Collaboration Spotting: Big Data Visual analytics

"The analysis of large graphs plays a prominent role in various fields of research and is relevant in many important application areas. Effective visual analysis of graphs requires appropriate visual presentations in combination with respective user interaction facilities and algorithmic graph analysis methods." **[Landesberger].**

Overall presentation

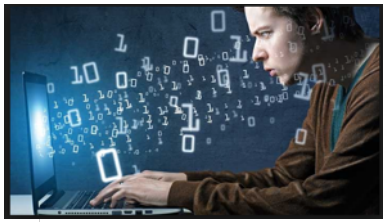
A. Agocs, D. Dardanis, R. Forster, M. Gazzari, J.-M. Le Goff, X. Ouvrard

CERN

Background

- **Collaboration Spotting (CS) is a graph-based interactive visualisation tool for multi-dimensional data networks**
  It aims at evolving towards a for visual analytics of Big Data.
- CS is particularly efficient in performing visual queries on complex and large multi-dimensional data networks
- Data Networks are stored in Neo4j Graph Databases
- CS intends to maximize human visual perception of the content of multi-dimensional data networks
- The current implementation of CS addresses
  - Publications/Patents (Technology monitoring via semantic searches)
  - LHCb process data
  - CERN procurement data

Big Data Analytics Cycle (Today)



Software developer

Domain expert

Forage for Data

Explore

Visualize

Cycle is managed by Data Scientist

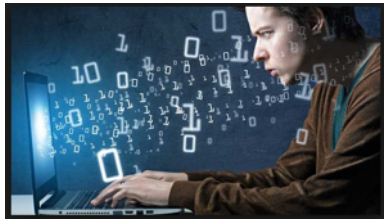
Task

Act

Develop Insights

Source: Tableau Software

Domain expert

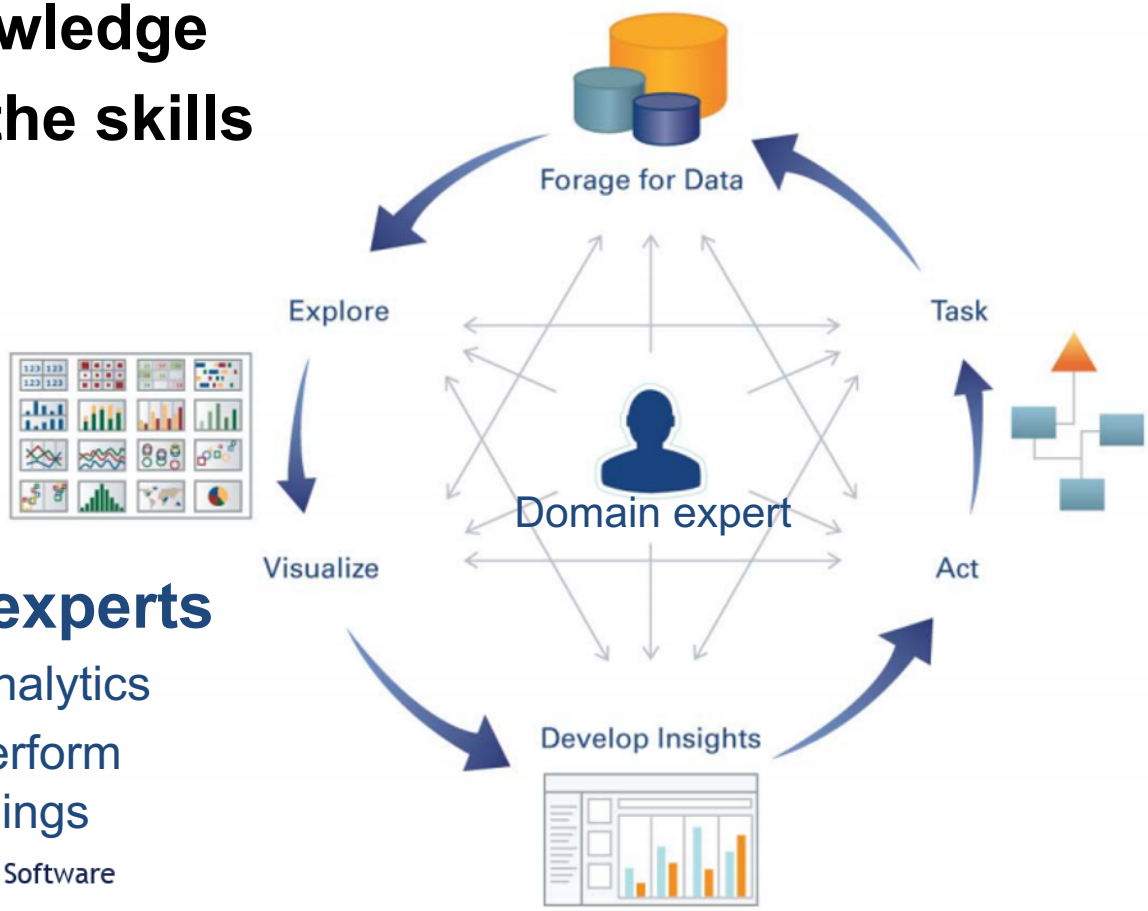


Software developer

Source: JIOX: Intelligence Tradecraft & Analysis

VISION➔ Expert at the centre of the cycle

- **Experts have the knowledge**
- **Data scientists have the skills**



- **➔Bring analytics to experts**
  - "Understand" results of analytics
  - "Instruct" computers to perform analytics according to findings

Source: Tableau Software

Big Data is organised in networks

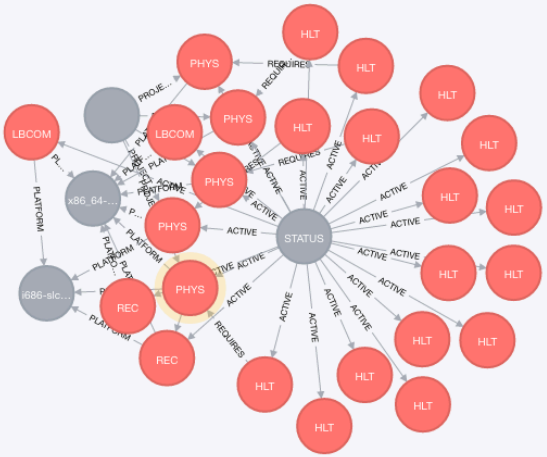
Big Data is distributed

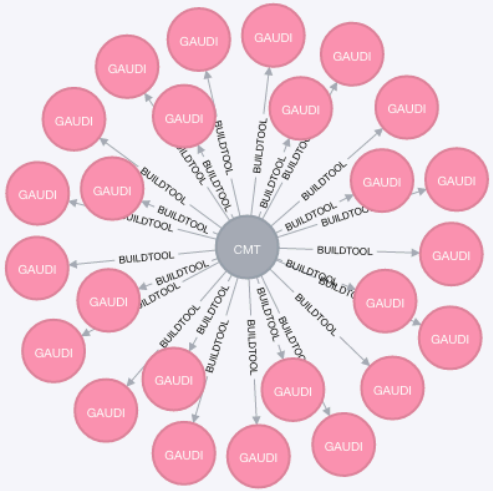
- Document systems with data and metadata in Database
- Database tables with metadata in schema

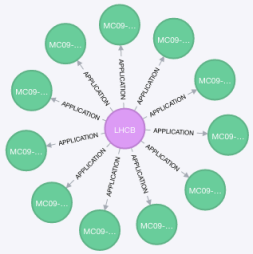
Big Data is strongly interconnected

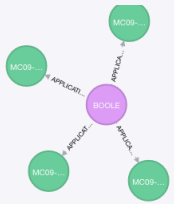
- Networks are **not always materialised** due to the distributed nature of data sources
- Ex: Publications and patents metadata

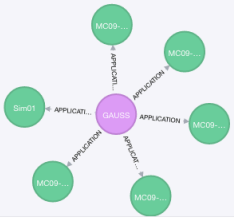
Networks in LHCb Neo4J DB & related schema for dependency data

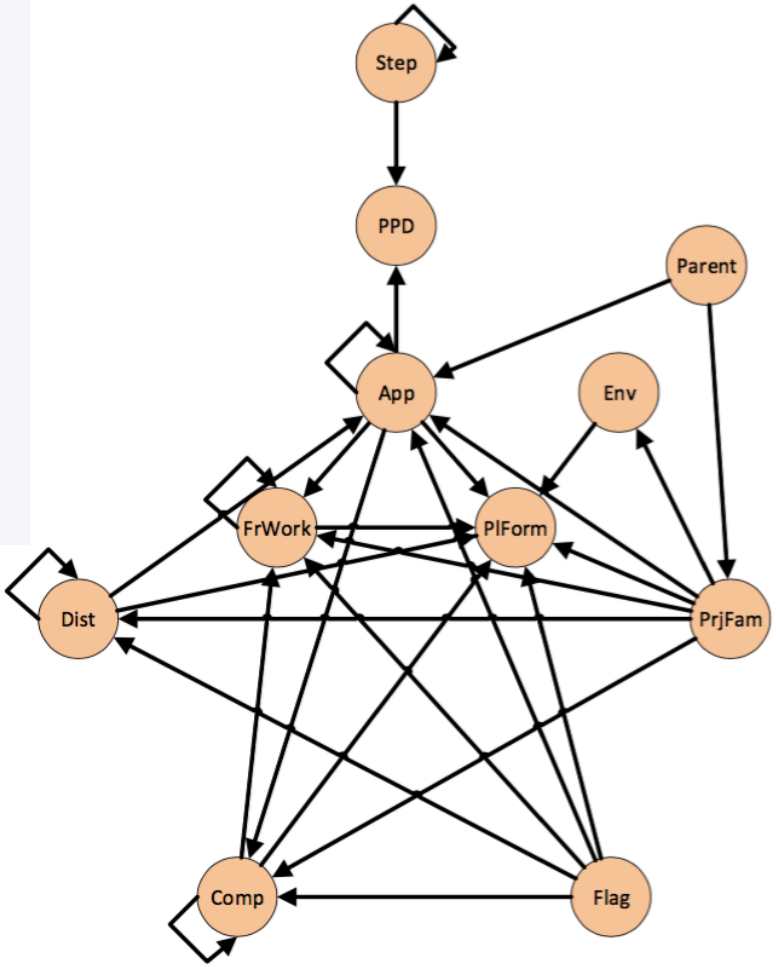












Label: Network dimension
Reachability graph: Graph of connected labels (Schema)

Graph visualisation features

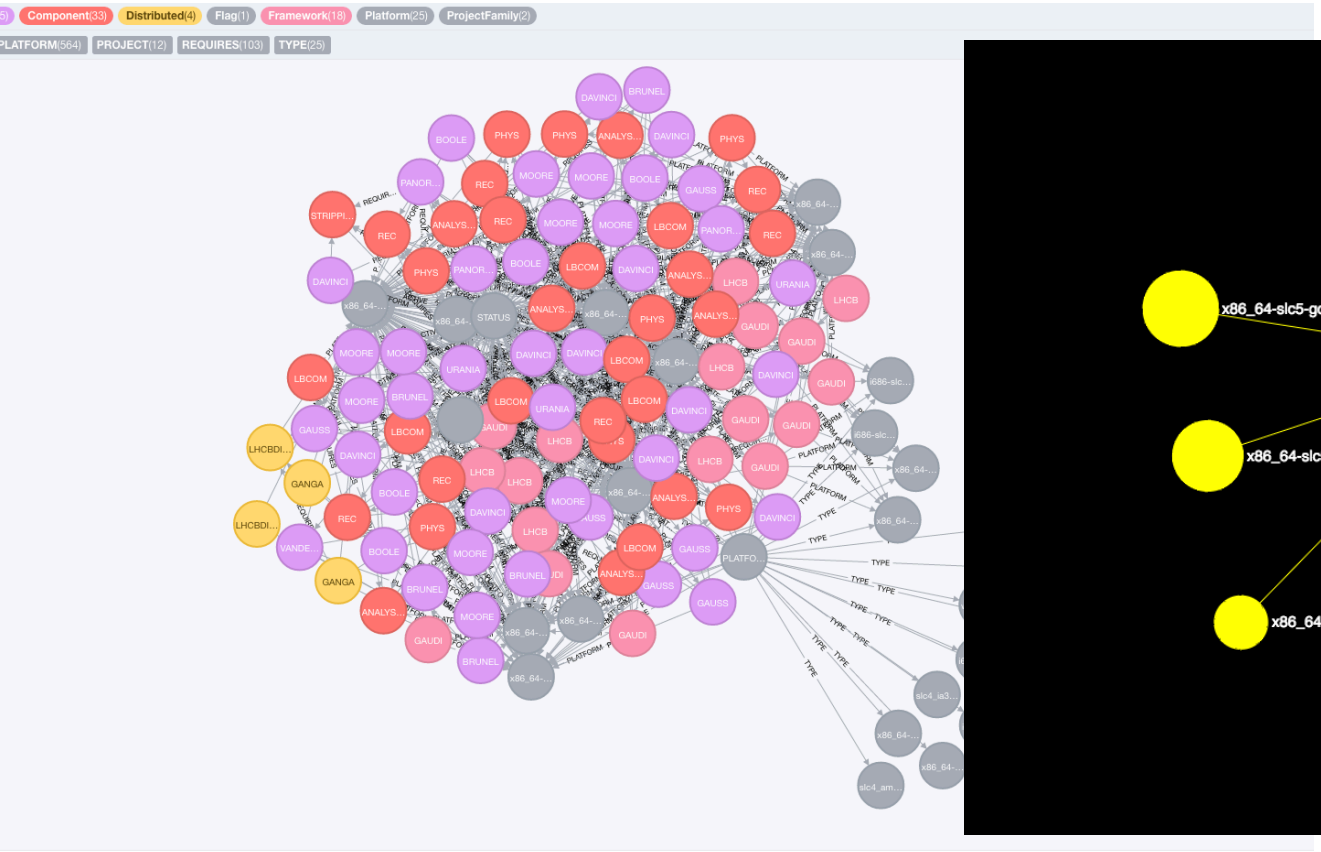
Maximizing human understanding

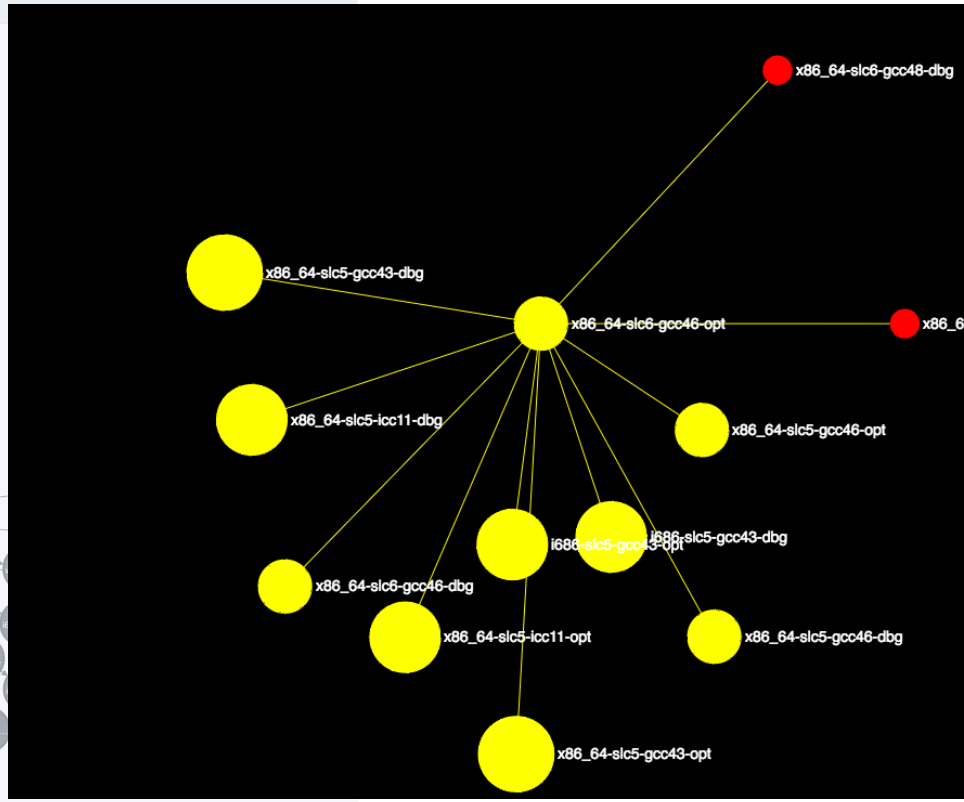
- Selecting network dimensions
- Traversing network dimensions
- Graphical queries
- Time/Frequency evolution

