Big Data analytics and Visualization

MTA Cloud symposium

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

MTA Head quarters, Budapest, 17 February 2017

Background information

- **Collaboration Spotting (CS) Platform** (V2) used to process examples
- CS is a **Visual Analytics** tool originally developed to analyse the technology landscape of key enabling technologies for the Particle Physics programme at CERN
  - Using Publications and Patent metadata
- The CS Platform has been used to visualize other datasets:
  - CERN procurement data
  - Ceased assets in collaborations with the UN-UNCRI
  - Neuro-science data in collaboration with Wigner

Characteristics of Big Data

- **Huge quantity**
- Processing and storage
- **Distributed sources**
- Access rights, security
- **Complexity**
- Valuable information may be hidden behind complexity
- **Interconnectivity**
- Unravelling new knowledge

➔ Data scientists are instrumental to analytics
➔ Domain experts are at the heart of the reasoning process

Big Data is organised in networks

Big Data is distributed

- **Document** systems with metadata in Database
- Database **tables** with metadata in schema

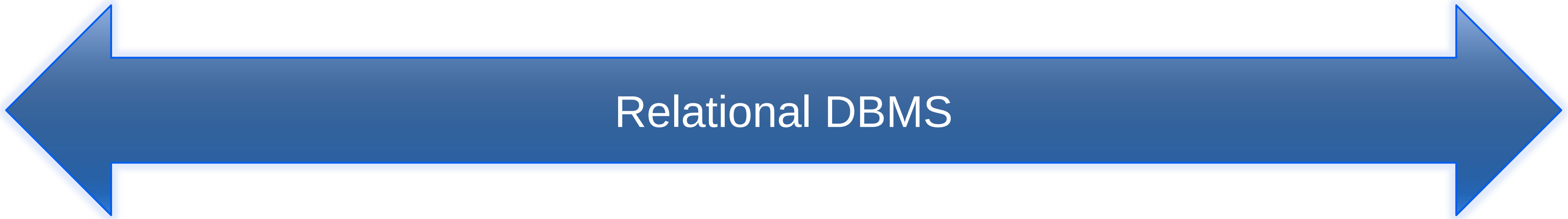
Big Data is strongly interconnected

- Connectivity **not materialised** due to the distributed nature of data sources
- Connectivity relates to the understanding of the data

Big Data Intrinsic vs additional value

- The **additional value** of Big Data comes from its **interconnectivity**

Discrete data **Connected data**



No-SQL **Graph DB**

Conventional analytics **Conventional + visual analytics**

Two Criteria:

Bottom-up VS Top Down

Discrete data VS highly interconnected data

When do domain experts really need to visualise Big Data Networks?

Top-Down VS Bottom-up

- **Process driven**
- Hypothesis
- Simulation software
- Validation with real data
- Review hypothesis

- ➔Experiments
- ➔Compare results with simulation

Typically hard sciences

- **Data driven**
- Extract features from data
- Generate hypothesis
- Run what-if scenario
- Validate with data

- ➔Big Data
- ➔Software for domain expert to make sense out of Big Data

Empirical approach

Discrete data — **Connected data**

Relational DBMS

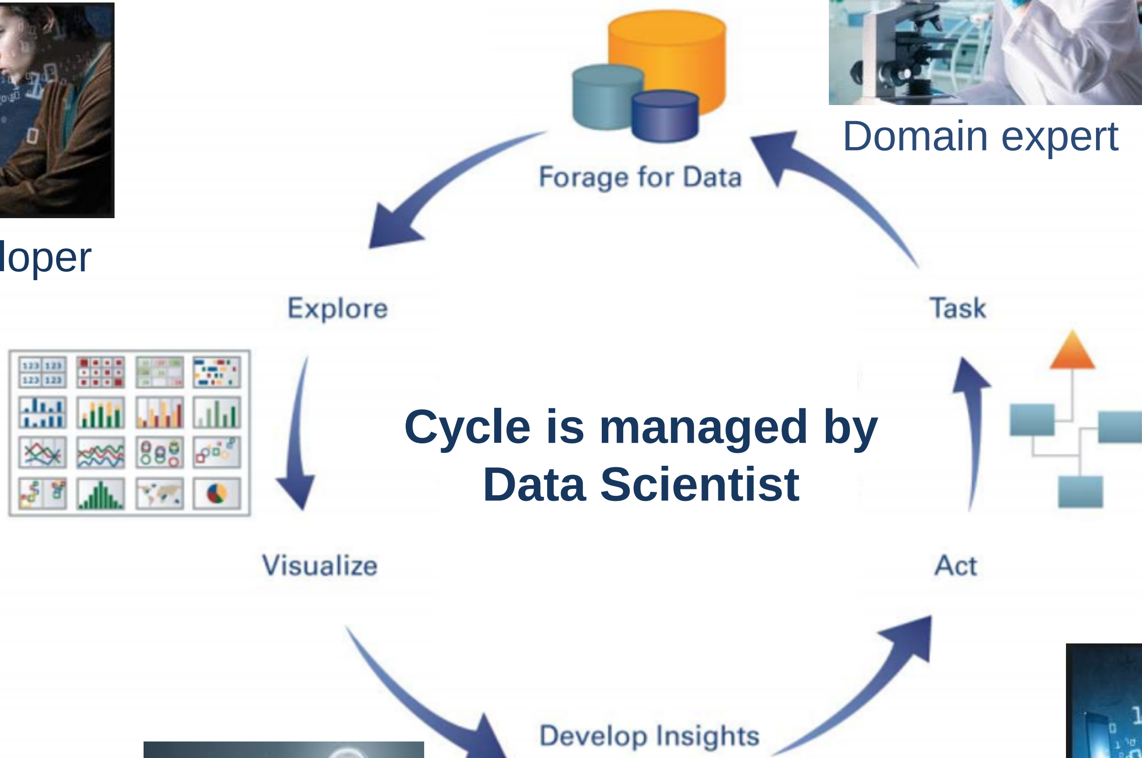
No-SQL — Graph DB

Domain Expert

Domain expert vs Data scientist

Software developer

Domain expert



Cycle is managed by Data Scientist

Source: Tableau Software

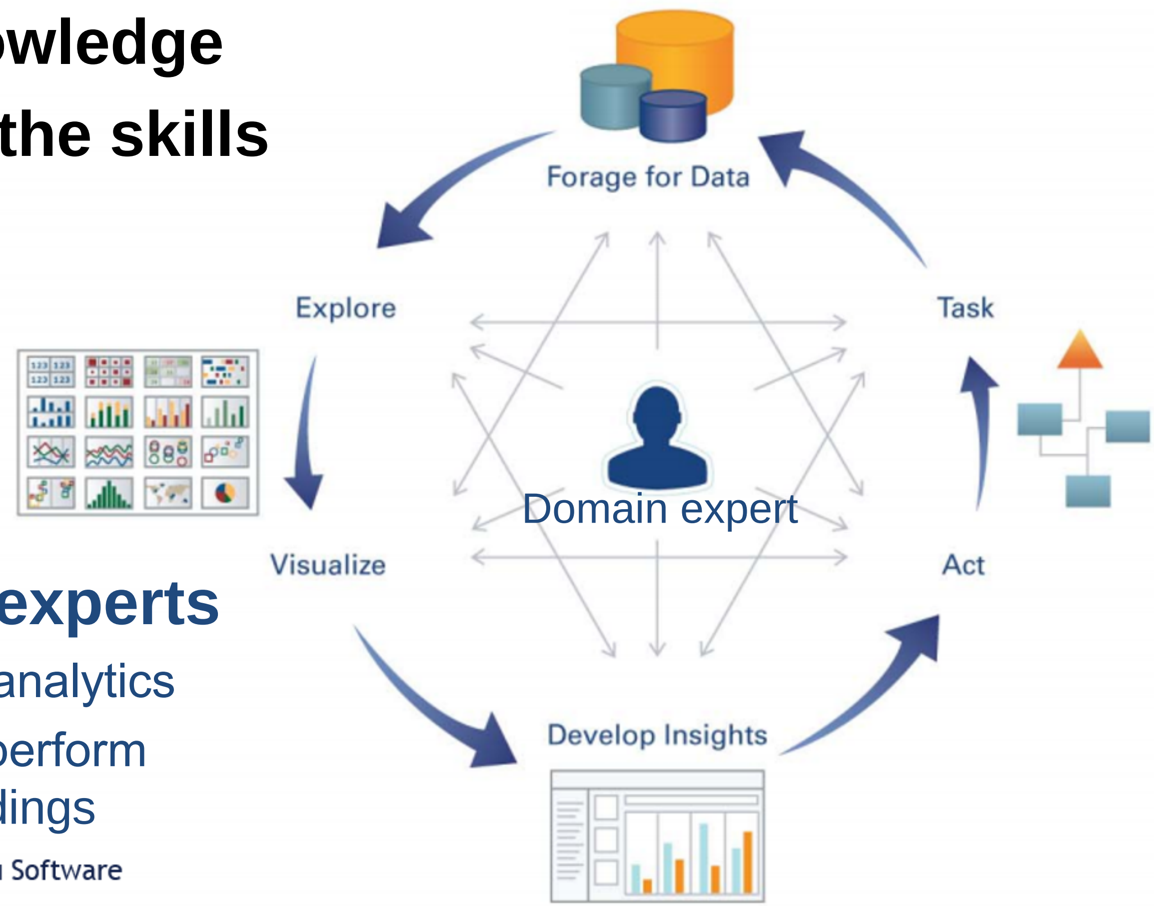
Domain expert

Software developer

Source: JIOX: Intelligence Tradecraft & Analysis

Challenge ➔ Bring domain experts at the centre of the visual analytics 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

Data scientists to build platforms that enable experts to perform analytics by themselves

What is required?

