T., Ziebarth, S., & Hoppe, H. U. (2014). Analysis of dynamic resource access patterns in a blended learning course. Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, Indianapolis, Indiana. 173-182.

Zeini, S., Göhnert, T., Lothar Krempel, L., & Hoppe, H.U. (2012). The impact of measurement time on subgroup detection in online communities. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). ACM.