Enhancing reasoning

- Viewing multiple data sources
- Looking for collaborations
- Sorting data
- Contextual visualisation & analytics

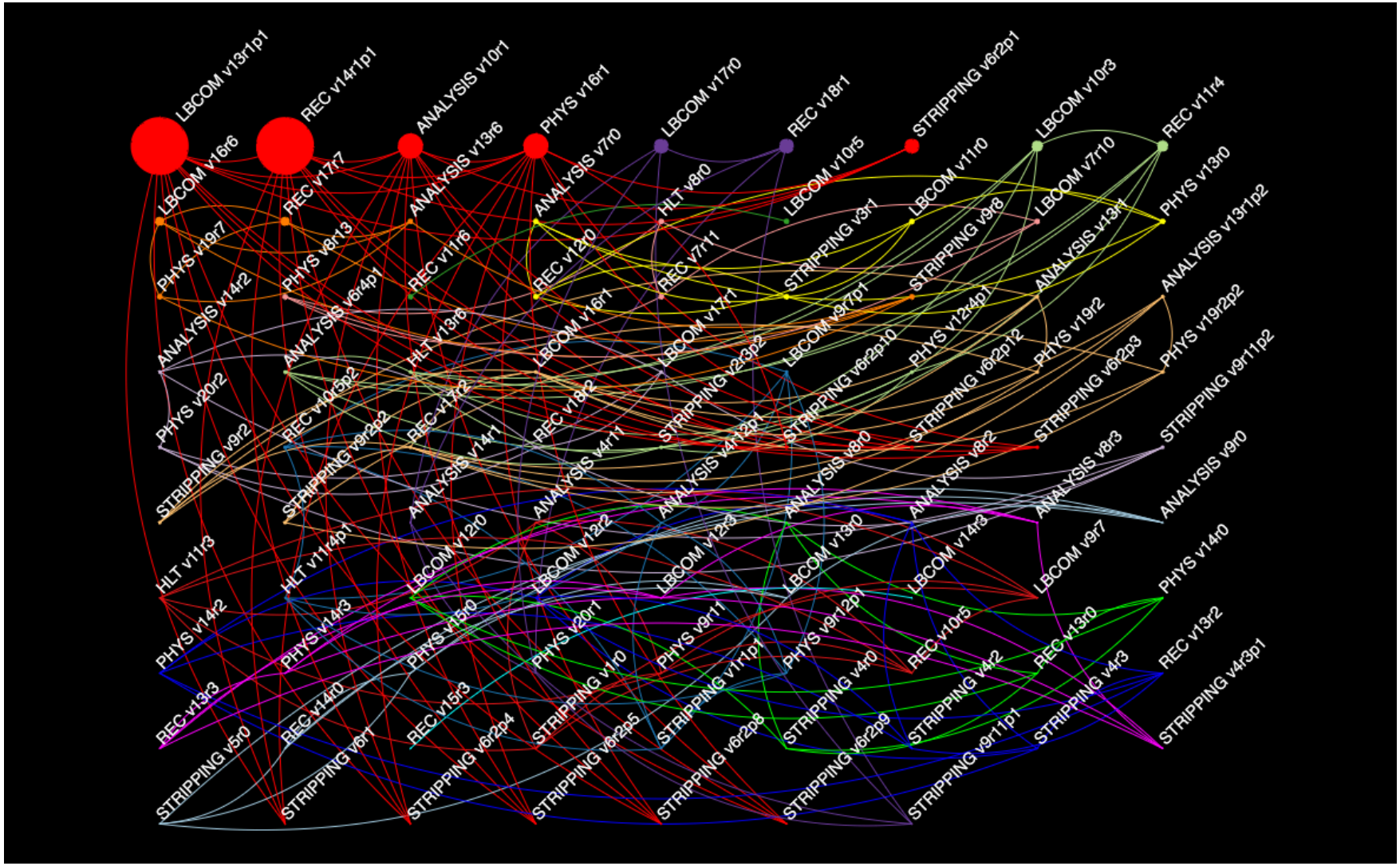
Navigation with CS eases the visual perception of the database content



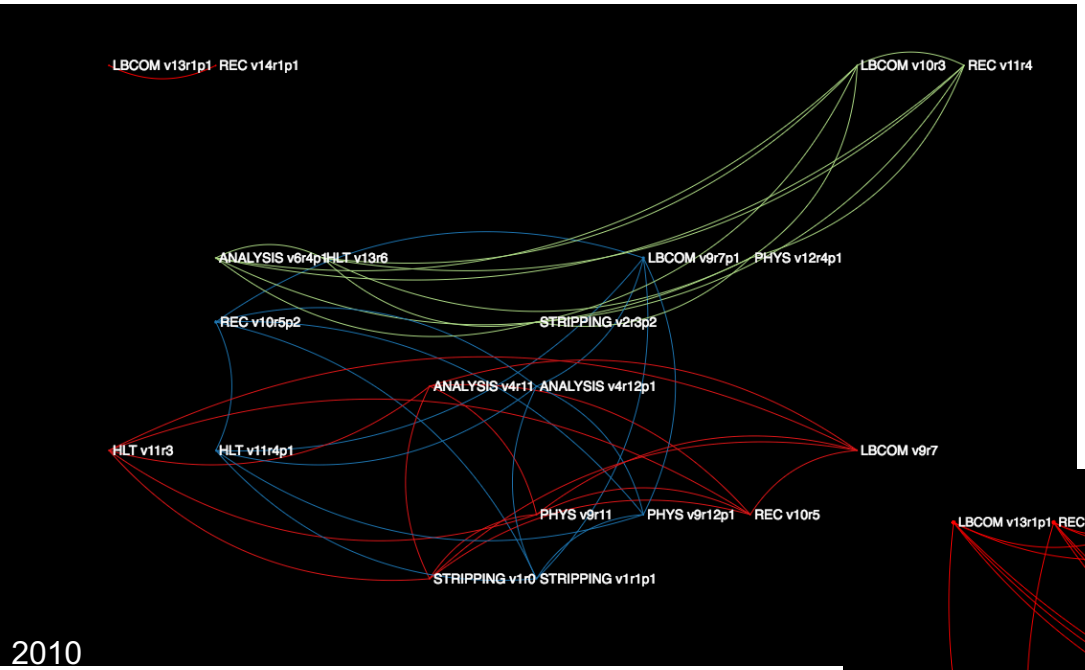


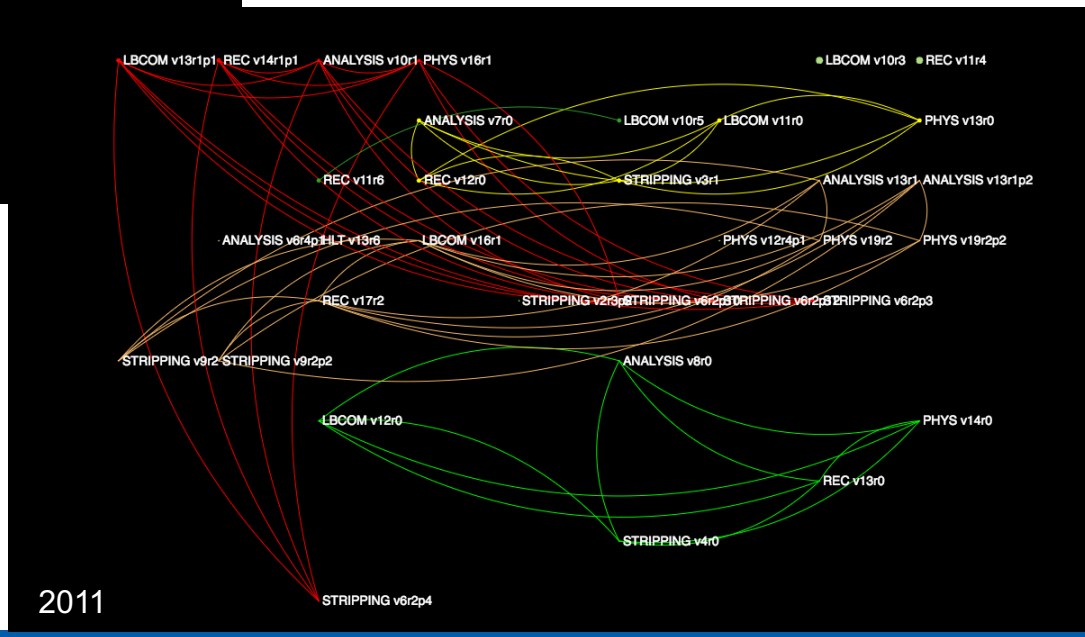
EX: Vertex x86_64-slot-goc46-opt in Neo4j, ☺→ Same in CS

Sorting is particularly easy with CS



Using the timeline

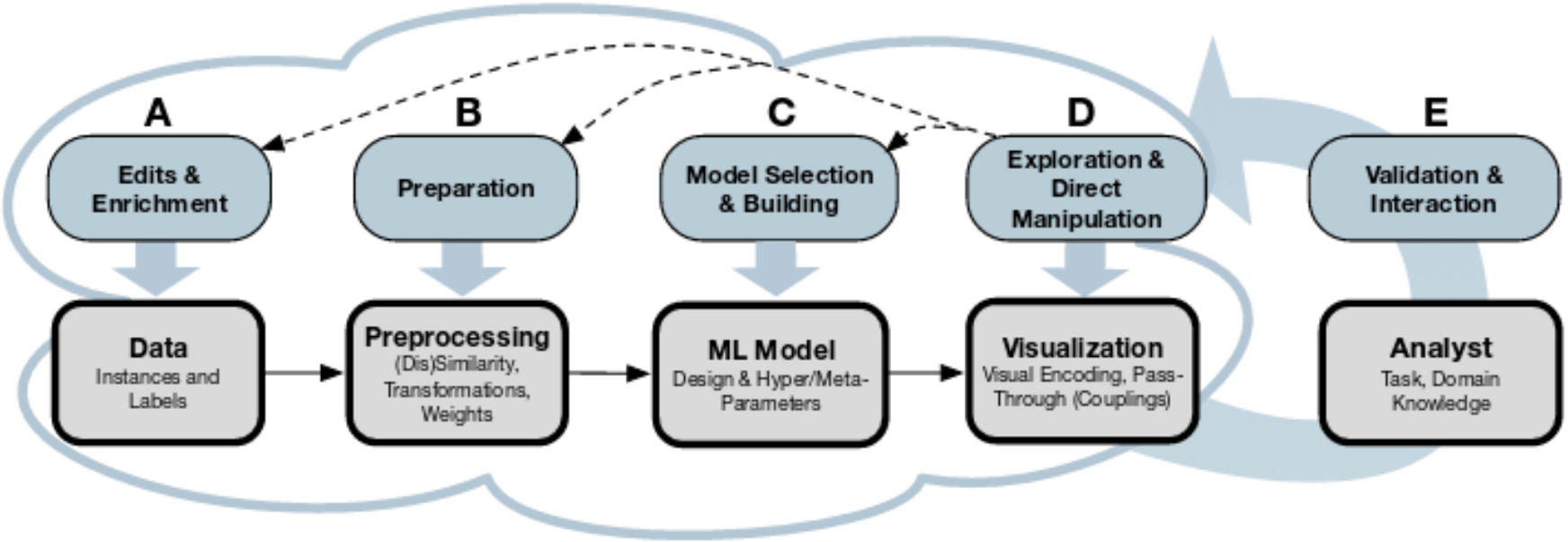




A single platform for visual analytics of multi-dimensional data networks

CS Concepts

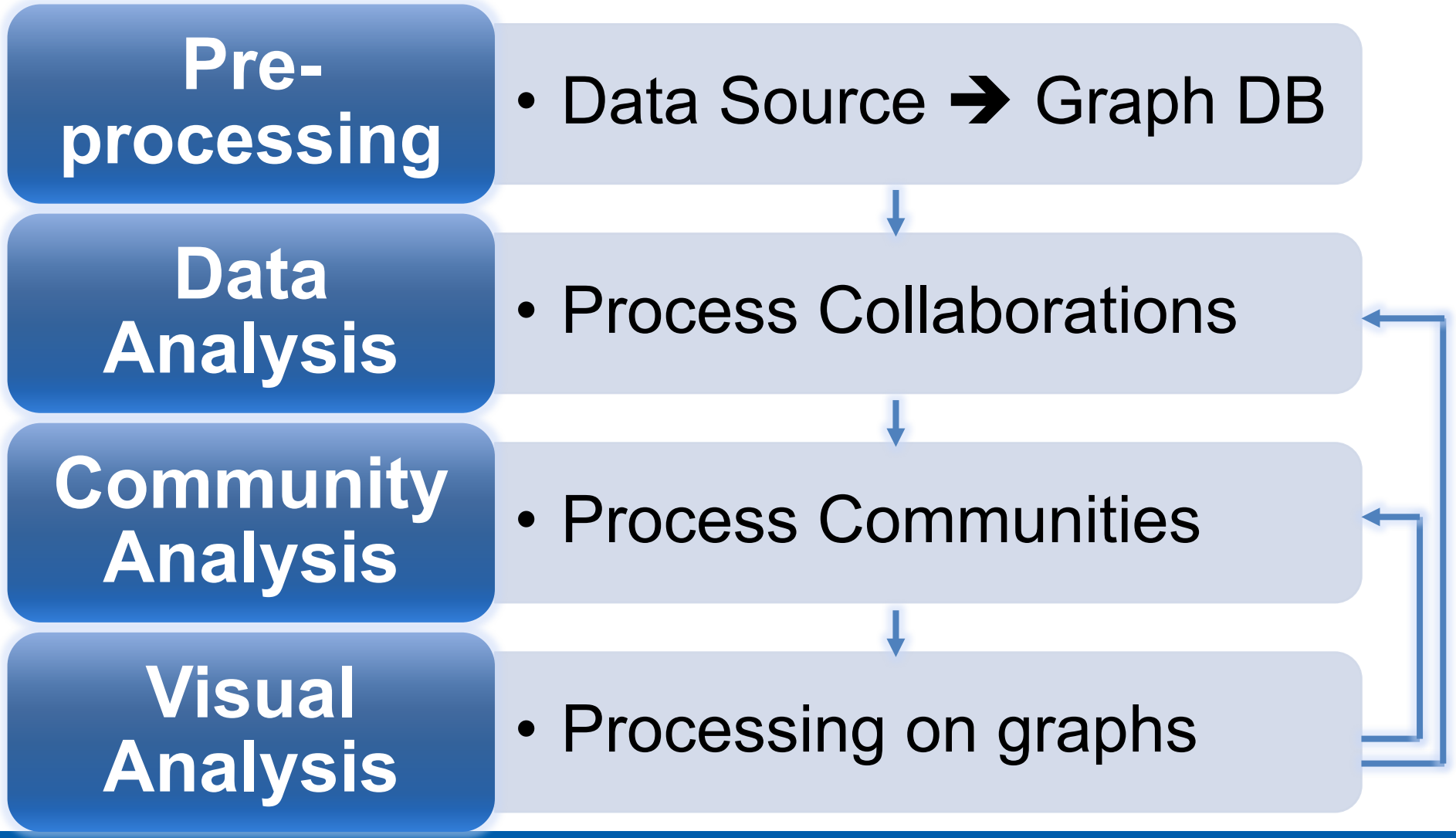
Collaboration Spotting Framework



The project follows the proposed conceptual framework of D. Sacha et al*.

*Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges
Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, Daniel Keim

CS analytics sequence



Vocabulary

Data

- Any set of labelled vertices and relationships in GDB
- A data instance is a labelled vertex or relationship in GDB

Visualisation Data

- Any set of labelled vertices and relationship in GDB
- A visual data instance is a labelled vertex or relationship in GDB

dimension (in data network)

- A label of a vertex or a relationship

Collaborations

- Results of the analysis of data instances
- A collaboration is a collection of visualisation data instances meeting a criteria of the data dimensions
- A collaboration corresponds to one and only one data instance
- There is a set of collections per visual dimension

Community

- Collection of visualisation data instances meeting a criteria of the visualisation dimensions

Pre-processing

Analysis of data source structure and content

- Ex: RDBMS: process schema
- Ex: Semi-structured data: Process tags
- Ex: Graph DB: Vertices and Relationships

Reachability graph

- Schema of graph DB describing the content of the subset of data source
- Schema of the multi-dimension data network resulting from the merging of various data sources

Pre-processing: Data source ➔ Multi-dimensional data network

Select Page | Save to EndNote online | Add to Marked List

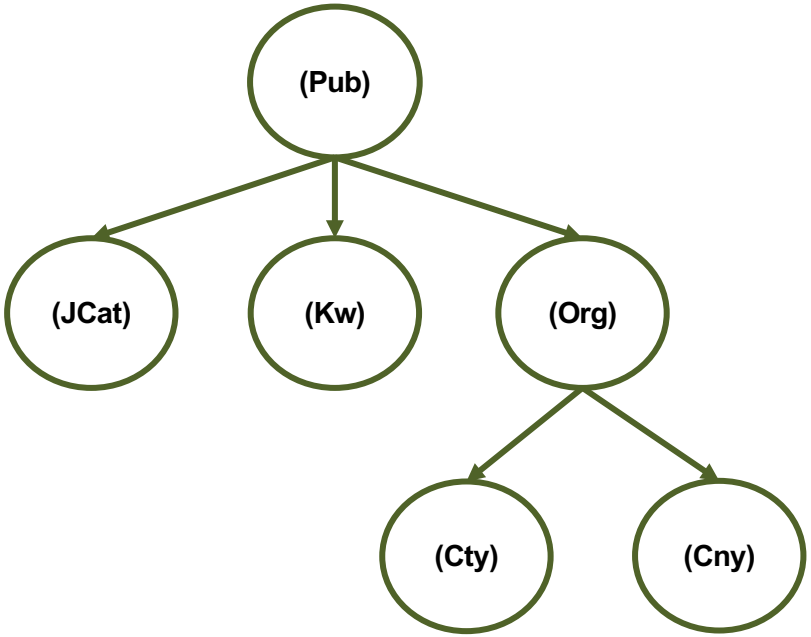
1. **Composition of oxygen precipitates in Czochralski silicon wafers investigated by STEM with EDX/EELS and FTIR spectroscopy**
   By: Kot, Dawid; Kissinger, Gudrun; Schubert, Markus Andreas; et al.
   PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 9 Issue: 7 Pages: 405-409 Published: JUL 2015
   Full Text from Publisher | View Abstract

2. **Correlation between Copper Precipitation and Grown-In Oxygen Precipitates in 300 mm Czochralski Silicon Wafer**
   By: Dong, P.; Ma, X. Y.; Yang, D.
   ACTA PHYSICA POLONICA A Volume: 125 Issue: 4 Pages: 972-975 Published: APR 2014
   Full Text from Publisher | View Abstract

3. **Morphology of Oxygen Precipitates in RTA Pre-Treated Czochralski Silicon Wafers Investigated by FTIR Spectroscopy and STEM**
   By: Kot, D.; Kissinger, G.; Schubert, M. A.; et al.
   ECS JOURNAL OF SOLID STATE SCIENCE AND TECHNOLOGY Volume: 3 Issue: 11 Pages: P370-P375 Published: 2014
   Full Text from Publisher | View Abstract

4. **Thermal deactivation of lifetime-limiting grown-in point defects in n-type Czochralski silicon wafers**
   By: Rougieux, F. E.; Grant, N. E.; Macdonald, D.
   PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 7 Issue: 9 Pages: 616-618 Published: SEP 2013
   Full Text from Publisher | View Abstract

5. **Phosphorus gettering of iron by screen-printed emitters in monocrystalline Czochralski silicon wafers**
   By: Pletzer, Tobias M.; Suckow, Stephan; Stegemann, Elmar F. R.; et al.
   PROGRESS IN PHOTOVOLTAICS Volume: 21 Issue: 5 Pages: 900-905 Published: AUG 2013

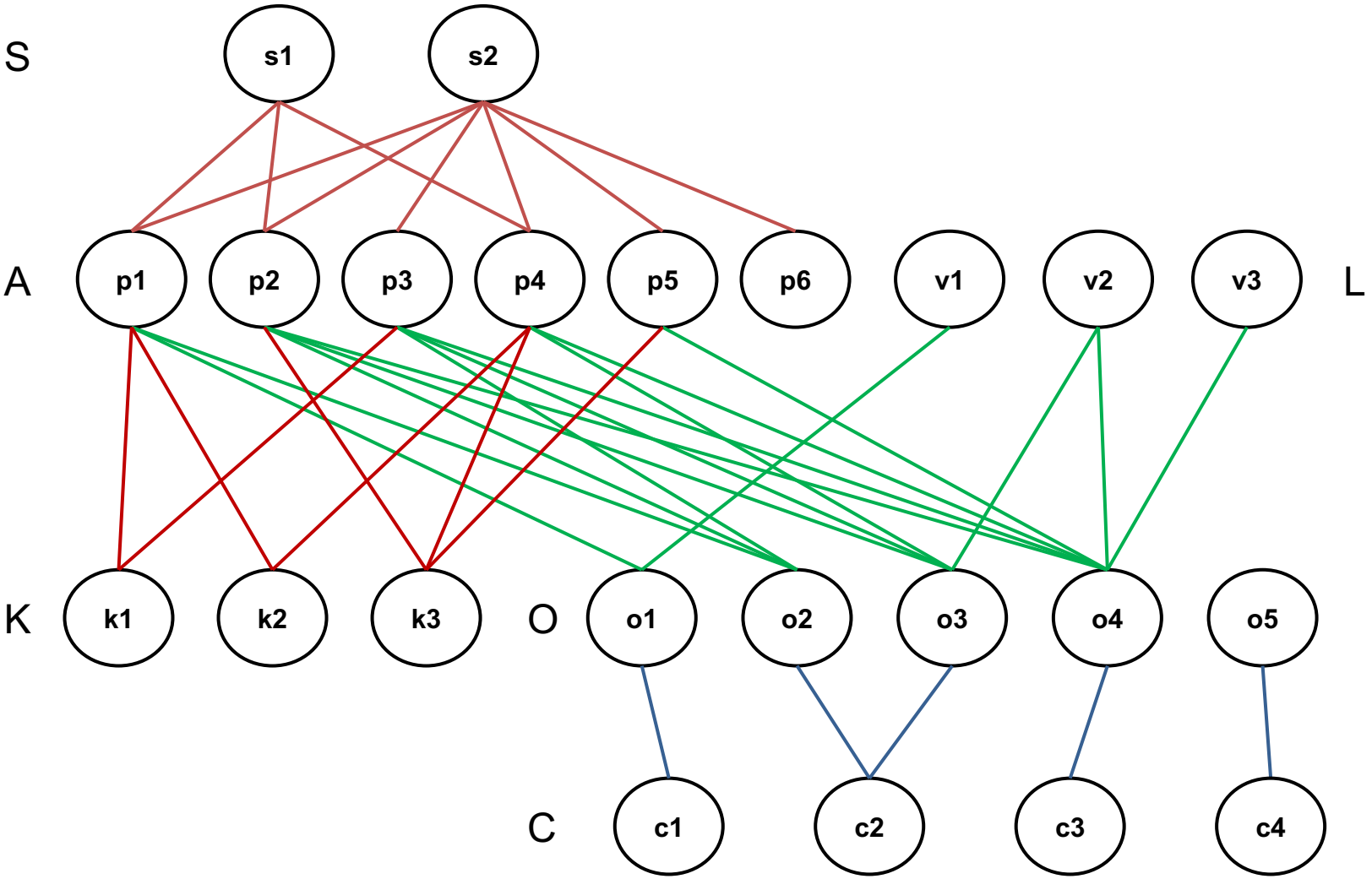


Document metadata

Reachability Graph: Graph of data types

JCat: Journal category, Kw: Keyword, Org: Organisation, Cny: Country, Cty: City

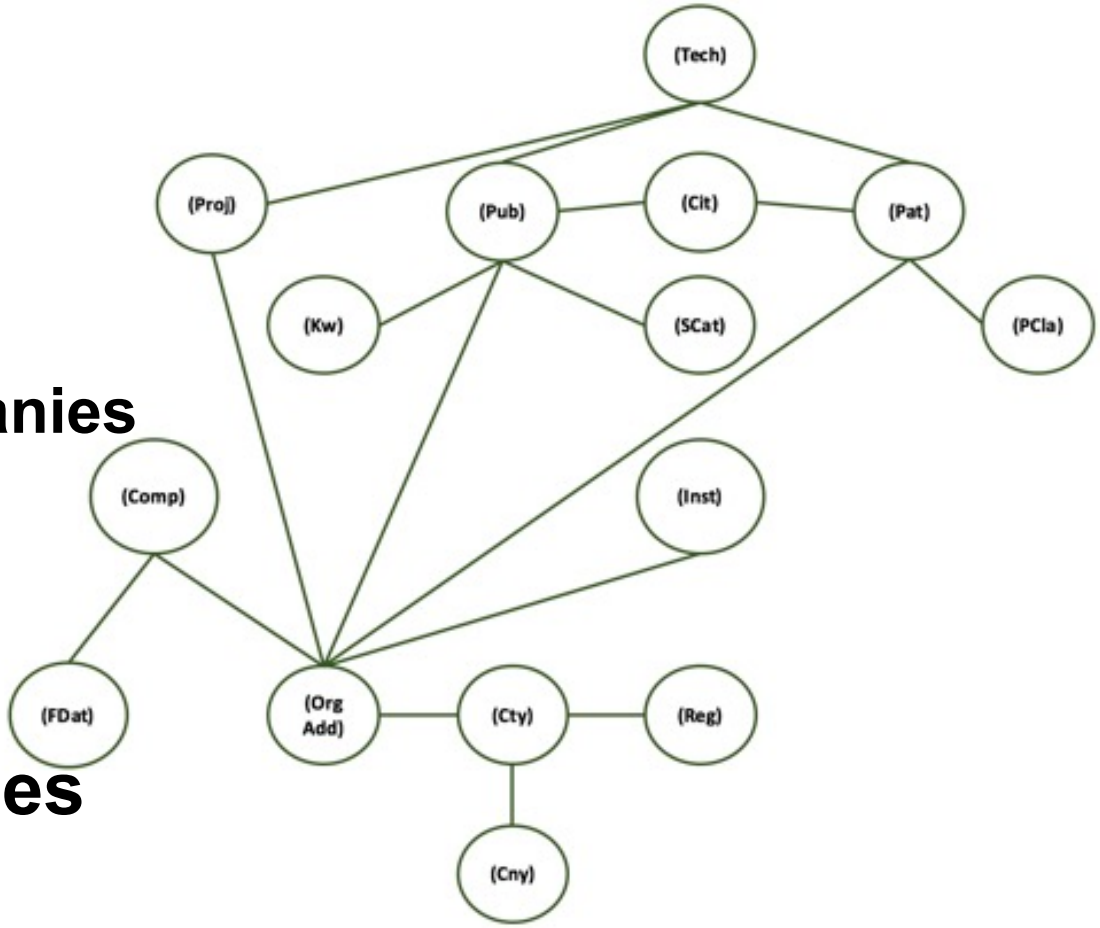
Graph of Metadata / Data



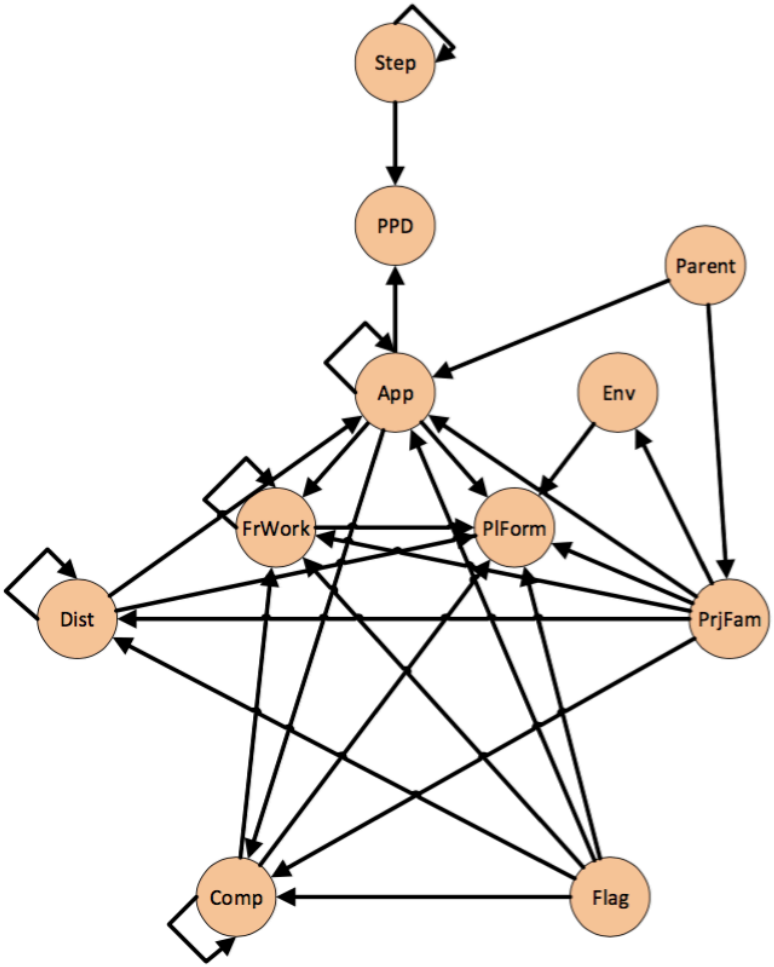
S: Categories, A: Pub/Pat, K: keywords, O: Organisations, C: Countries, L: Cities

Example: Merging data sources

- **Data sources**
  - Publications/**Patents**
    - **Citations**
    - **Institutions/Companies**
  - **EU projects**
  - **Financial data**
  - **Geolocation data**
- **Technology searches**
  (resulting from processing)



Construct reachability graph



Data analysis

Processing specific to a particular dataset

- Publications/Patents ➔ Semantic search

Results added to the Graph DB

- Creation of new labels (if needed)

Compute Collaborations according to criteria

- Ex: Co-publishing/co-patenting
  - Collaborations of organisation, KW, Sub Cat, etc. for each Pub/Pat
- Ex: Synonyms
  - Collaborations of KW.