Network Data and Domain independent

- Support interconnectivity
- Support Cross Domain applications

Smart Data management concepts and tools

Scalable and flexible

- Support any combination of data structures

Easily accessible and navigable to Experts

- Support visualisation of network content

Smart graphic management concepts and tools

Enhance value of Data Network for Experts

- Support navigation of network content
- Support queries of network content

A Domain independent platform

Smart Data Management

Directed graphs are natural representations of large and interconnected datasets

- Complexity
- Interconnectivity
- Scalability
- Multi dimensional

Schema is embedded in the data

- Nodes' labels
- Compact graph structure
- Graph query language
- No schema evolution

Graphs of connected elements constitute multi-dimensional networks

- Schema: labels and edges (interconnectivity)
  - Labels ←→ Graph dimensions
  - Edges ←→ Directed relationships between Labels
- Data graph: vertices and edges
  - Vertices: data instances and dimension instances
  - Edges: Directed relationships between vertices

CERN

→ Graph Databases offer a natural support for storing network information

→ Label property graph data model

rk from two data

Journal of Crystal Growth

Volume 452, 15 October 2016, Pages 22–26

ELSEVIER

ACCGE-20-
Growth and E
on Organome

Analysis of the effe
fields on melt/cryst
growth

Parthiv Daggolu[a], Jae Woo Ryu[a]

Show more

http://dx.doi.org/10.1016/j.jcrysg

Highlights

- Effect of CUSP magneti
- Symmetric CUSP provi
- Asymmetric CUSP with
- 2D model predicts the i
- 3D unsteady model allo

Abstract

With the use of 300 mm sili
manufacturing, the Czochr
achieve higher quality and
models combined with 3D
time and money in the proc
growth is controlled by a C
turbulence. MF can be opt

Bibliographic data: RO122515 (B1) — 2009-07-30

In my patents list | Previous | 2/24 | Next | Report data error | Print

PROCESS FOR MANUFACTURING A HIGH EFFICIENCY SOLAR CELL ON MONOCRYSTALLINE SILICON

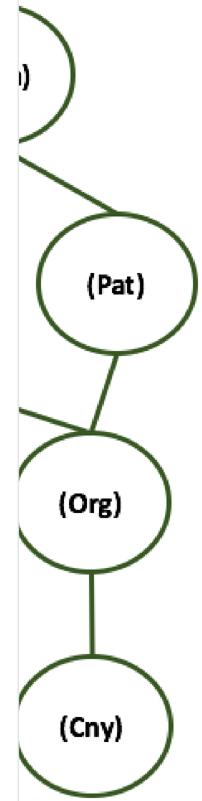
Page bookmark	RO122515 (B1) - PROCESS FOR MANUFACTURING A HIGH EFFICIENCY SOLAR CELL ON MONOCRYSTALLINE SILICON
Inventor(s):	MANEA ELENA [RO]; PODARU CECILIA [RO]; BUDIANU ELENA [RO]; MUNIZER PURICA [RO]; CORACI ANTONIE [RO]; POPESCU ALINA [RO] +
Applicant(s):	INSTITUTUL NATIONAL ICCF [RO] +
Classification:	**- international:** *H01L27/142; H01L31/04; H01L31/042; H01L31/06*
	- cooperative: Y02E10/50
Application number:	RO20060000749 20060927
Priority number(s):	RO20060000749 20060927

Abstract of RO122515 (B1)

Translate this text into

Select language | patenttranslate powered by EPO and Google

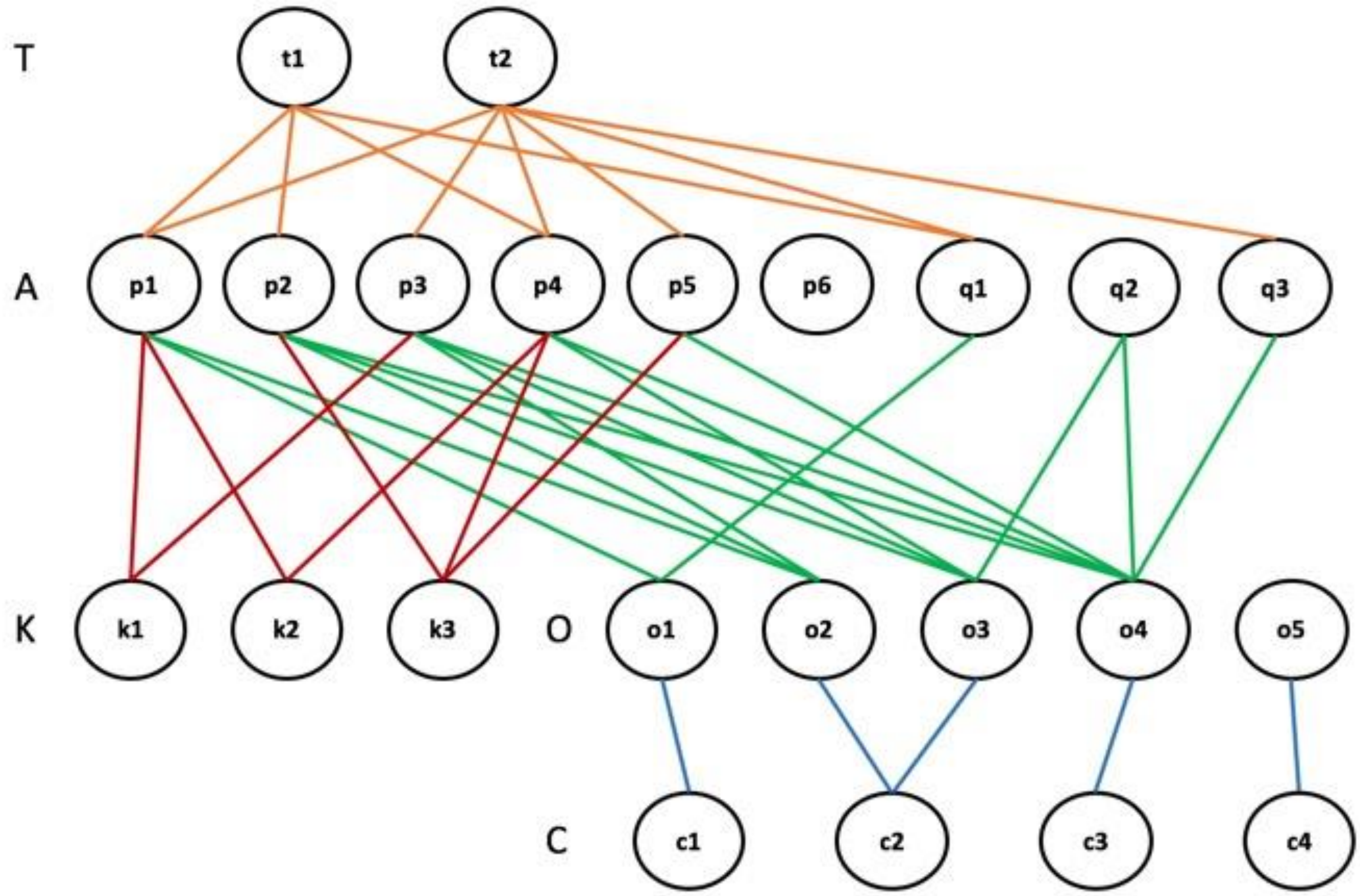
The invention relates to a process for manufacturing a solar cell by using a monocrystalline silicon substrate. According to the invention, the process consists in manufacturing a Czochralsky -type substrate silicon wafers (1), said substrate exhibiting a resistance of 1 ... 2 ohm x cm, orientation <100>, till the thickness of the oxide (2) reaches 1.8 nium, at a temperature of 1,000A C and for 300 min/vapours, followed by the removal of the oxide from the wafers back side by means of photolithographic process, and then the wafers are enriched with boron from a solid source at a temperature of 1,110A C for 20 min/source, in order to obtain a protection oxide (3) with a thickness of 0.6 nium, followed by the application of a photolithographic process for texturing the wafer front side, obtained by silicon etching through the oxide masking layer, by using the "honeycomb-like array" topography, at the end of the etching, the oxide used as masking layer being lifted, leaving the active area free, followed by forming the junction n+ (5) in the active area carried out by a prediffusion from liquid source of POCl3 at a temperature of 1,000A C for 10 min/vapours, a junction of 0.8 nium, a V/I ratio = 0.5 ohm and an anti-reflex oxide (6) of 95 mium being obtained.



Document metadata | **Graph of data types**

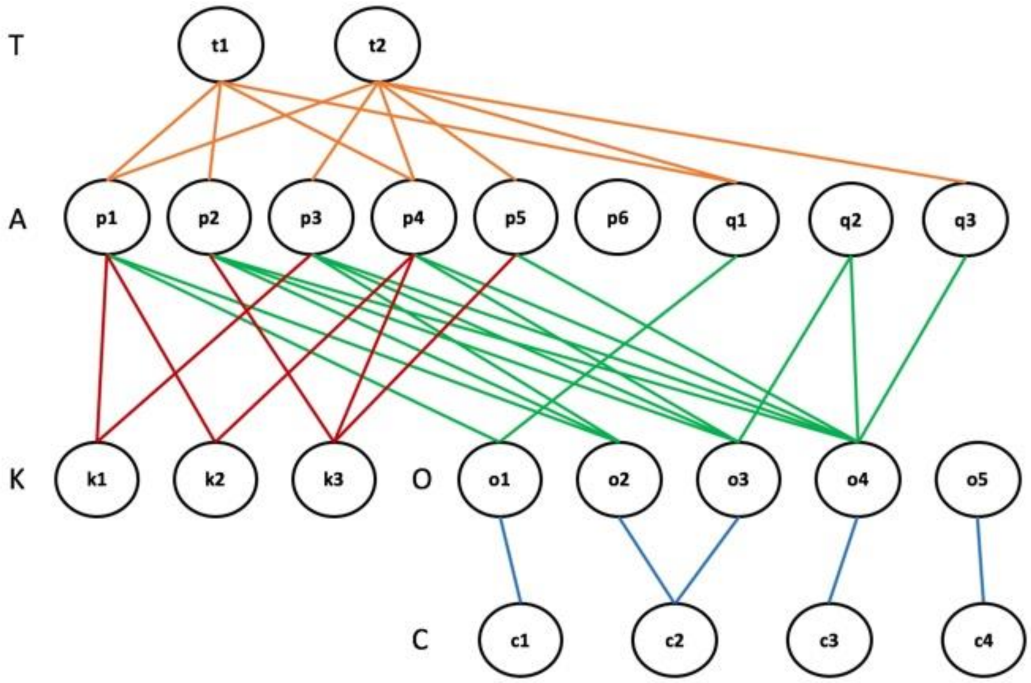
SCat: Journal category, Kw: Keyword, Org: Organisation, Cny: Country, Tech: Technology