Community Analysis

Build communities from collaborations according to criteria

- Communities = how collaborations are organised and interconnected
- Results: Connected Components as a partition of the set of vertices
- Ex: Pub: Louvain ➔ Organisations publishing more often together
- Ex: Tech: Louvain ➔ Technologies having more papers in common

Labelling of communities according to criteria

- Ex: Pub: Community = Organisations with common pub/pat
- Ex: Tech: semantic search ➔ Community = technologies corresponding to pub/pat having common terms

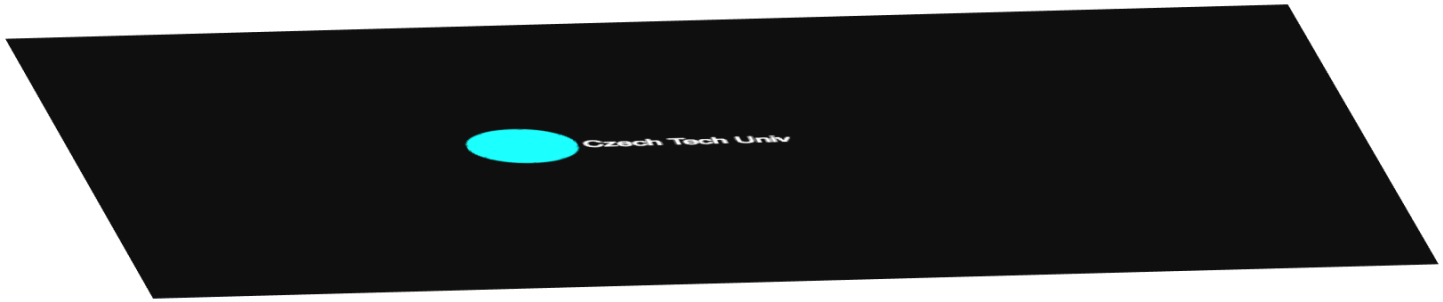
Build compound graph information

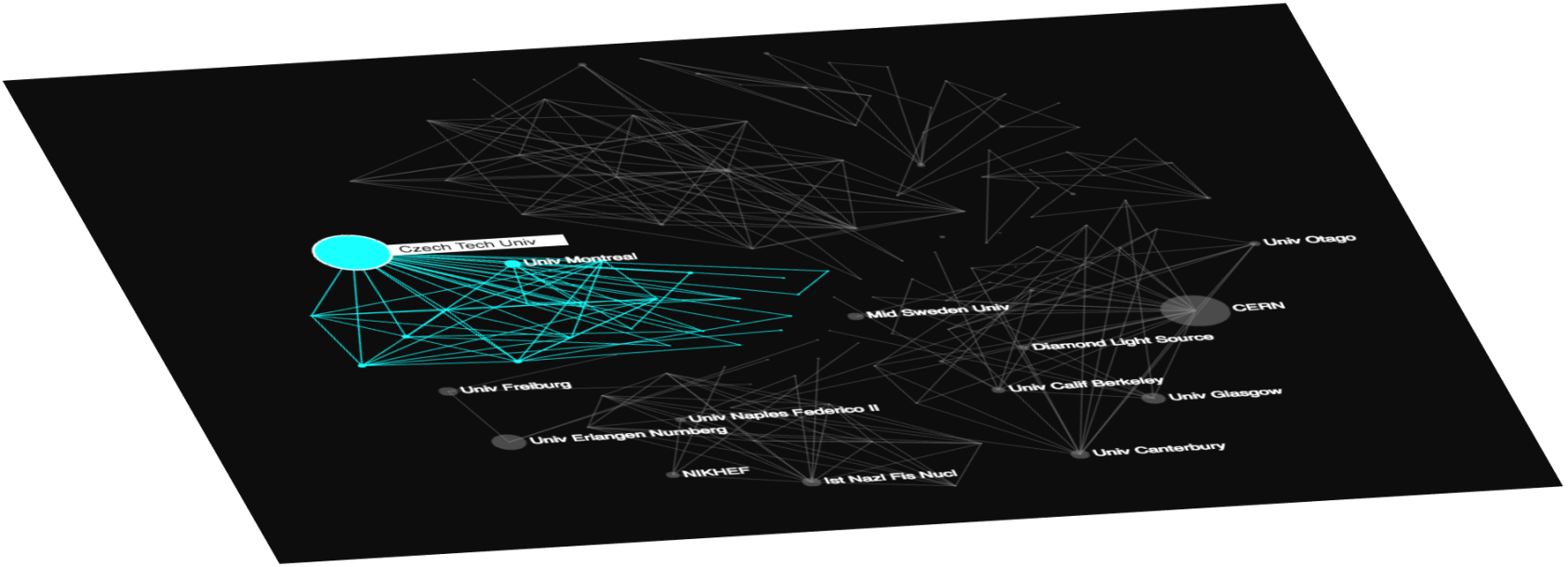
- All info on vertices/edges and collaborations

Compound graph(view)

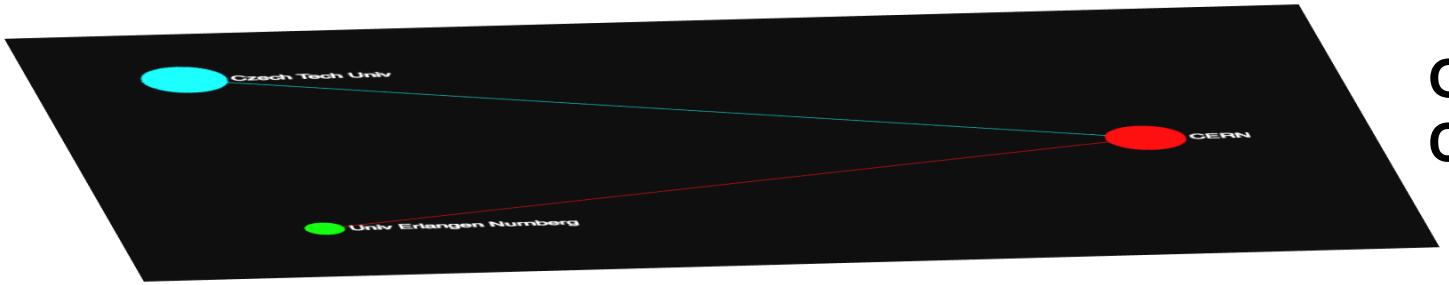
- Combination of a tree and a graph
  - **Tree**: Hierarchy = Vertex → Cluster mapping after applying Community Analysis
  - **Layers**:
    - Vertex layer: Multi-dimensional collaboration layout
      - Containing all vertices and edges from collaborations according to a selected view
    - Cluster layer: Community layout
      - Containing all communities represented as coloured clusters and cluster interconnections (edges)
    - Root Layer: Connected component layout
      - One vertex hierarchically linked to all the linked clusters per connected component.

Tree: Hierarchy
Vertices to Communities

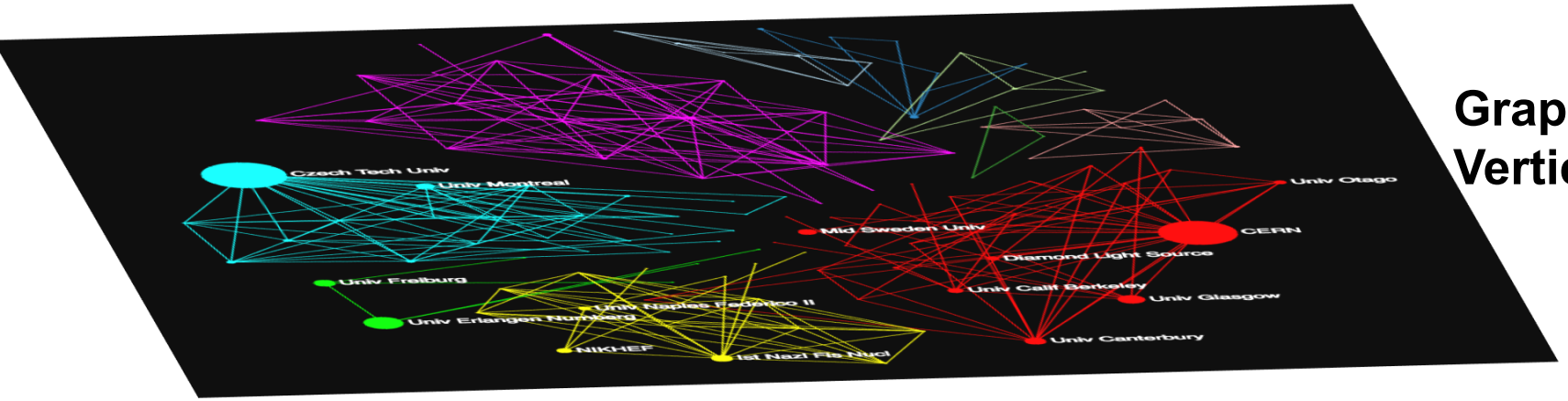




Multi-layer information



Graph layer: Communities



Graph layer: Vertices

Layers in Compound Graph

Vertex layer

- Graph of labelled vertices from selected dimensions for visualisation
- Communities as collections of labelled vertices (colours)
- Collaborations representing data as hyper-edges
- Collaborations representing collection of labelled vertices as vertices

Community layer

- A vertex represents a community, i.e. vertices link together

Connected Component layer (Root)

- A vertex represents a connected component i.e. all the communities linked together in the community layer

Visual representation of data, collaborations, communities

Visual dimensions (Data for visualisation)

- Vertices→ data instances
- Vertex information→ Information on collaborations obtained from the analysis of data

Collaborations

- Visual data → vertices
- Data for analysis → hyper-edges

Communities

- Visual data → Clusters, a colour represents a cluster
- Data for analysis → content of vertices in clusters

Visual Analysis

Default compound graphs

- Vertex layer
  - Colours: communities
  - Sizes: Proportional to |data instances|
  - Ex: Nb of pub/pat

Process graph parameters (colour, size, shape, labels)

- Using data and/or attributes in vertices,
  - Ex: Red for companies and Blue for institutions
  - Ex. CERN procurement: Well balanced vs poorly balanced countries
- Using collaborations resulting from the analysis of visual data
  - Replace vertices with collaborations

Users

- **Data Scientist**
  - Manages Reachability Graph
  - Defines/Specifies Expert's options and r/w access:
  - Maintains/updates CS environment incl. GraphDB (network)
  - Manages/Implements analytics modules
- **Expert**
  - Configures his personal Graph environment
    - Selects views (one visual dimension = one view)
    - Combines views
    - Specifies his analysis options out of the system possibilities
    - Specifies his community analysis options
      - Criteria to compute communities
    - Specifies his visual analysis options for communities
      - Meaning of vertex size, colour, etc.
    - Specifies his graph options

Setting up user visual environment

Reachability Graph ➔ Navigation graph

- Subset of DB schema optimized for navigation purposes

Visual dimensions (user selection)

- Organisations, People, Countries, Software components, activity codes, etc.

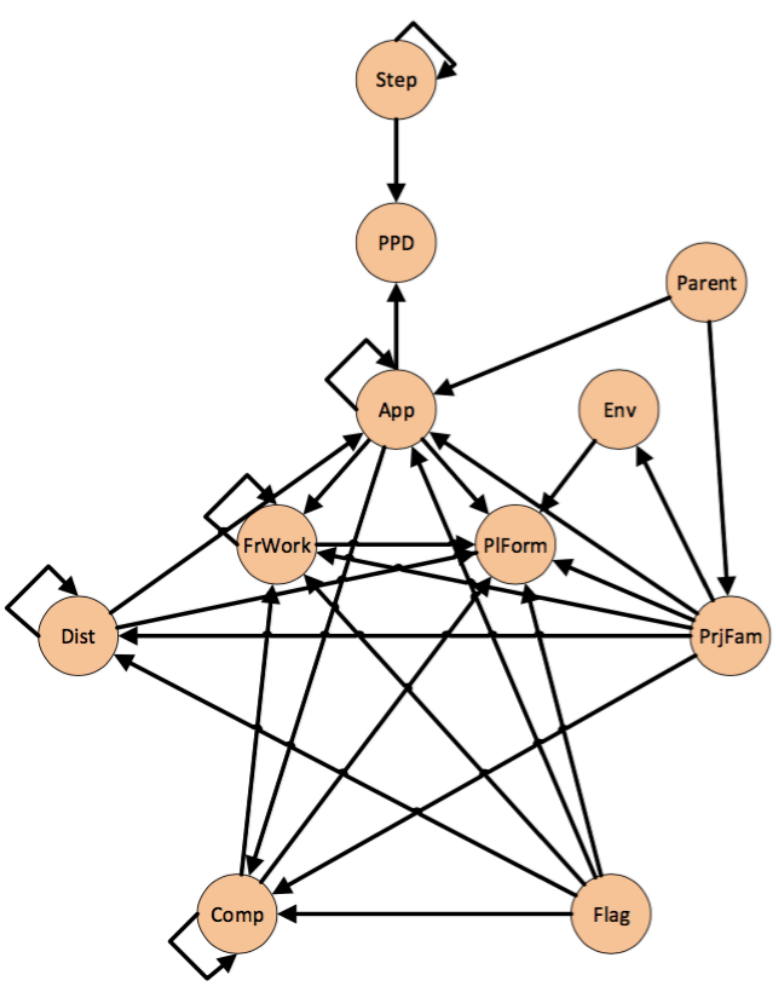
Data Dimensions

- Publications, patents, projects, supplier records, UNICRI specific data records, etc.