Graph of Network

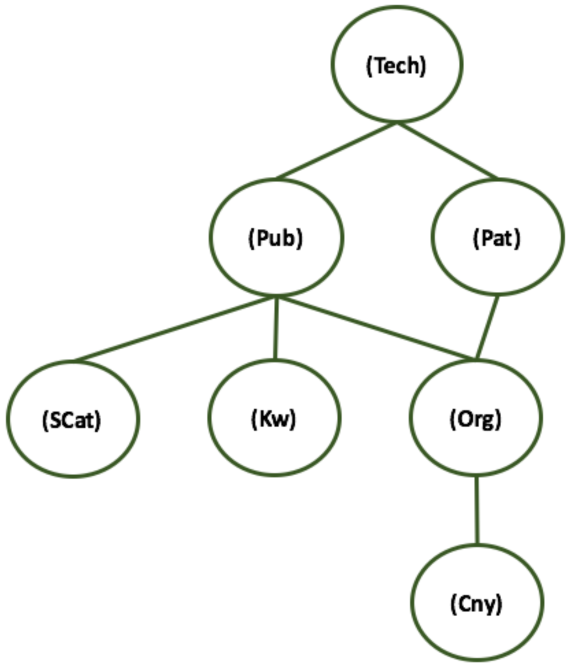


T: Technologies, A: Pub/Pat, K: keywords, O: Organisations, C: Countries

Data Graph & Graph Schema

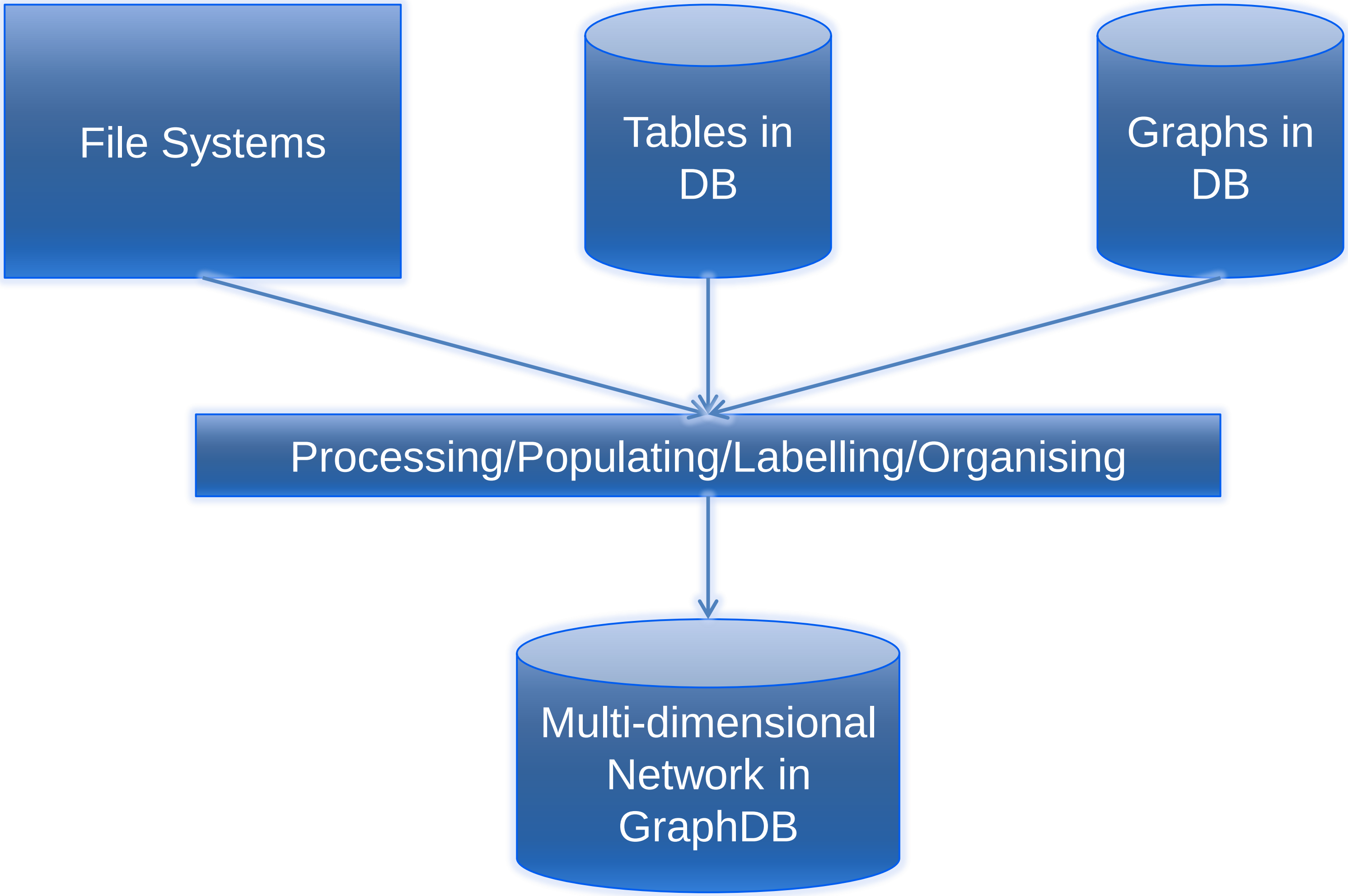


Graph of data network



Reachability Graph

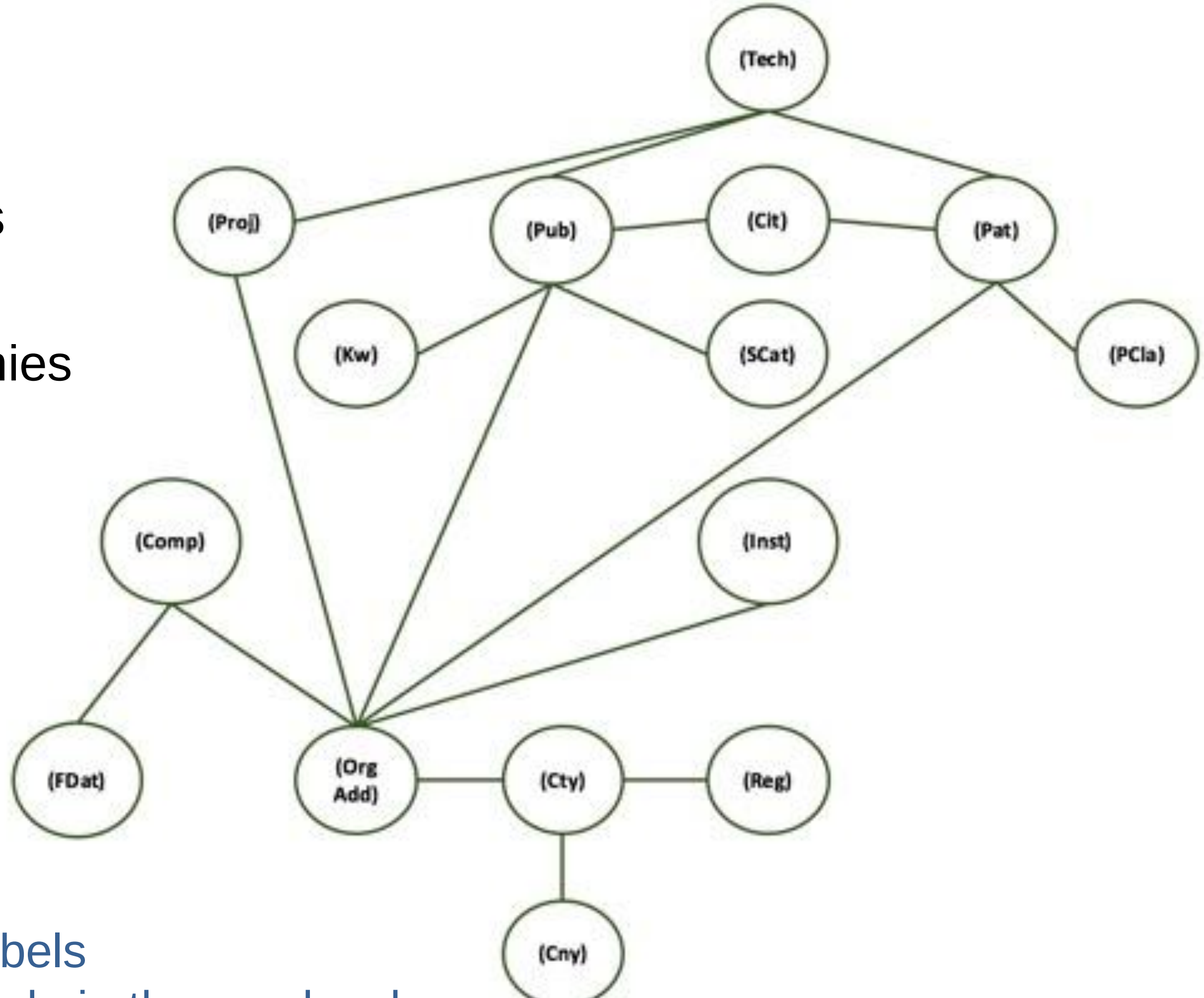
Building multi-dimensional networks



➔ **No limitations on the sixe of a network!**

Combining data sources ➔ Enriching networks ➔ More interconnectivity

- **Data sources**
  - Publications/Patents
    - Citations
    - Institutions/Companies
- **Data sources**
  - EU projects
  - Financial data
  - Geolocation data



Schema: Graph of datatypes/labels
Dimension: a datatype i.e. a node in the graph schema

➔ **No limitations on the extension of the network's schema!**

Smart Graphic Concepts and Management tools

Graphs are excellent for visualising networks

- Retain complexity
- Singularities
- Clusters/communities/patterns

Graphs contain many visual information

- Vertex label, shape, size and colour to visualise properties of datasets
- Edges colours to highlight clusters

➔ Visualisation enhances the perceptual reasoning potential of analytics

Smart Graphic Concepts and Management tools(2)

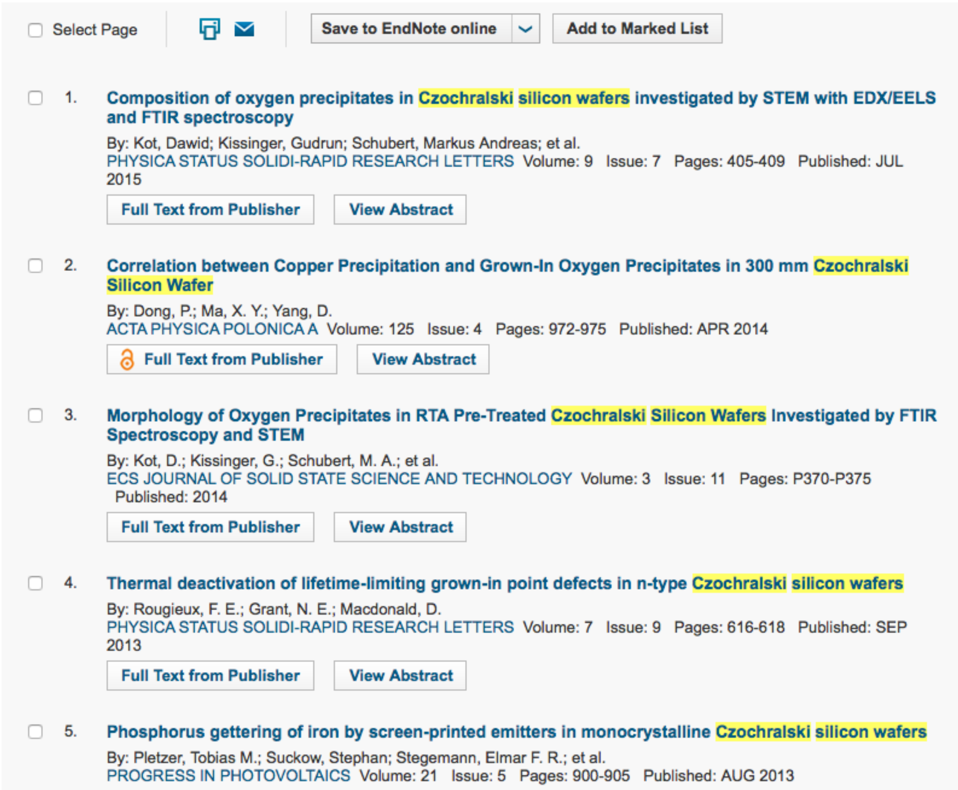
Maximizing human understanding

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

Enhancing reasoning

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

Selecting Visualisation dimensions



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

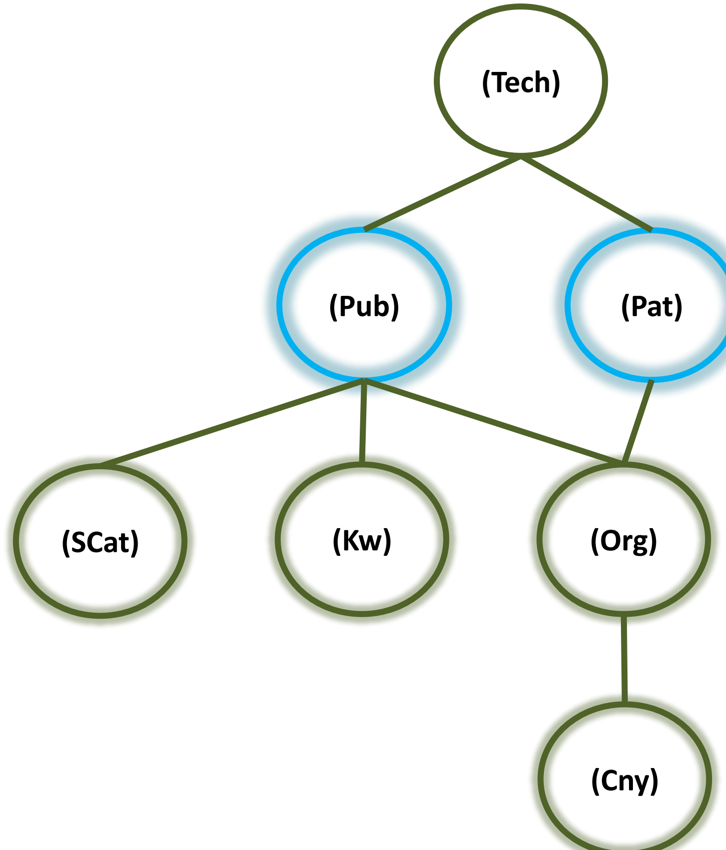
1. **Composition of oxygen precipitates in Czochralski silicon wafers investigated by STEM with EDX/EELS and FTIR spectroscopy**
By: Kot, Dawid; Kissinger, Gudrun; Schubert, Markus Andreas; et al.
PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 9 Issue: 7 Pages: 405-409 Published: JUL 2015
Full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