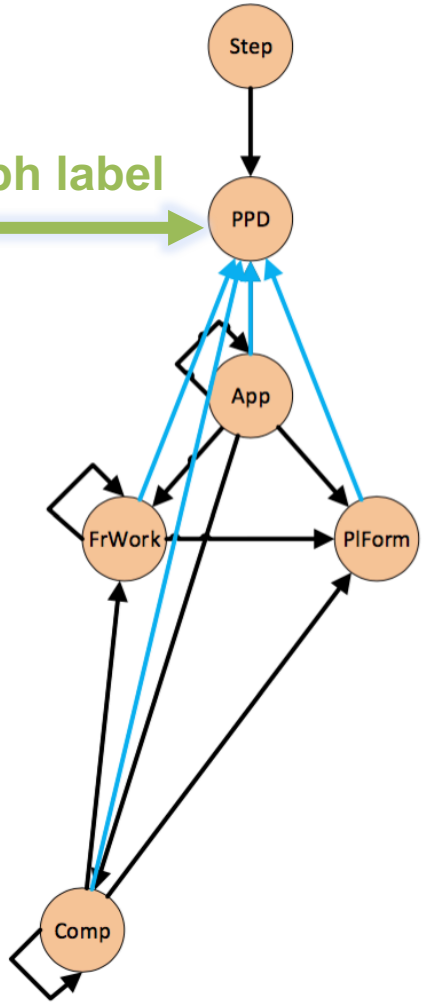
Entry graph (user specified)

- Visual dimension of the front graph
- Technology, Processing Pass Descriptions, Procurement Data

Construct reachability graph



Entry graph label



Reachability graph → Available dimensions for navigation

Blue edges added to support navigations from entry graph

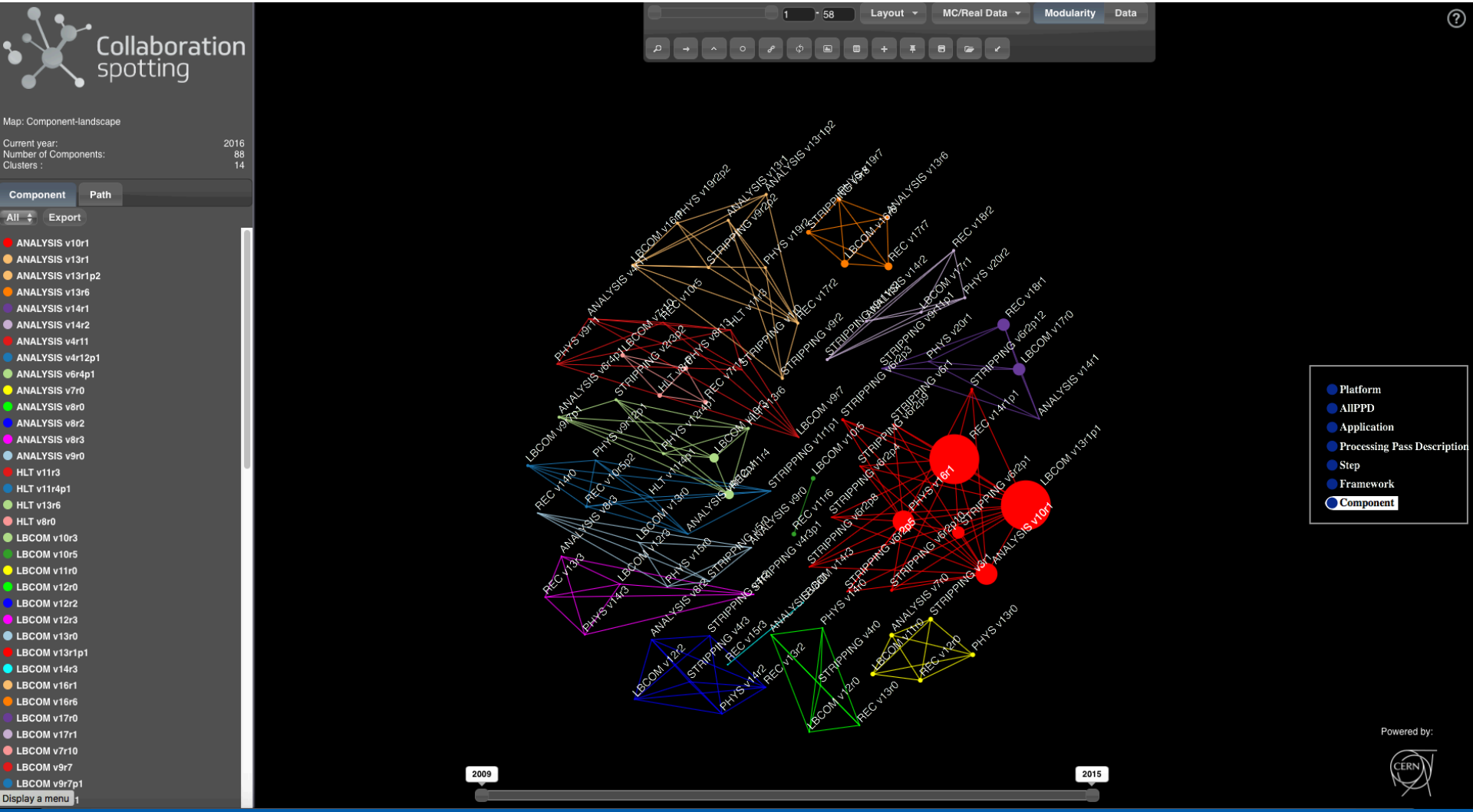
CS supported Graph Visual representations

- Static graph with timeline window
- Node-link using different layout techniques
  - Clique representation (currently available)
    - Force Atlas (currently available)
    - Circular representation
  - Extra node representation (hyper-graph)
    - Force Atlas
    - Circular representation

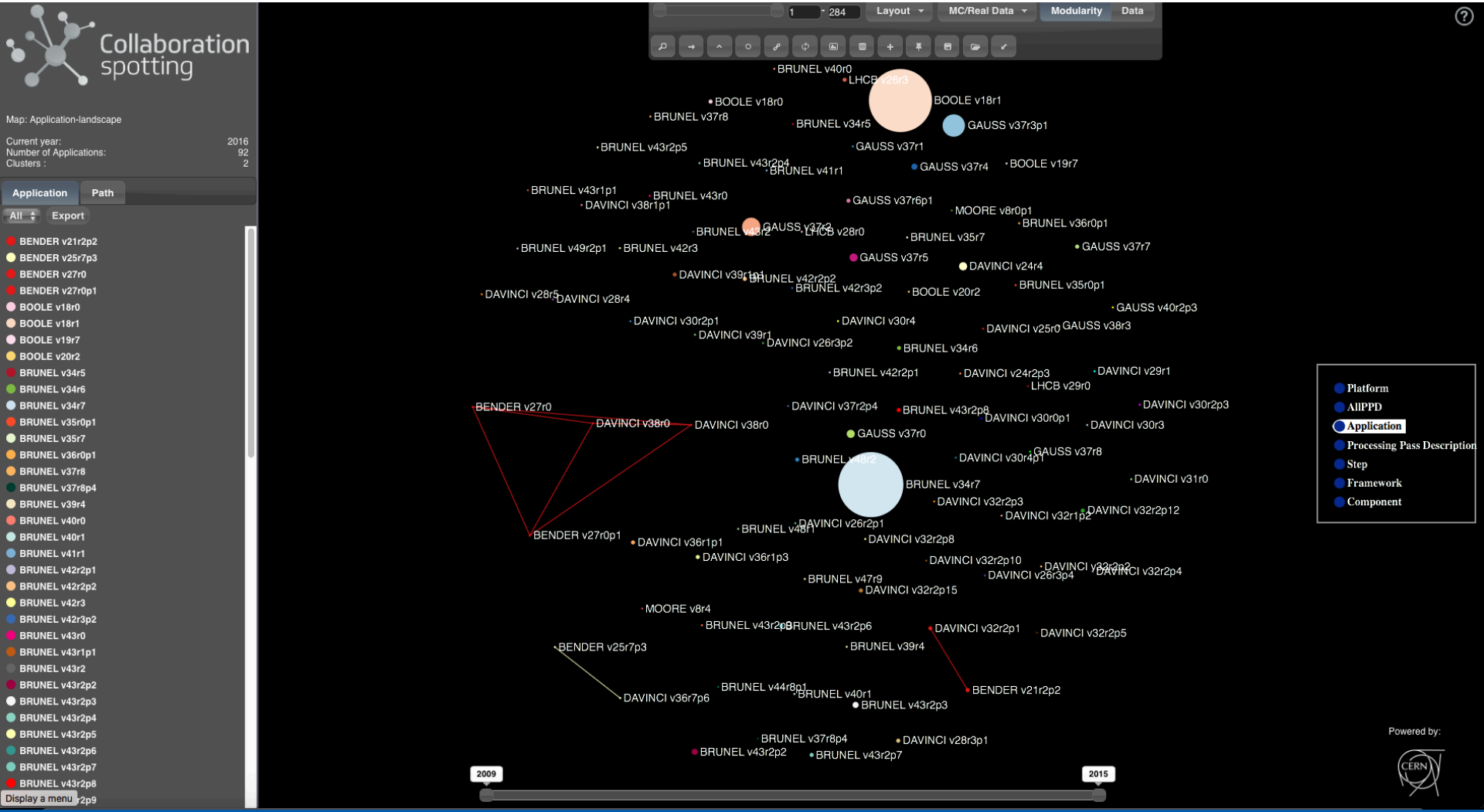
Entry graph in LHCb

1 vertex (all PPD) and **Navigation options** (as defined in the navigation graph)

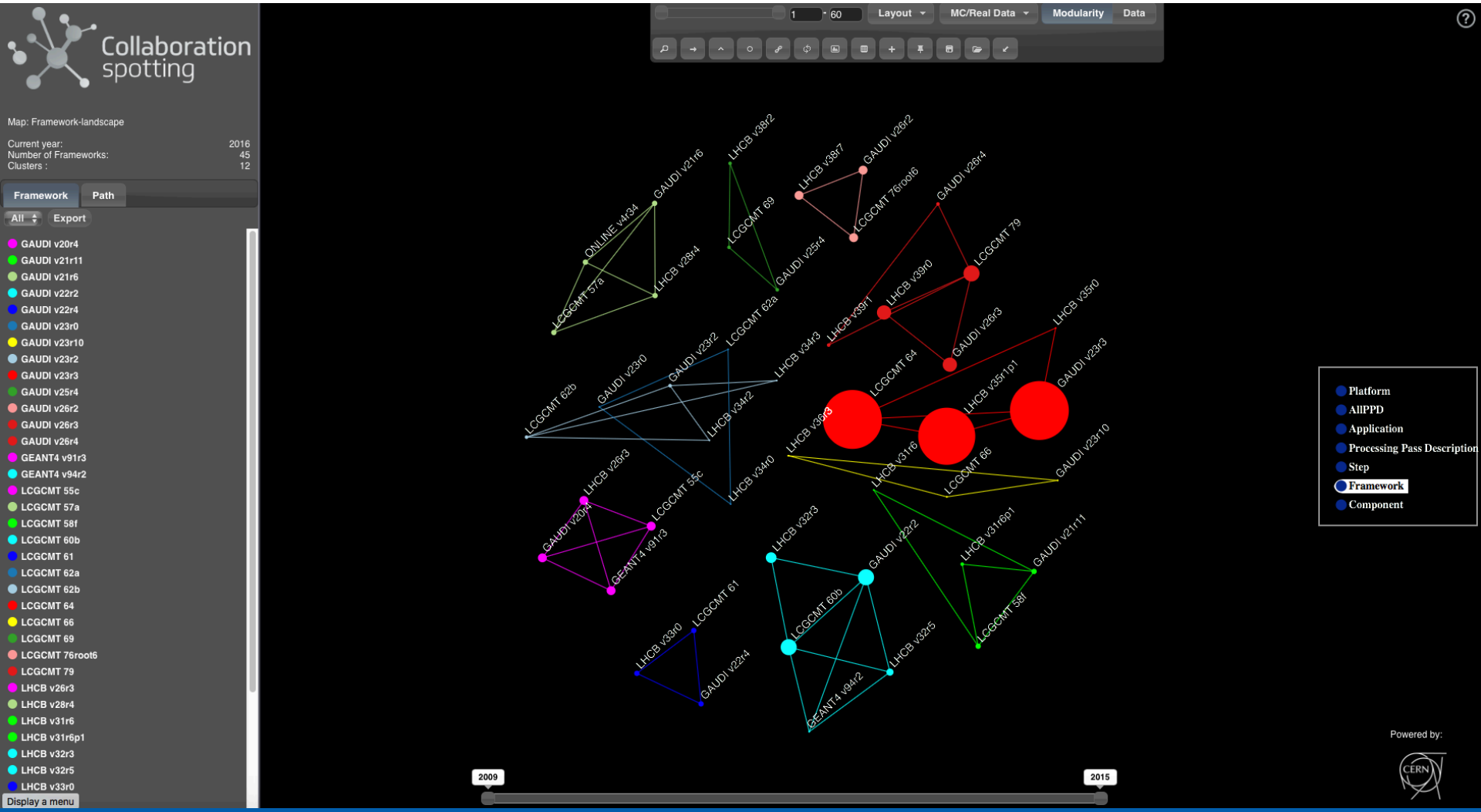
Components


Collaboration spotting
Map: Component-landscape
Current year: 2016
Number of Components: 88
Clusters : 14
Component
Path
All
Export
ANALYSIS v10r1
ANALYSIS v13r1
ANALYSIS v13r1p2
ANALYSIS v13r6
ANALYSIS v14r1
ANALYSIS v14r2
ANALYSIS v4r11
ANALYSIS v4r12p1
ANALYSIS v6r4p1
ANALYSIS v7r0
ANALYSIS v8r0
ANALYSIS v8r2
ANALYSIS v8r3
ANALYSIS v9r0
HLT v11r3
HLT v11r4p1
HLT v13r6
HLT v8r0
LBCOM v10r3
LBCOM v10r5
LBCOM v11r0
LBCOM v12r0
LBCOM v12r2
LBCOM v12r3
LBCOM v13r0
LBCOM v13r1p1
LBCOM v14r3
LBCOM v16r1
LBCOM v16r6
LBCOM v17r0
LBCOM v17r1
LBCOM v7r10
LBCOM v9r7
LBCOM v9r7p1
Display a menu
Layout
MC/Real Data
Modularity
Data
Platform
AllPPD
Application
Processing Pass Description
Step
Framework
Component
2009
2015
Powered by:
CERN

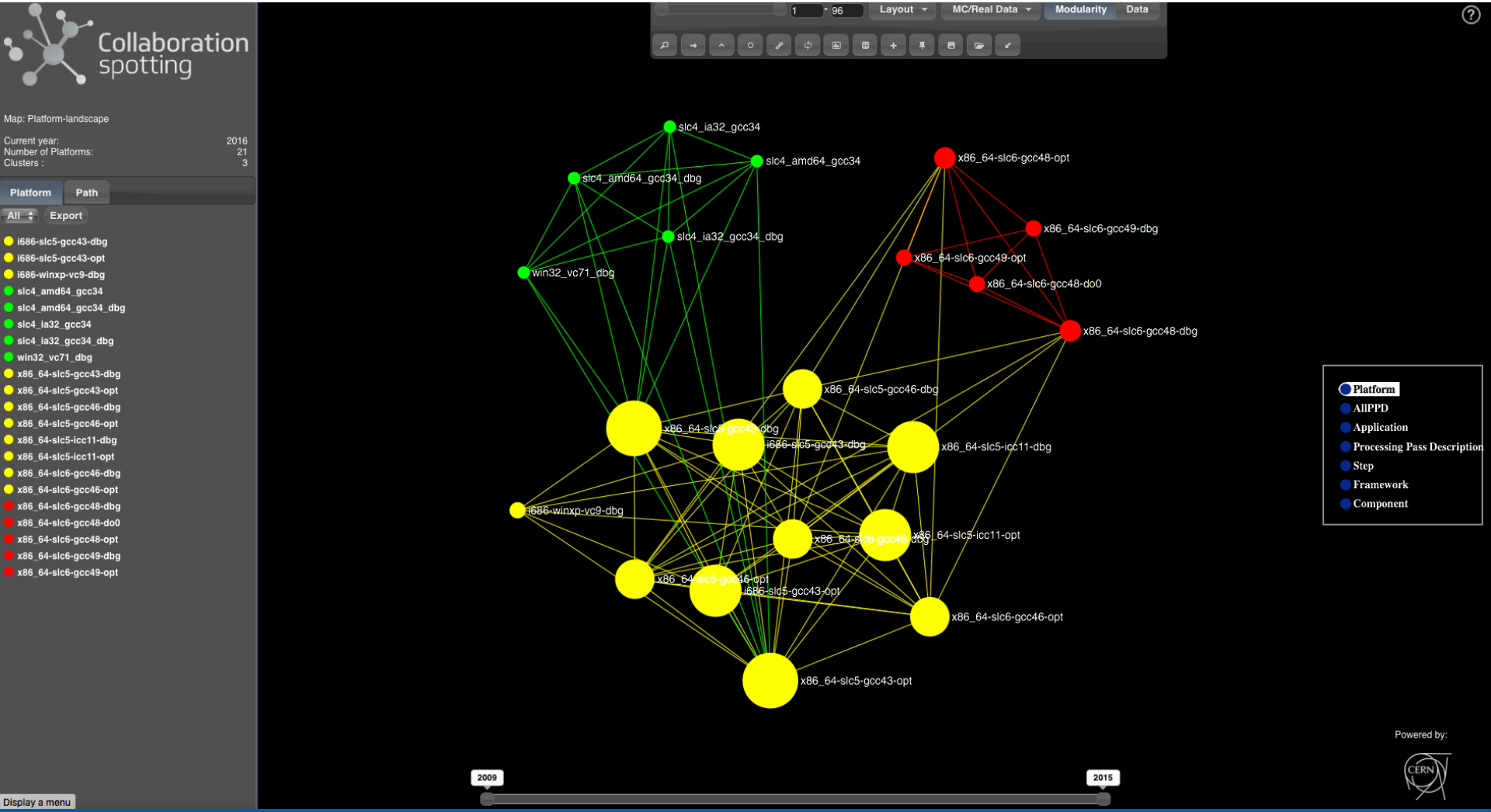
Applications



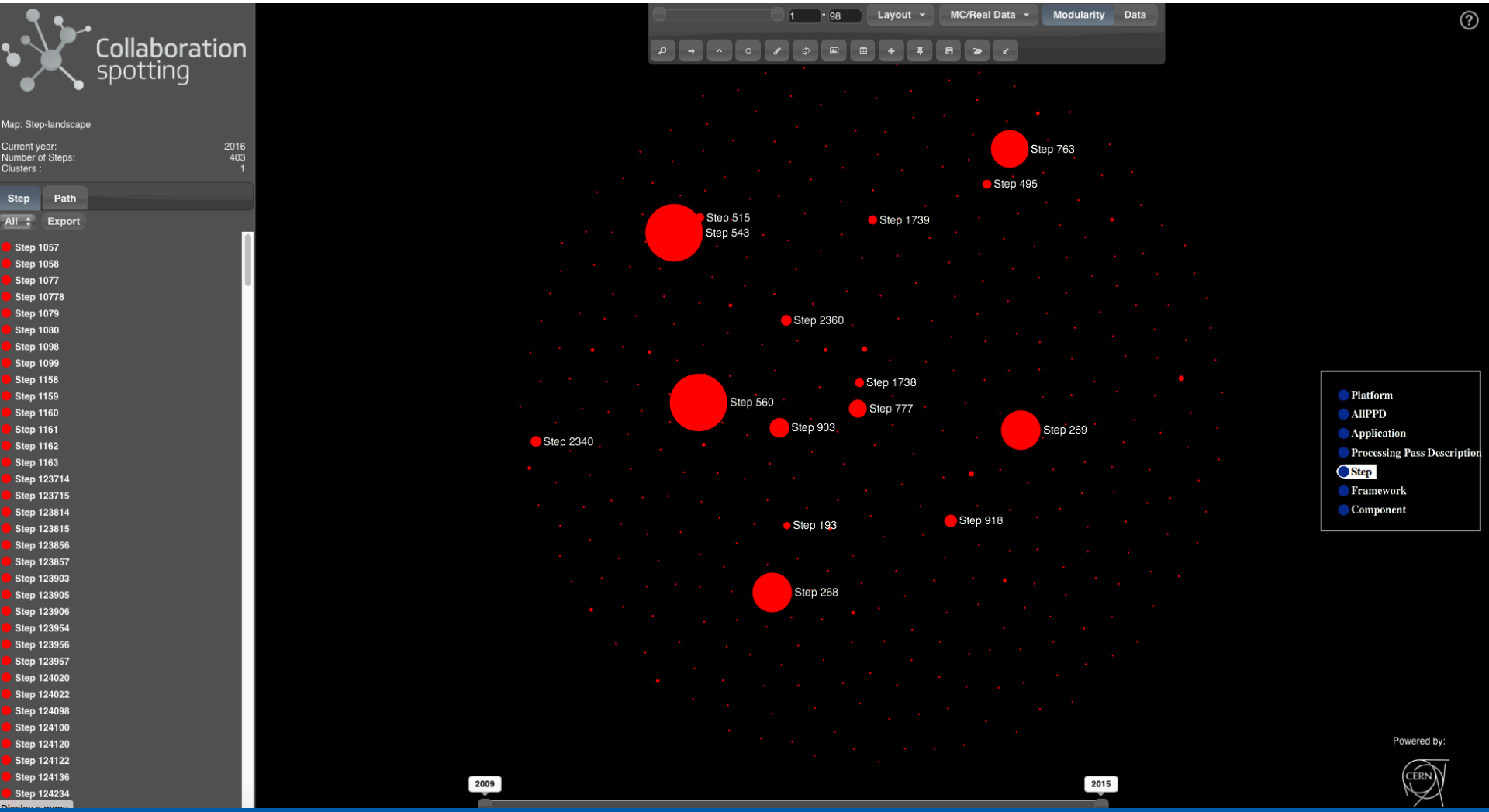
Frameworks


Collaboration spotting
Map: Framework-landscape
Current year: 2016
Number of Frameworks: 45
Clusters : 12
Framework
Path
All
Export
GAUDI v20r4
GAUDI v21r11
GAUDI v21r6
GAUDI v22r2
GAUDI v22r4
GAUDI v23r0
GAUDI v23r10
GAUDI v23r2
GAUDI v23r3
GAUDI v25r4
GAUDI v26r2
GAUDI v26r3
GAUDI v26r4
GEANT4 v91r3
GEANT4 v94r2
LCGCMT 55c
LCGCMT 57a
LCGCMT 58f
LCGCMT 60b
LCGCMT 61
LCGCMT 62a
LCGCMT 62b
LCGCMT 64
LCGCMT 66
LCGCMT 69
LCGCMT 76root6
LCGCMT 79
LHCB v26r3
LHCB v28r4
LHCB v31r6
LHCB v31r6p1
LHCB v32r3
LHCB v32r5
LHCB v33r0
Display a menu
1
60
Layout
MC/Real Data
Modularity
Data
Platform
AllPPD
Application
Processing Pass Description
Step
Framework
Component
Powered by:
2009
2015

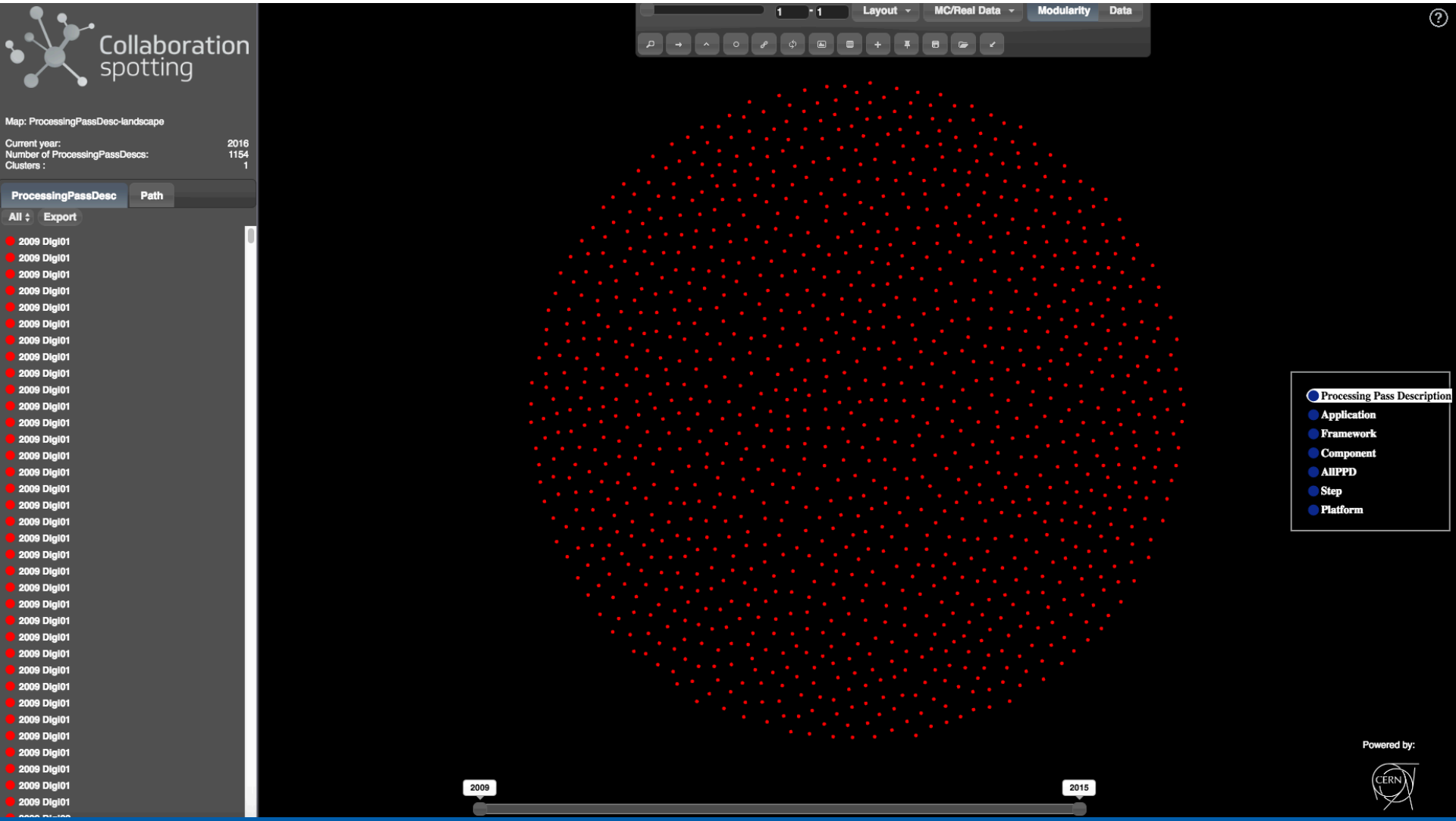
Platforms


Collaboration spotting
Map: Platform-landscape
Current year: 2016
Number of Platforms: 21
Clusters : 3
Platform
Path
All
Export
i686-slc5-gcc43-dbg
i686-slc5-gcc43-opt
i686-winxp-vc9-dbg
slc4_amd64_gcc34
slc4_amd64_gcc34_dbg
slc4_ia32_gcc34
slc4_ia32_gcc34_dbg
win32_vc71_dbg
x86_64-slc5-gcc43-dbg
x86_64-slc5-gcc43-opt
x86_64-slc5-gcc46-dbg
x86_64-slc5-gcc46-opt
x86_64-slc5-icc11-dbg
x86_64-slc5-icc11-opt
x86_64-slc6-gcc46-dbg
x86_64-slc6-gcc46-opt
x86_64-slc6-gcc48-dbg
x86_64-slc6-gcc48-do0
x86_64-slc6-gcc48-opt
x86_64-slc6-gcc49-dbg
x86_64-slc6-gcc49-opt
Display a menu
1
96
Layout
MC/Real Data
Modularity
Data
slc4_ia32_gcc34
slc4_amd64_gcc34
slc4_amd64_gcc34_dbg
slc4_ia32_gcc34_dbg
win32_vc71_dbg
x86_64-slc6-gcc48-opt
x86_64-slc6-gcc49-dbg
x86_64-slc6-gcc49-opt
x86_64-slc6-gcc48-do0
x86_64-slc6-gcc48-dbg
x86_64-slc5-gcc46-dbg
x86_64-slc5-gcc43-dbg
i686-slc5-gcc43-dbg
x86_64-slc5-icc11-dbg
i686-winxp-vc9-dbg
x86_64-slc6-gcc46-dbg
x86_64-slc5-icc11-opt
x86_64-slc5-gcc46-opt
i686-slc5-gcc43-opt
x86_64-slc6-gcc46-opt
x86_64-slc5-gcc43-opt
Platform
AllPPD
Application
Processing Pass Description
Step
Framework
Component
Powered by:
2009
2015

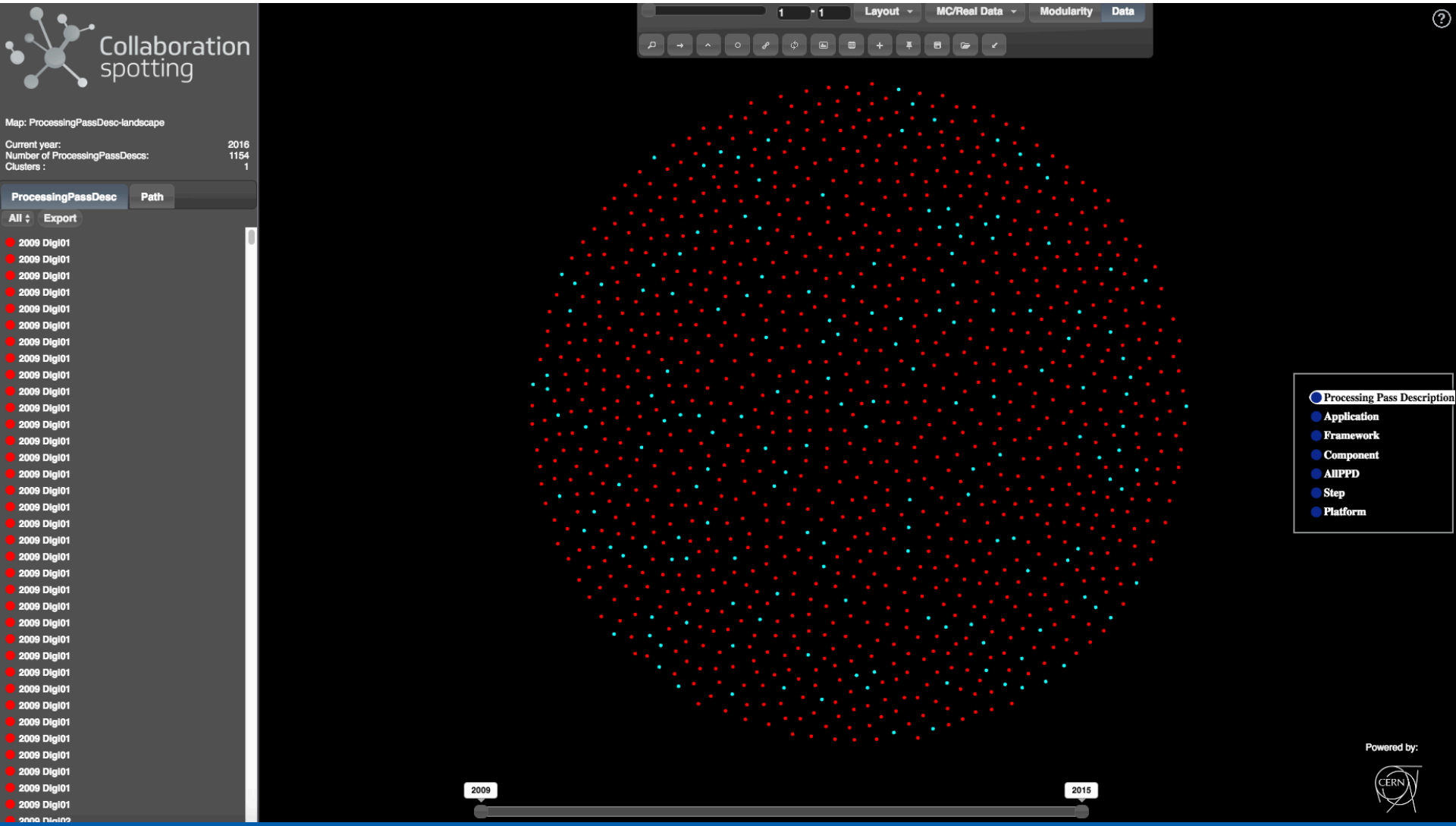
Steps


Collaboration spotting
Map: Step-landscape
Current year: 2016
Number of Steps: 403
Clusters : 1
Step
Path
All
Export
Step 1057
Step 1058
Step 1077
Step 10778
Step 1079
Step 1080
Step 1098
Step 1099
Step 1158
Step 1159
Step 1160
Step 1161
Step 1162
Step 1163
Step 123714
Step 123715
Step 123814
Step 123815
Step 123856
Step 123857
Step 123903
Step 123905
Step 123906
Step 123954
Step 123956
Step 123957
Step 124020
Step 124022
Step 124098
Step 124100
Step 124120
Step 124122
Step 124136
Step 124234
1
98
Layout
MC/Real Data
Modularity
Data
Step 763
Step 495
Step 515
Step 543
Step 1739
Step 2360
Step 1738
Step 560
Step 777
Step 903
Step 269
Step 2340
Step 193
Step 918
Step 268
Platform
AllPPD
Application
Processing Pass Description
Step
Framework
Component
Powered by:
2009
2015