Reference dimensions for Analytics

Pub: Publications, Pat: Patents (Attributes: Title and abstract are used for semantic searches)

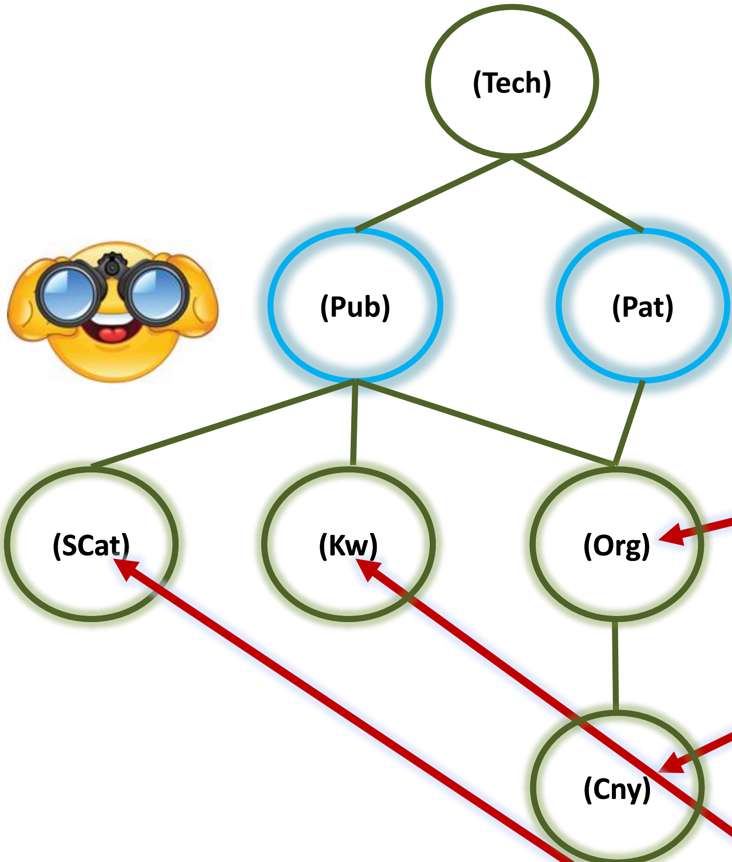
Visualisation dimensions of Analytics results:

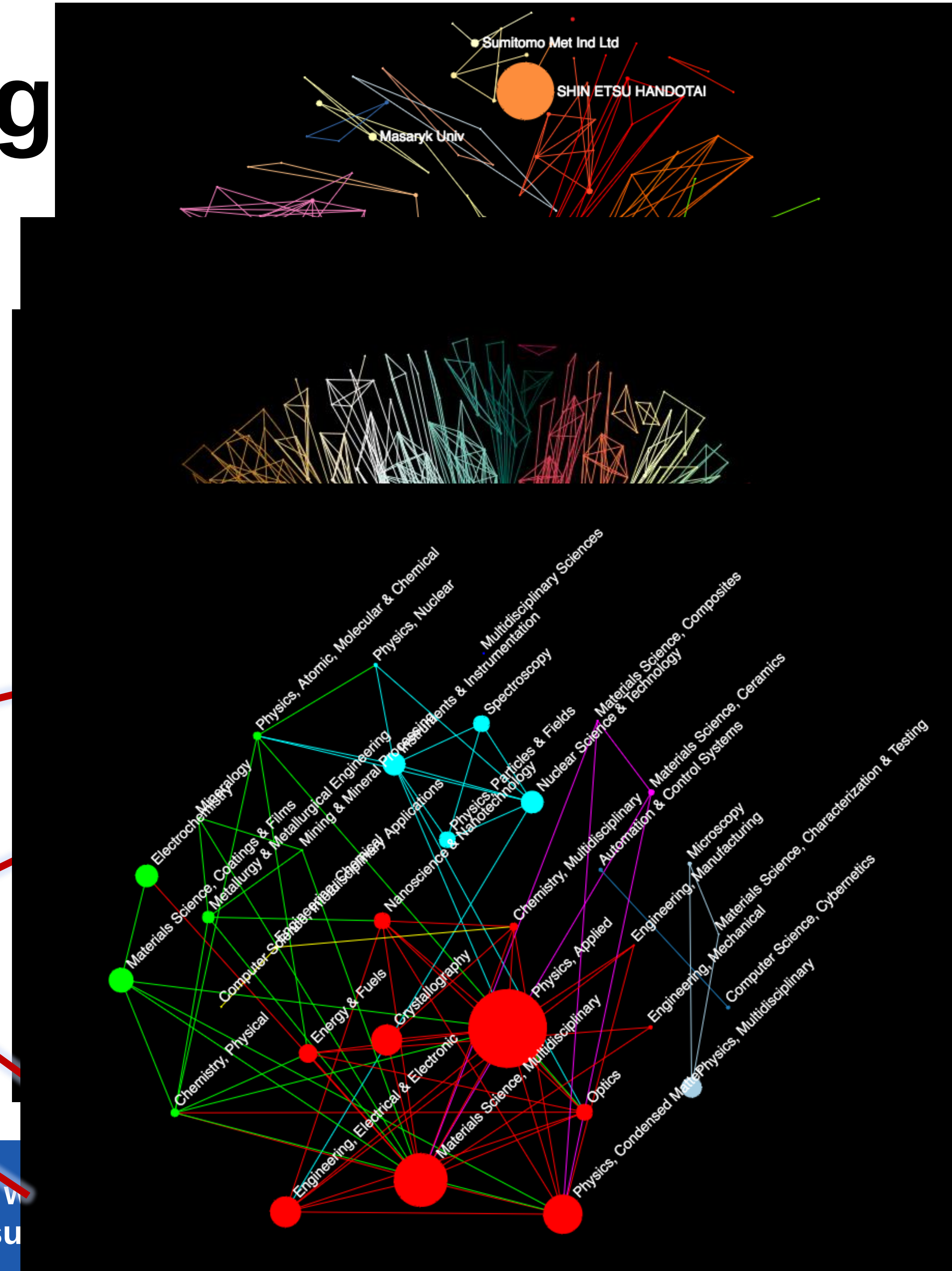
SCat: Journal category, Kw: Keyword, Org: Organisation and Cny: Country)

CERN

Technology Search: Czochralski Silicon wafer
Pub/Pat: documents found in search results

Traversing


(Tech)
(Pub)
(Pat)
(SCat)
(Kw)
(Org)
(Cny)


Sumitomo Met Ind Ltd
SHIN ETSU HANDOTAI
Masaryk Univ
Physics, Atomic, Molecular & Chemical
Physics, Nuclear
Multidisciplinary Sciences
Instruments & Instrumentation
Spectroscopy
Materials Science, Composites
Materials Science, Ceramics
Automation & Control Systems
Materials Science, Characterization & Testing
Mineralogy
Electrochemistry
Materials Science, Coatings & Films
Metallurgy & Metallurgical Engineering
Mining & Mineral Processing
Nuclear Science & Technology
Physics, Particles & Fields
Nanoscience & Nanotechnology
Chemistry, Multidisciplinary
Microscopy
Engineering, Manufacturing
Materials Science, Multidisciplinary
Engineering, Mechanical
Computer Science, Cybernetics
Physics, Applied
Crystallography
Energy & Fuels
Chemistry, Physical
Engineering, Electrical & Electronic
Optics
Physics, Condensed Matter
Physics, Multidisciplinary

How to scale up the “graph” approach for very large multi-dimensional networks?

Visual analytics for large datasets

Visual analytics features

- **Visual analytics <u>does not replace</u> Big Data analytics ➔ Visualize results**
- **Maintain visual perception quality and user interactivity**
  - No matter the **size**
  - No matter the **diversity** (dimensions)
  - No matter the **interconnectivity**

➔ Data sampling & filtering

➔ Visualize subsets of network dimensions

➔ View data from different perspectives

Visual analytics Needs

- Visualize part of data network with respect to particular references and from different perspectives
  - **Reference**: Data dimensions (labels)
  - **Perspective**: Visual dimensions (labels)
- Need to navigate across visual dimensions
  - => Visual queries
- Need to get contextual statistics
  - In the context of a particular view
- Need to change Data Reference while navigating
  - Queries adapted to change of reference

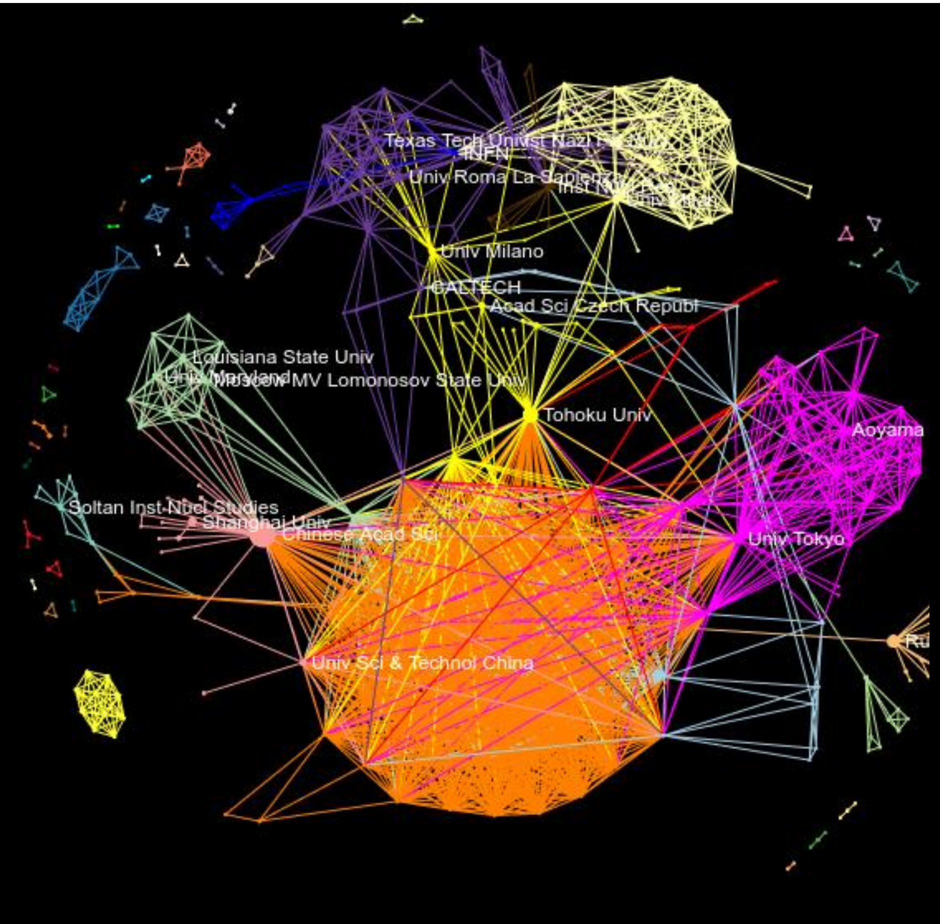
Visual analytics Features (2)

- **Structural vs Behavioural**
  - Understand from the data how something is working
- **Visualization**
  - Maximum number of collaborations that can be processed (~100k) to feed visualization
  - Maximum number of vertices and edges one can visualize within a graph (~ 10k)
  - Maximum number of Clusters one can visualize within a graph (~10k)
- **Data quality**
  - Can the data be trusted?
  - How complete is the dataset under study?

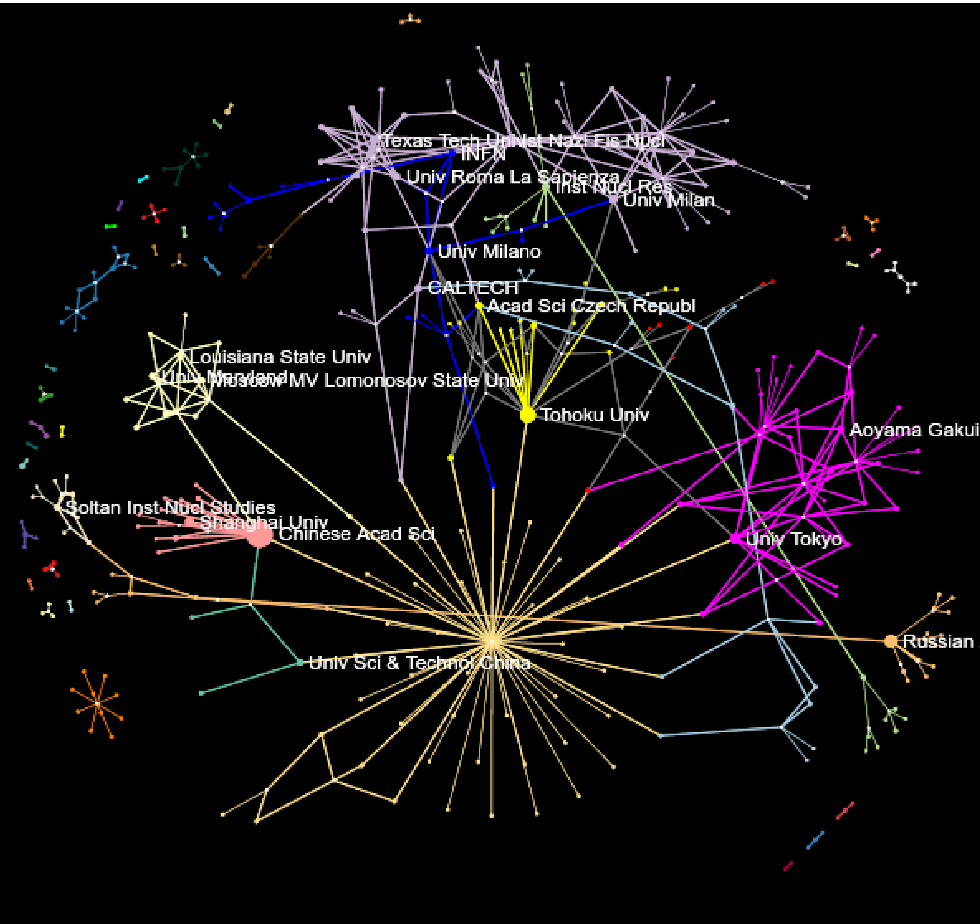
Visual analytics Needs(2)

- Need to visualize **processes**, **interactions** in addition to **structure** of data network
  - **Connectivity** graphs **AND**
  - **Causality** graphs ➔ directed edges
- For large graphs:
  - Replace vertices with communities in complex graphs
  - Compound graph approach
- For graphs built out of large collaborations
  - Replace 2-adic calculations with m-adic
- Example
  - Neuro science: paths of length 2 to visualize input/process/target flows

Reduce visual complexity & faster graph processing: Hyperedges vs edges



Organisation landscape graph view

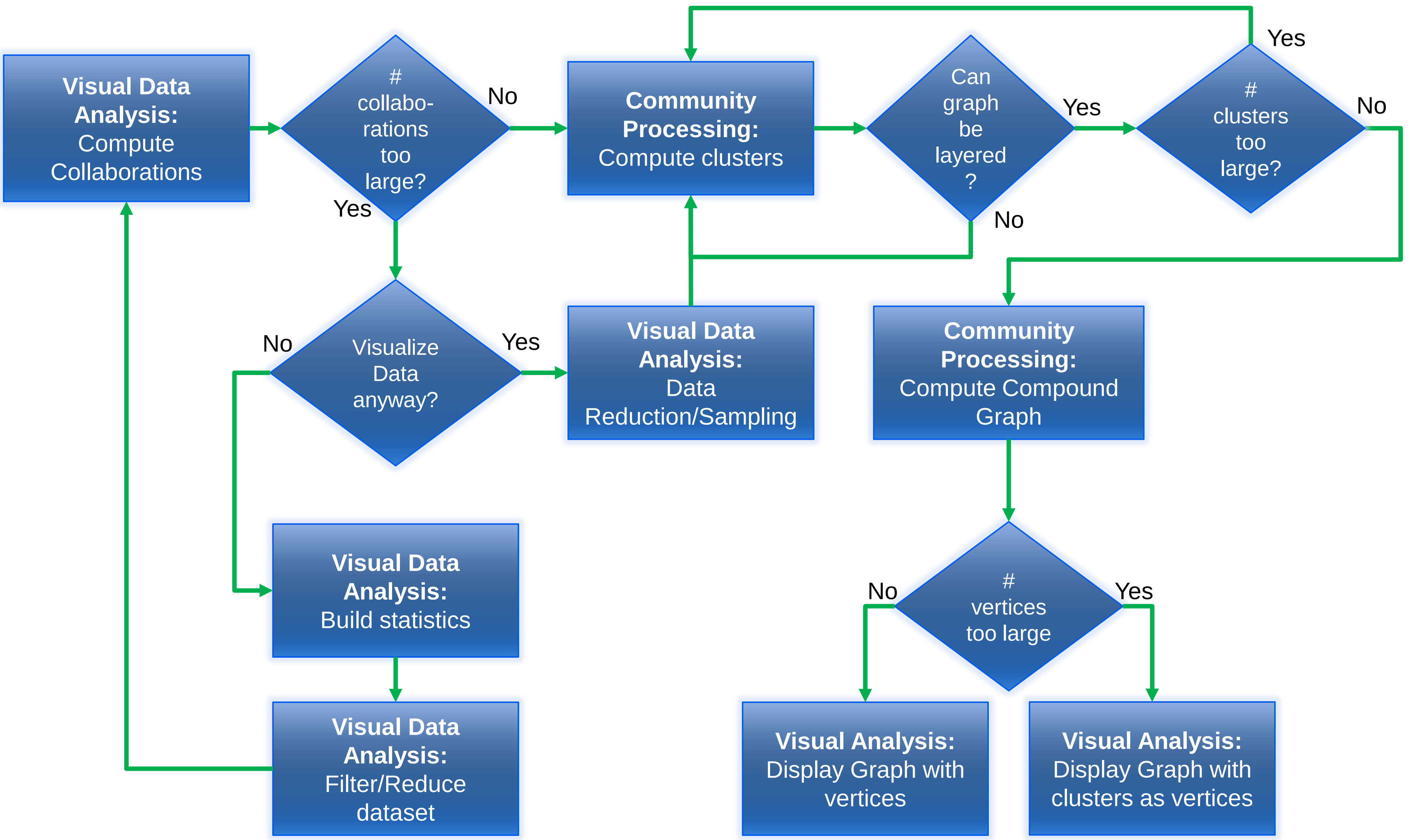


Organisation landscape hypergraph view

Tailor visualisation to data

- **STRATEGY**: Combining various techniques to support quality visual perception and user interactions according to data and graph sizes
  - Statistics
  - Data sampling & Reduction
  - Compound graphs
  - 2-adic vs n-adic node-link graph representation

Combining techniques for visualisation



➔**Objective: Reduce dataset and graph content with very minimal loss in visual perception**

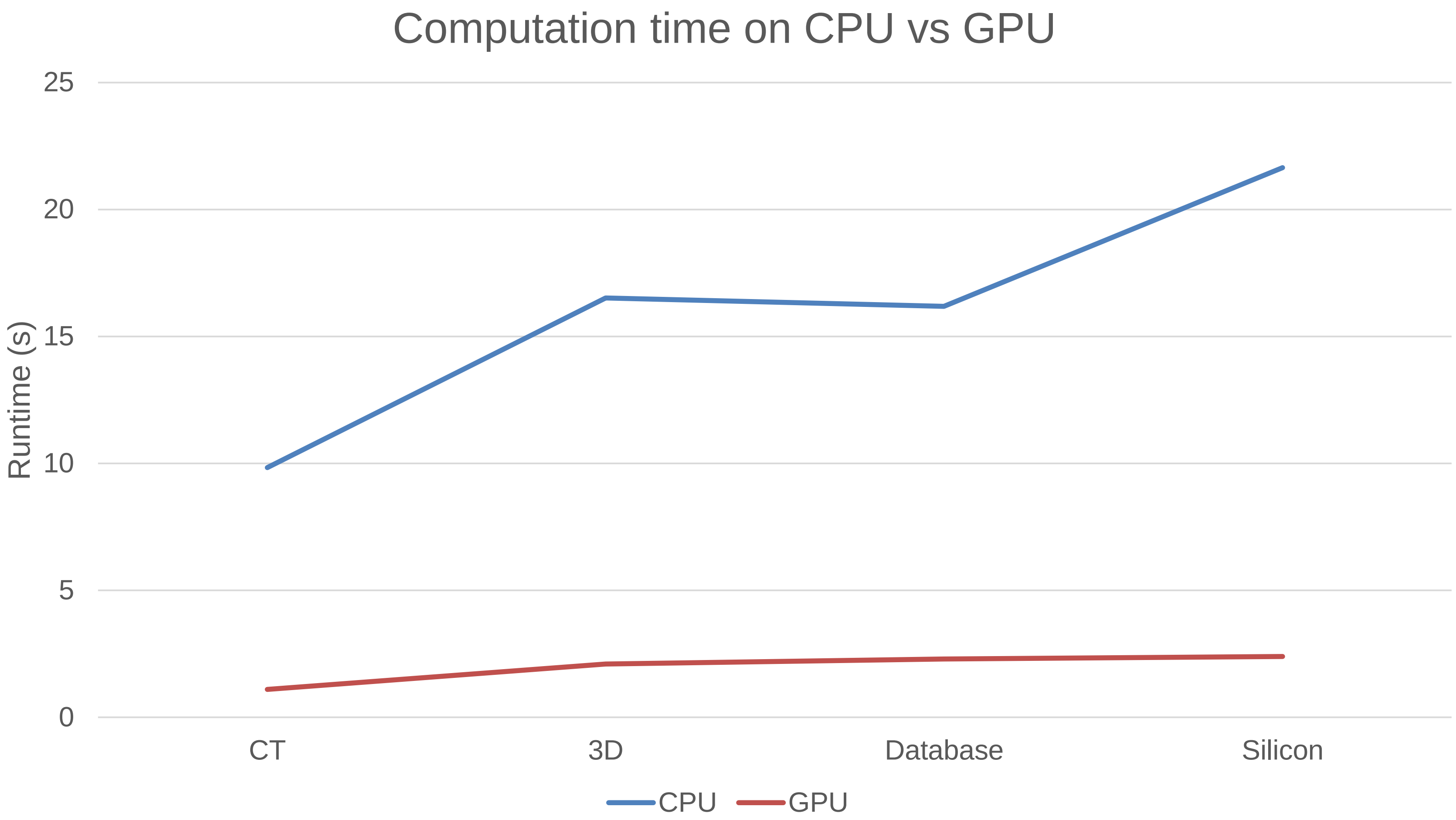
Computing requirements for visualization