PPD(Modularity)

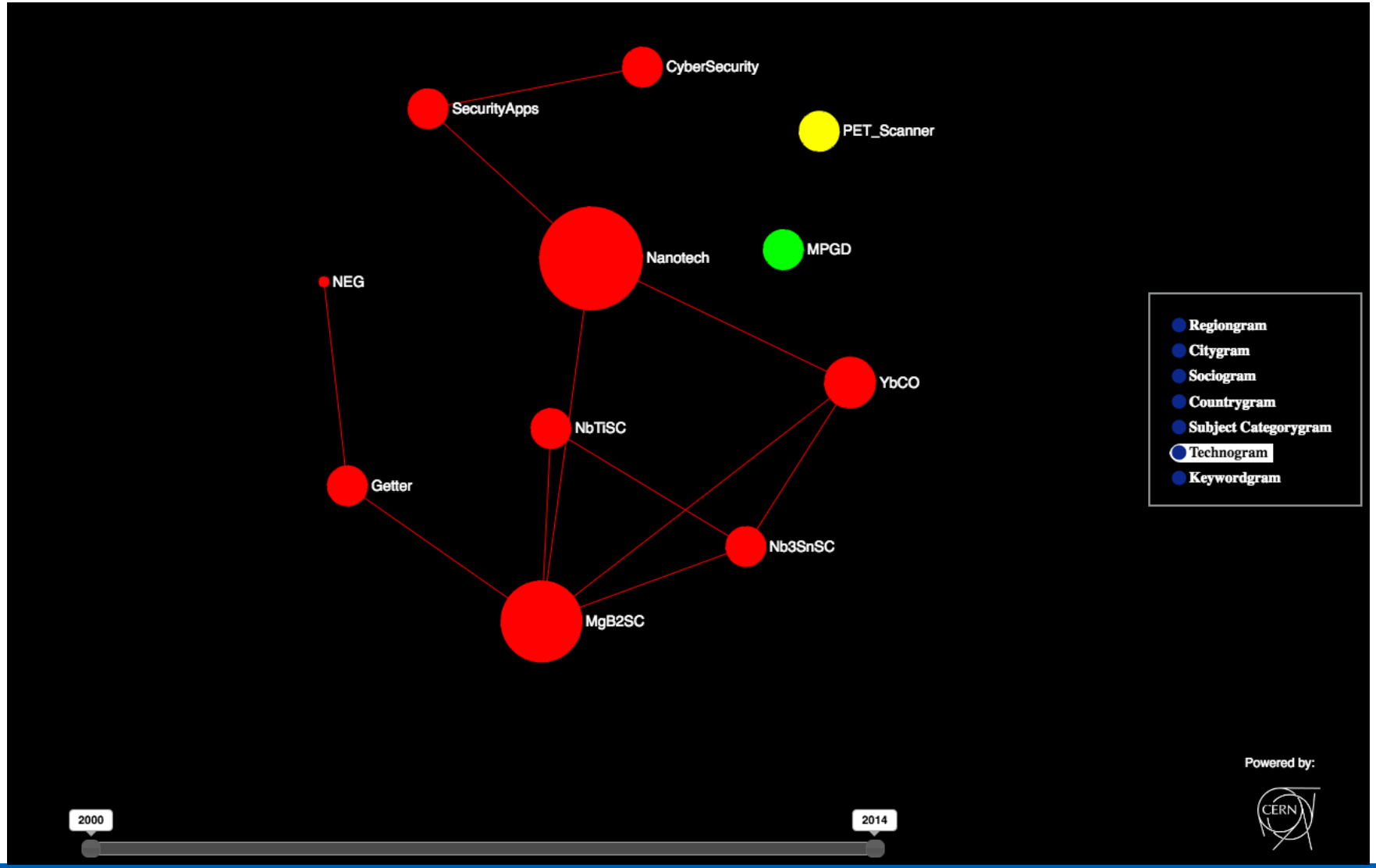


PPD(Data)



Blue: Real Data
Red: Monte Carlo

Entry graph in Tech monitoring



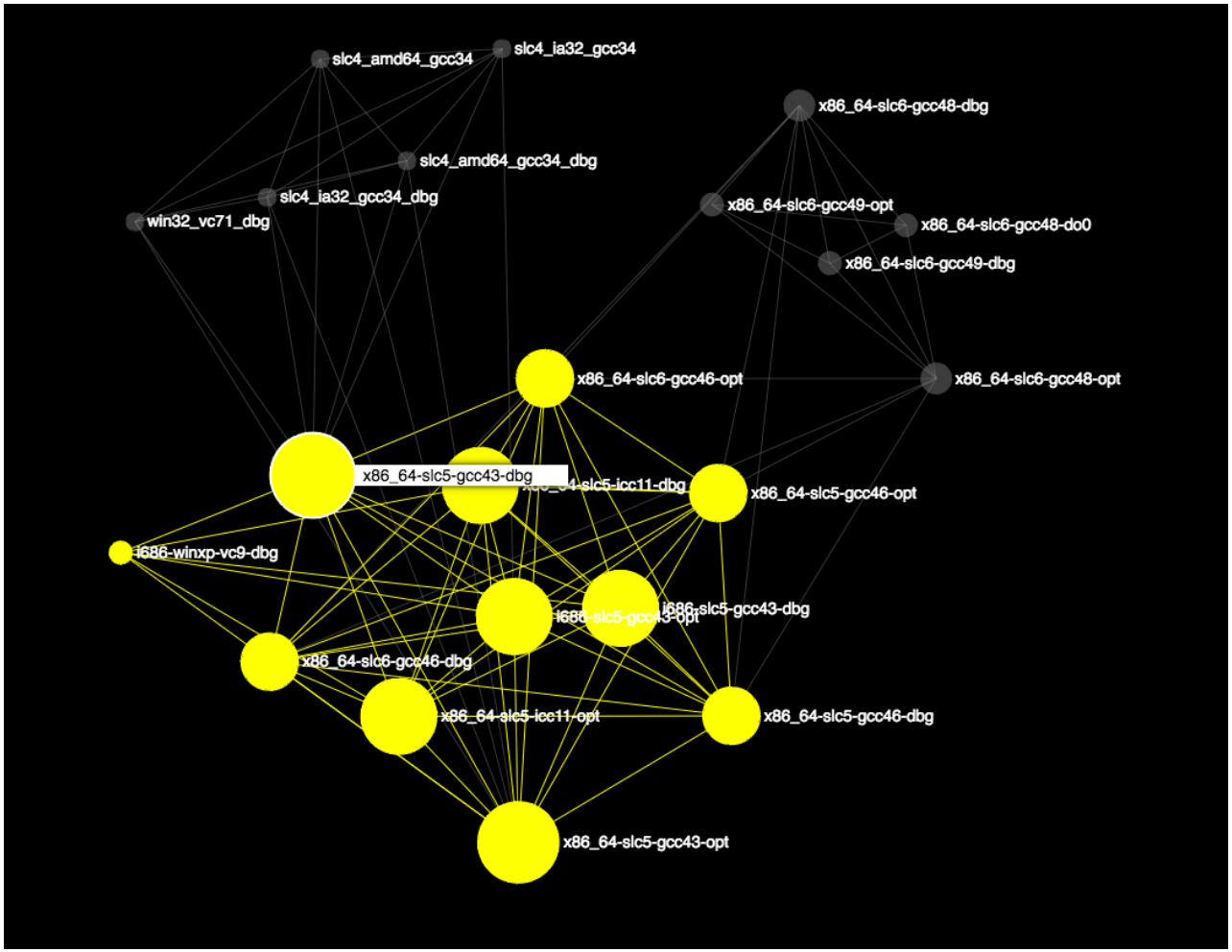
A Vertex = a semantic search

Graph navigation operations

Hovering:	• Highlight clusters
Left click:	• Node egocentric view
Right click:	• Access to other dimensions from a node
Right pane	• Navigation across dimensions
Ctrl click:	• Multiple vertex selection
Shift click	• Cluster selection

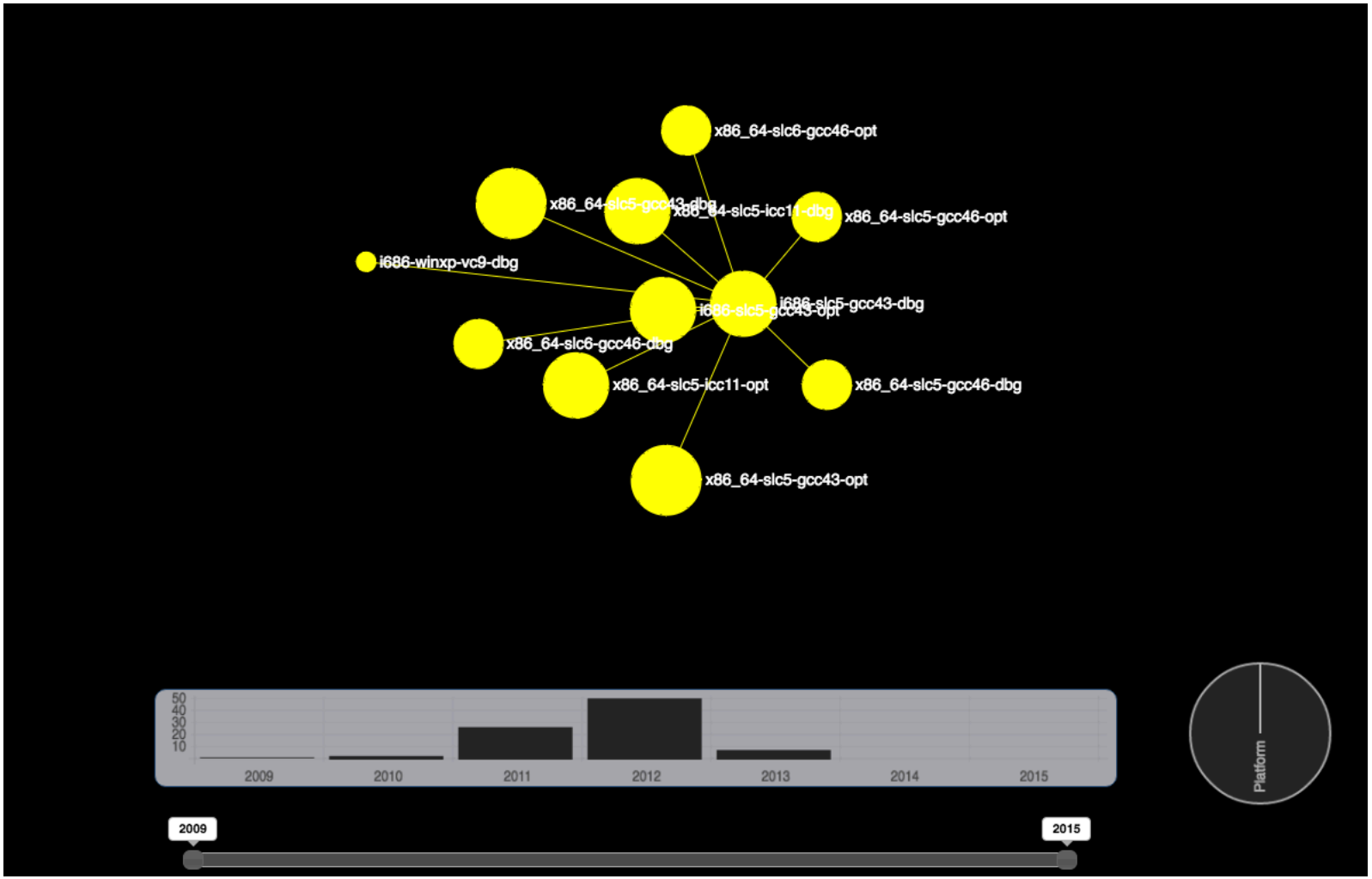
Hovering:

- **Highlight clusters**



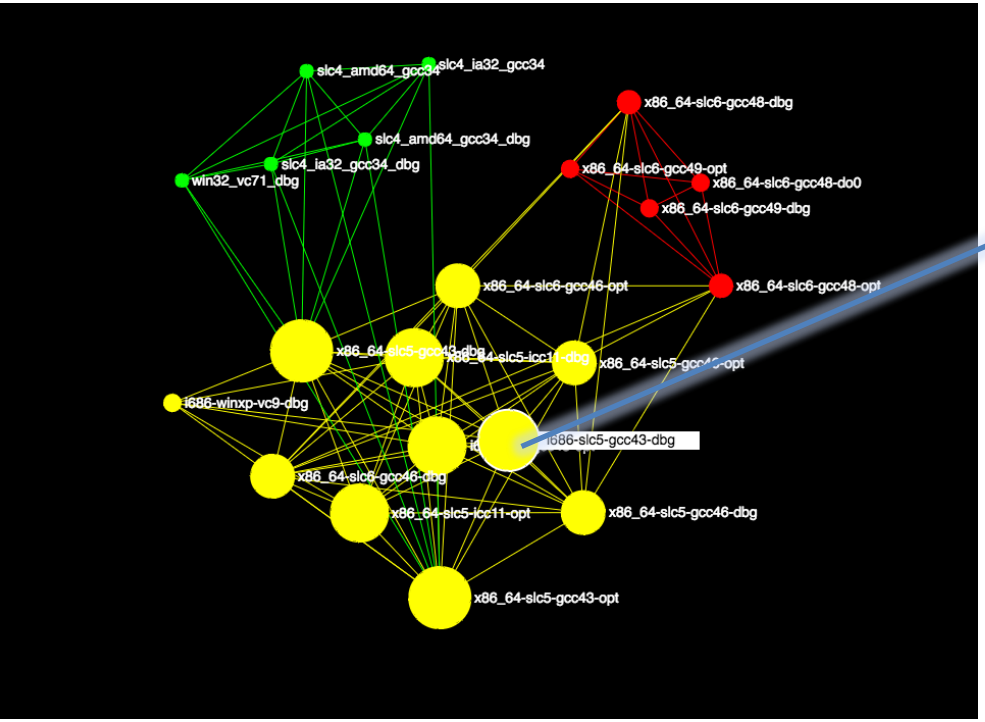
Left click:

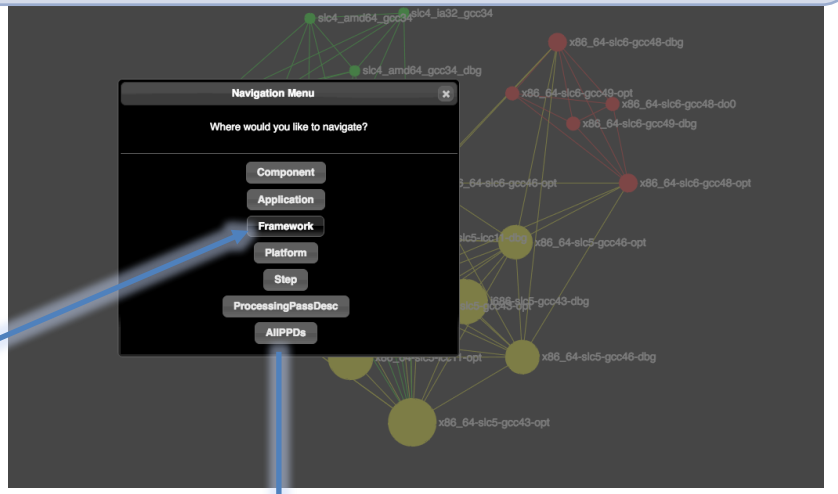
- **Node egocentric view**

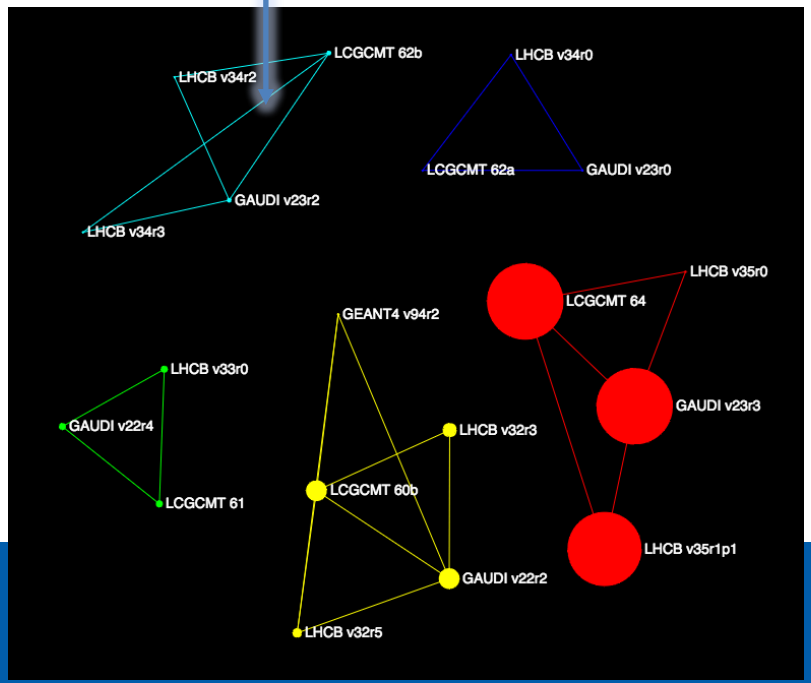


Right click:

- **Access to other dimensions from a node selection**

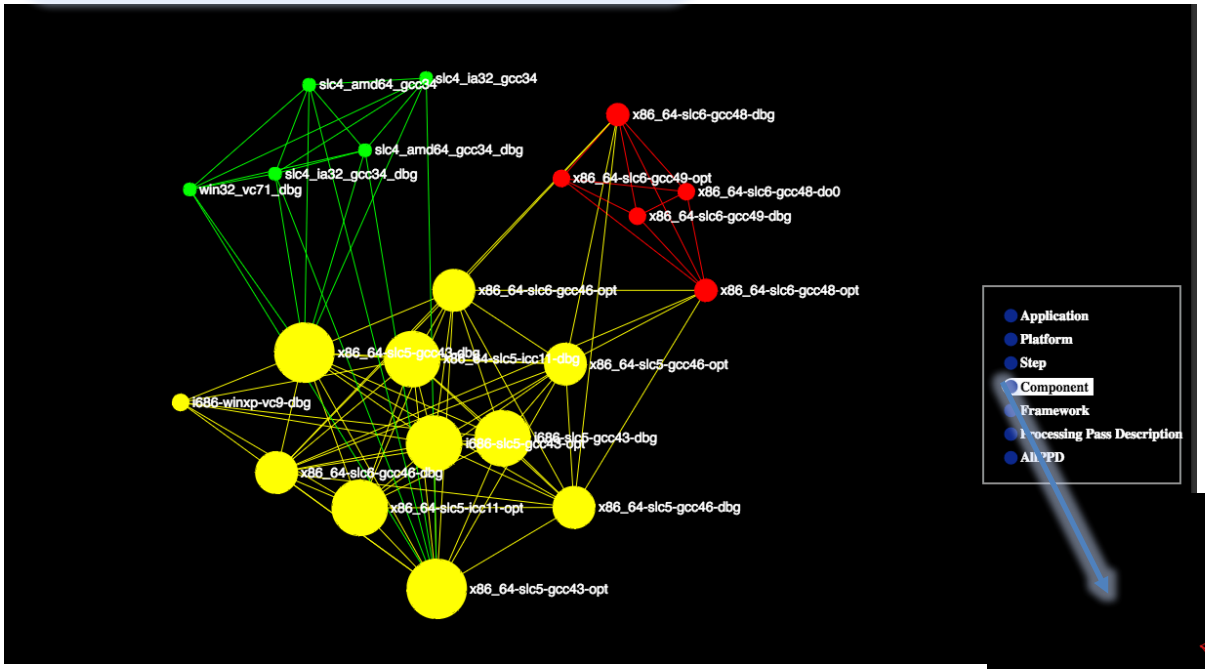


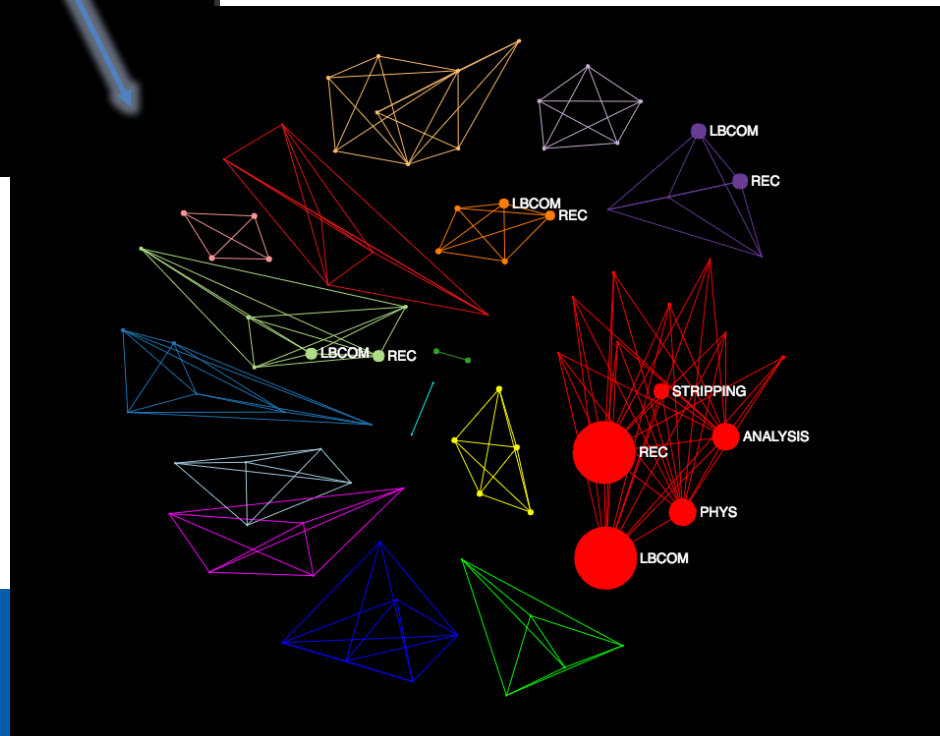




Right pane

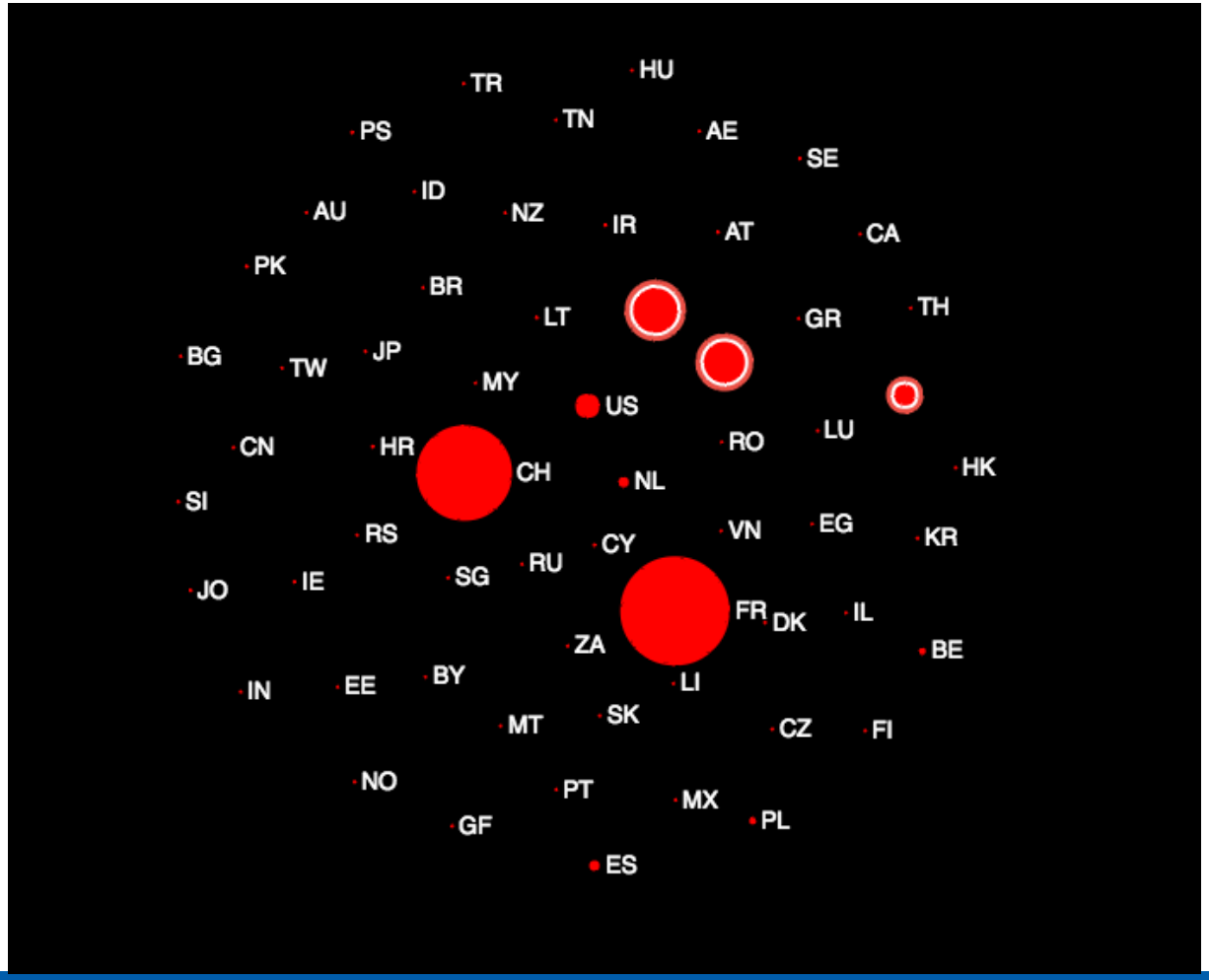
- **Navigation across dimensions**





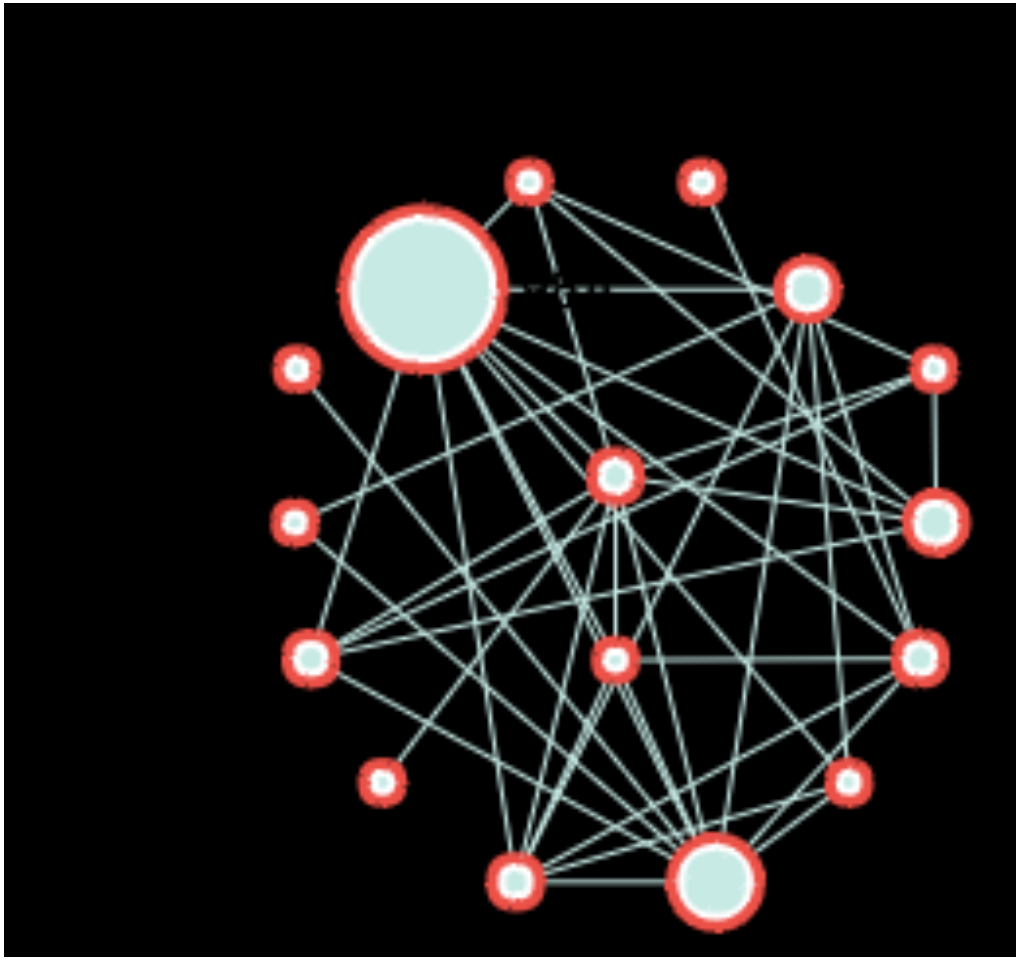
Ctrl click

- **Multiple vertex selection**



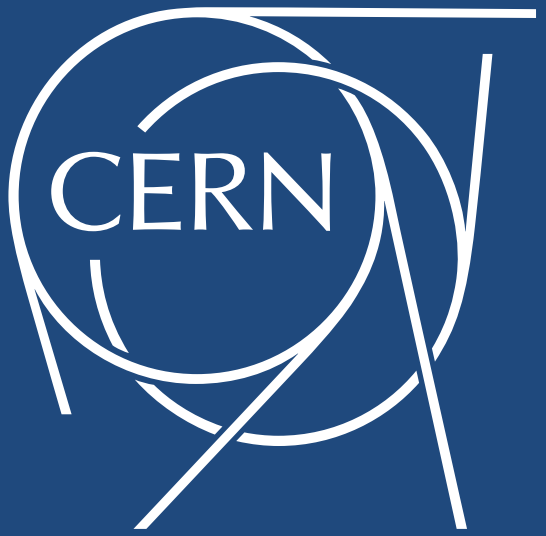
Shift click

- **Cluster selection**



Conclusion

- **CS V2 (Current version) demonstrated on**
  - Publications, patents
  - CERN procurement data
  - LHCb computing process data
  - Deployable to other data sources
- **CS V3 Platform supporting**
  - Data Manager / Expert concept
  - Full data analysis chain
  - Compound graph navigation-based mechanisms



Thank you for your attention!