- Service users within a few seconds
- Heavy computing at the backend to process clusters, optimize layout and support visual navigation
- Need for Cloud computing
  - Using machines with 4 CPU cores (8 threads), 8 GB of memory
- CPU vs GPU
  - Comparing them using consumer level hardware (Intel Core i7, GeForce GTX 980)

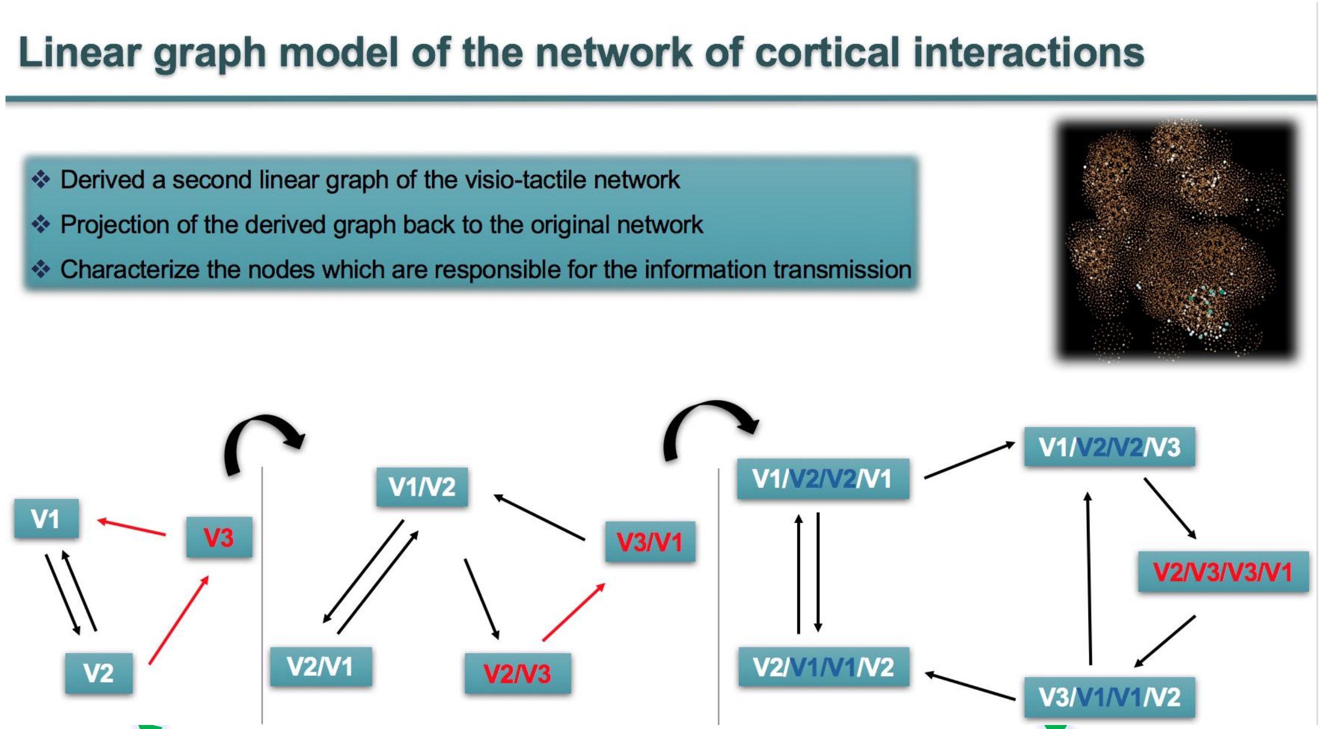
Computing requirements for visualization

- Computation on the CPU
  - Graphs with tens of thousands of nodes and hundreds of thousands of edges, computing requires *~17 seconds*.
  - Further optimization can be achieved by further distributing the computation among multiple machines
- Computation on the GPU
  - Same graphs compute ~8 times faster (*~2 seconds*)
  - Distribution among multiple GPUs is a further possible optimization

Computing requirements for visualization



The macaque case

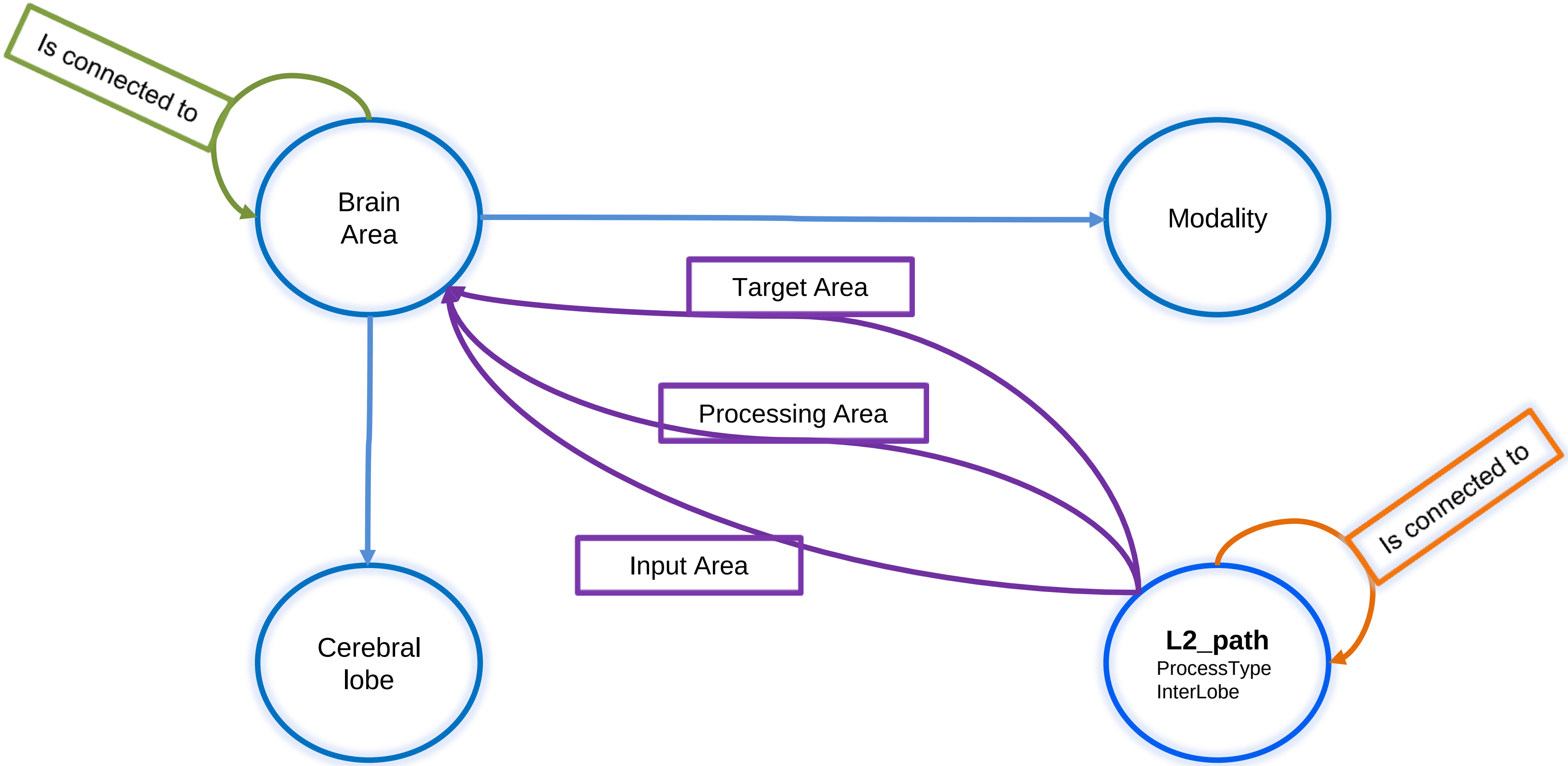


g2 is too large for visual perception
➔ Communities
172 clusters
10668 edges

g_0: directed graph of brain area interconnectivity*
(**42** vertices = areas, **601** edges= interactions)

g_2: directed graph of cortical interactions*
(Input/Processing/Target)
(**9869** vertices = IPT flows, **166219** edges = common interactions)

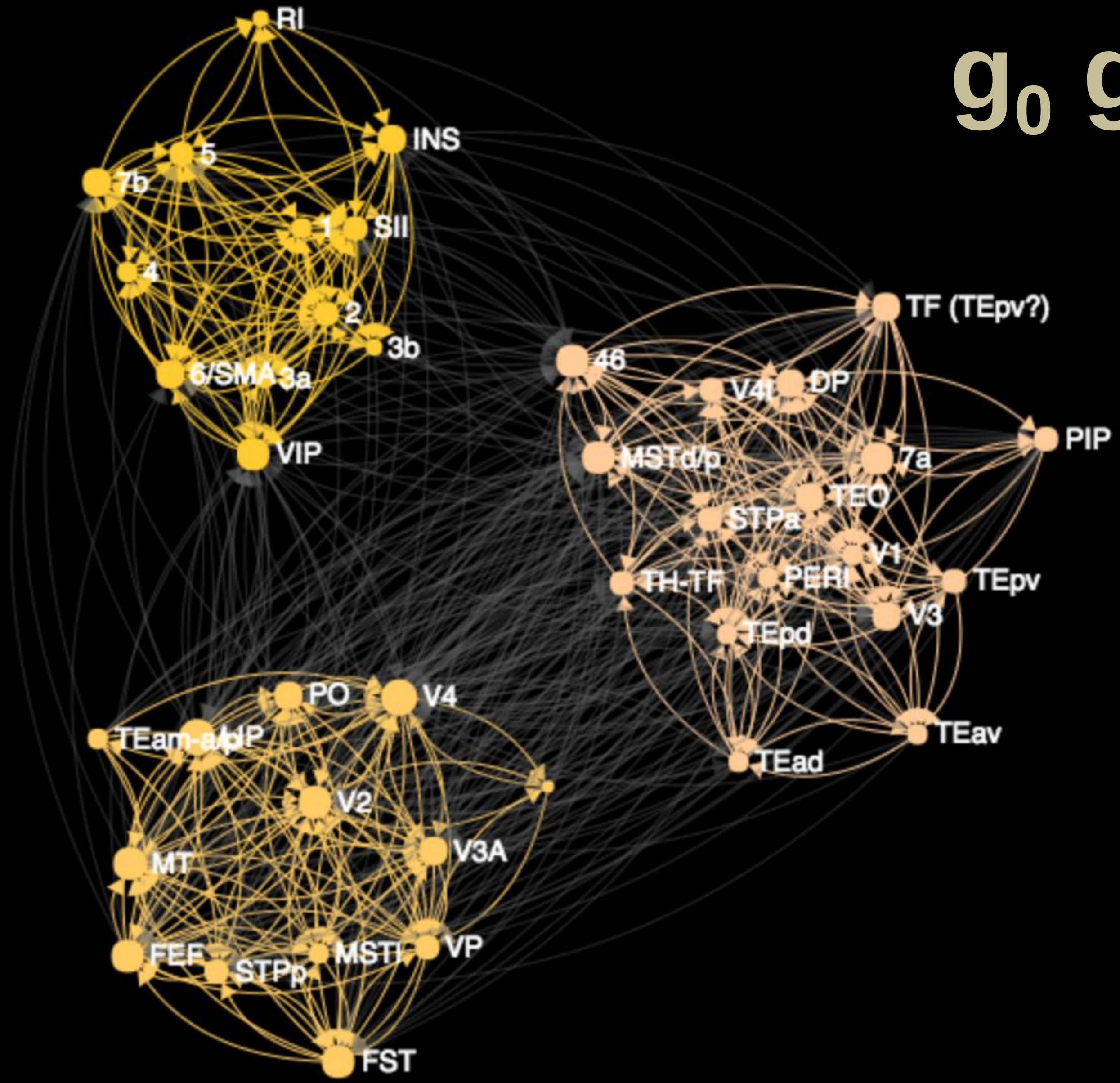
Constructed Reachability Graph



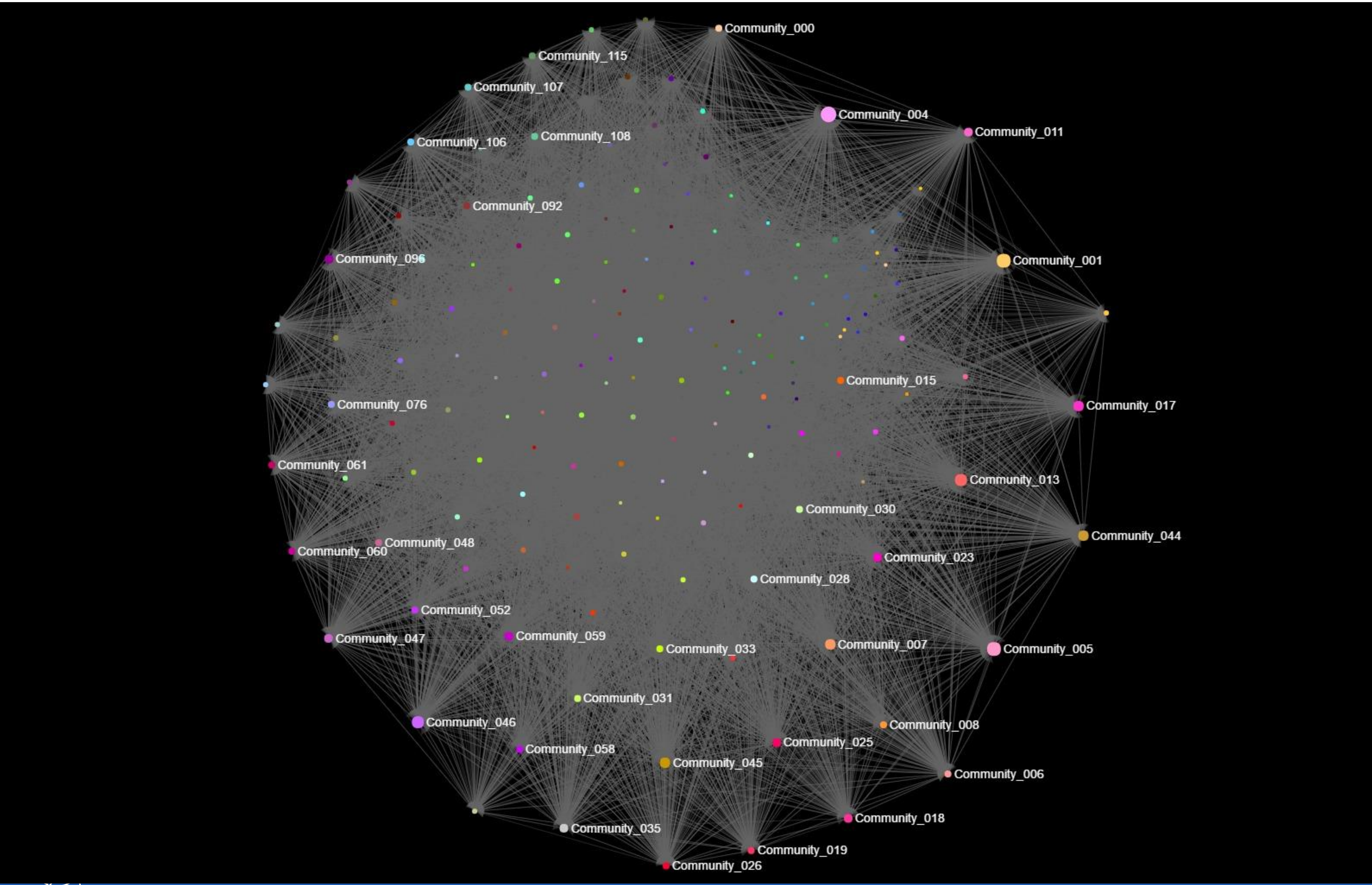
→ g_0 edges
→ g_2 edges
→ g_2 → g_0 connections

Are “type of processing” and “Interactive lobe”
- Vertex attributes?
- Visual dimensions?

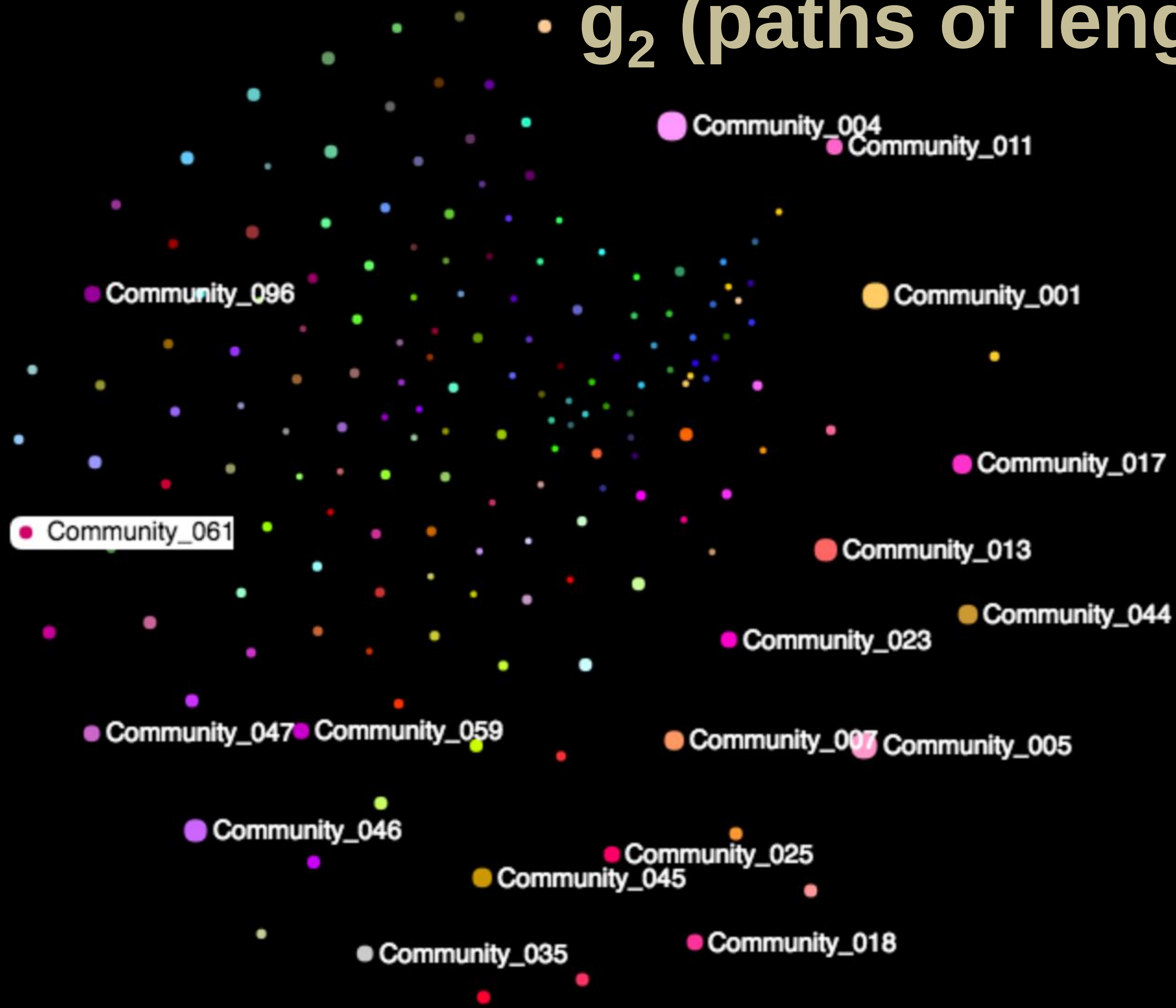
g_0 graph


RI
INS
5
7b
1
SII
4
2
3b
6/SMA
3a
VIP
TF (TEpv?)
46
V4t
DP
PIP
MSTd/p
7a
TEO
STPa
V1
TH-TF
PERI
TEpv
V3
TEpd
TEav
TEad
PO
V4
TEam-a/p
V2
V3A
MT
FEF
STPp
MSTl
VP
FST

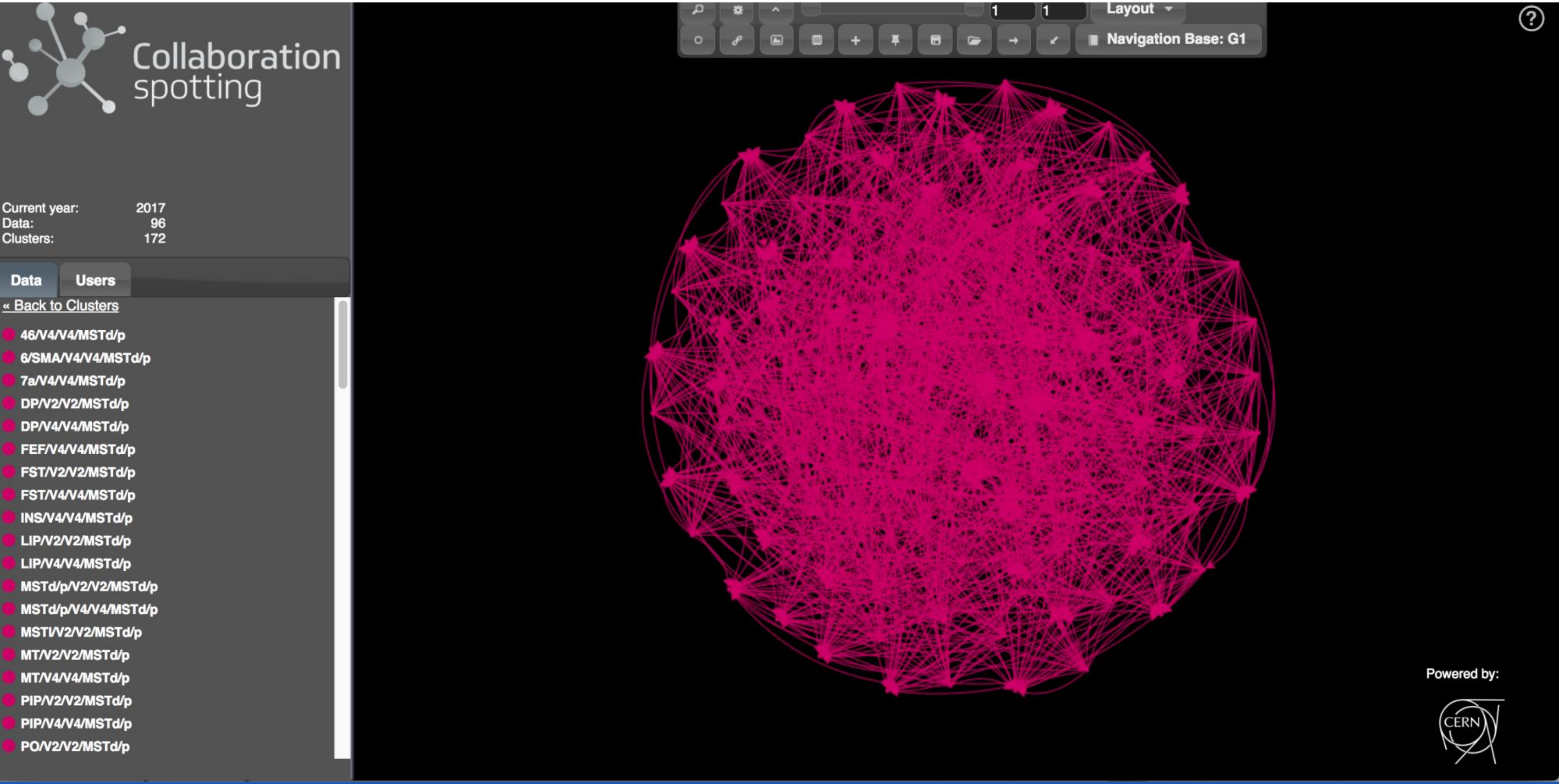
g_2 (with intercluster edges)

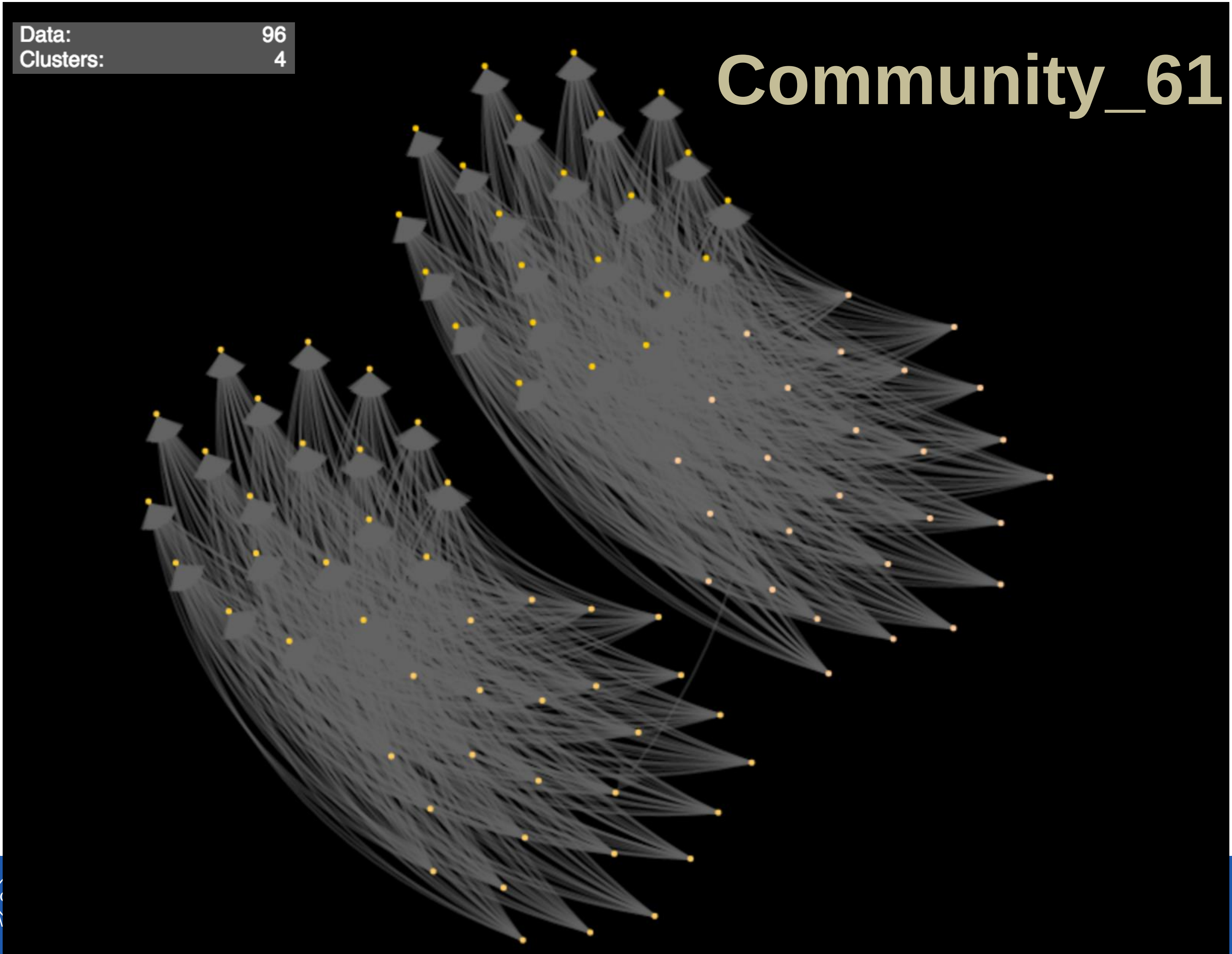


g_2 (paths of length 2

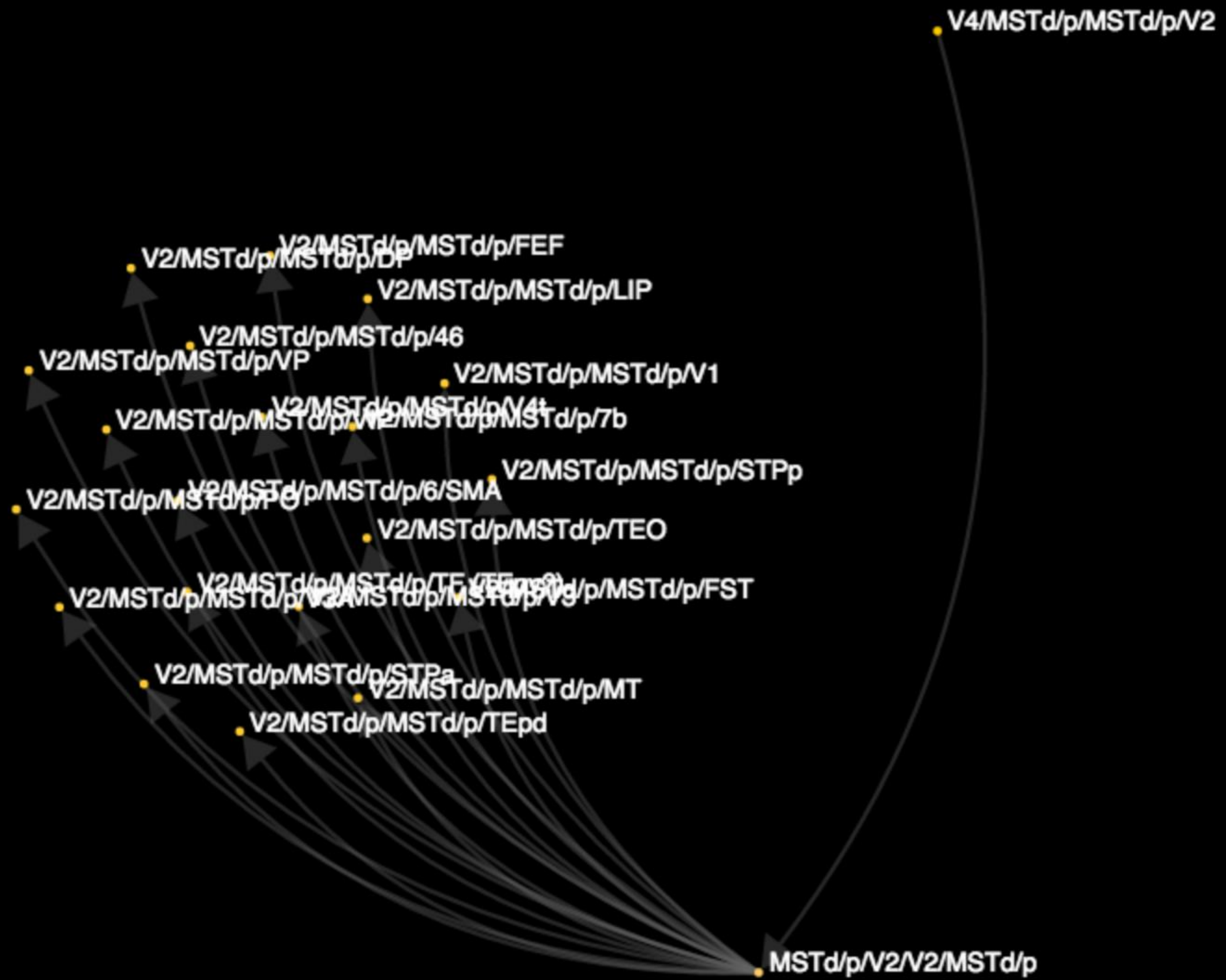


Community_61 Egocentric




Data: 96
Clusters: 4
Community_61

Community_61


V4/MSTd/p/MSTd/p/V2
V2/MSTd/p/MSTd/p/FEF
V2/MSTd/p/MSTd/p/DP
V2/MSTd/p/MSTd/p/LIP
V2/MSTd/p/MSTd/p/46
V2/MSTd/p/MSTd/p/VP
V2/MSTd/p/MSTd/p/V1
V2/MSTd/p/MSTd/p/V4t
V2/MSTd/p/MSTd/p/7b
V2/MSTd/p/MSTd/p/STPp
V2/MSTd/p/MSTd/p/6/SMA
V2/MSTd/p/MSTd/p/PO
V2/MSTd/p/MSTd/p/TEO
V2/MSTd/p/MSTd/p/FST
V2/MSTd/p/MSTd/p/STPa
V2/MSTd/p/MSTd/p/MT
V2/MSTd/p/MSTd/p/TEpd
MSTd/p/V2/V2/MSTd/p

Conclusion

- **To visualize Big Data Analytics output you need**:
  - **Graphs** to store your data networks and their schema
  - **Graphs** to view network structure through selected dimensions
  - **Graphs** to navigate across dimensions to provide contextual data to visualisation tools
- **To maintain visual perception you need to combine various techniques**
  - Statistics, sampling, compound graph, layered graph
- **To support structural and behavioural visualisation you need to explore**
  - Clustering algorithms supporting directed edges
  - Processes, interactions in relation with the data

Thank you for your attention!