Textual Analysis of ICALEPCS and IPAC Conference Proceedings

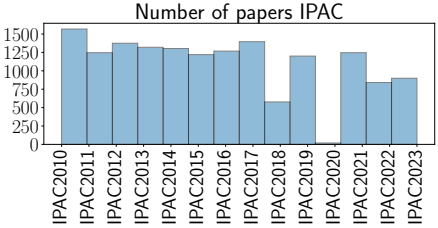
Revealing Research Trends, Topics, and Collaborations for Future Insights and Advanced Search

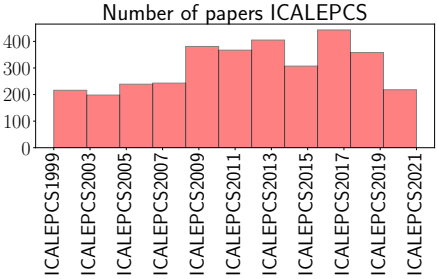
Antonin Sulc, Annika Eichler, Tim Wilksen
Cape Town,

HELMHOLTZ

Introduction

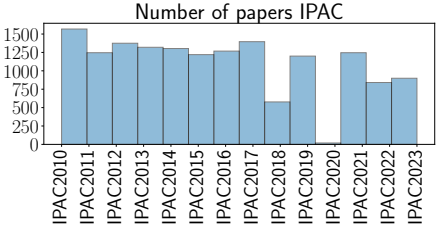
> There are almost 19000 submissions on **IPAC** (15478, 2010 - 2023) and **ICALEPCS** (3375, 1999 - 2021).

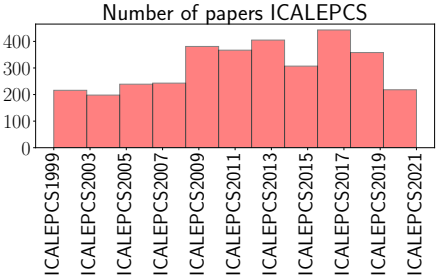




Introduction

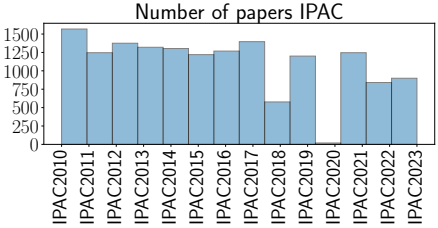
- There are almost 19000 submissions on **IPAC** (15478, 2010 - 2023) and **ICALEPCS** (3375, 1999 - 2021).
- Many of them are highly relevant to the community.

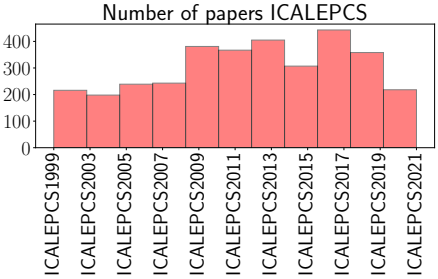




Introduction

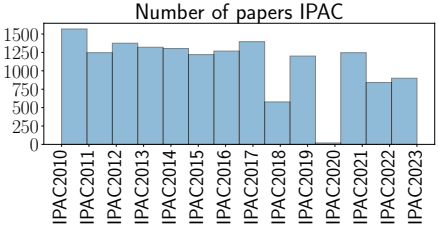
- There are almost 19000 submissions on **IPAC** (15478, 2010 - 2023) and **ICALEPCS** (3375, 1999 - 2021).
- Many of them are highly relevant to the community.
- It is really hard to understand the entire landscape.

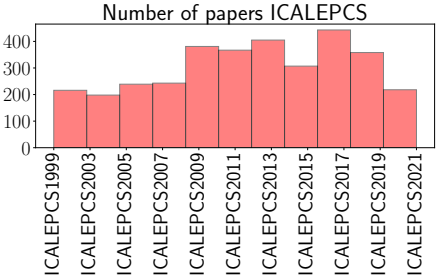




Introduction

- There are almost 19000 submissions on **IPAC** (15478, 2010 - 2023) and **ICALEPCS** (3375, 1999 - 2021).
- Many of them are highly relevant to the community.
- It is really hard to understand the entire landscape.
- Searching through submissions is also a challenge.





Approach

1 Collected the PDFs from public sources (JACOW)

PyMuPDF

Space charge lenses use a confined electron cloud for the focusing of ion beams. The electron density gives the focusing strength whereas...

Search

Topic Modeling

Citation graphs

Bipartite Graphs

Language Processing

Approach

1. Collected the PDFs from public sources (JACOW)
2. Extracted text with `Pymupdf` (can handle two page layout)

PyMuPDF

```
Space charge lenses
use a confined
electron cloud
for the focusing
of ion beams.  The
electron density
gives the focusing
strength whereas...
```

Search

Topic Modeling

Citation graphs

Bipartite Graphs

Language Processing

Approach

1. Collected the PDFs from public sources (JACOW)
2. Extracted text with `Pymupdf` (can handle two page layout)
3. Applied NLP techniques.

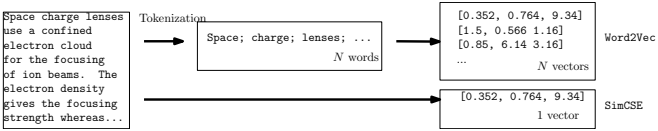
PyMuPDF

Space charge lenses use a confined electron cloud for the focusing of ion beams. The electron density gives the focusing strength whereas...

- Search
- Topic Modeling
- Citation graphs
- Bipartite Graphs
- Language Processing

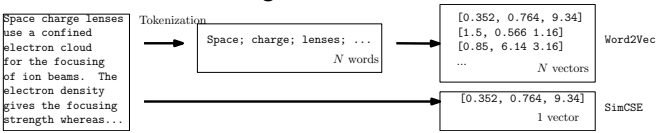
Analyses

1 Search & Embedding

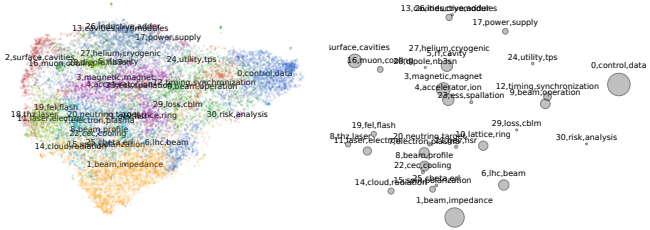


Analyses

1 Search & Embedding

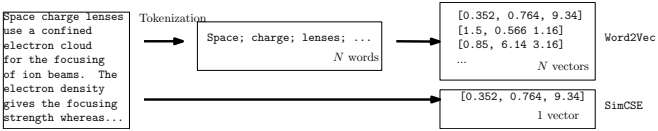


2 Unsupervised topic modeling with `BERTopic`

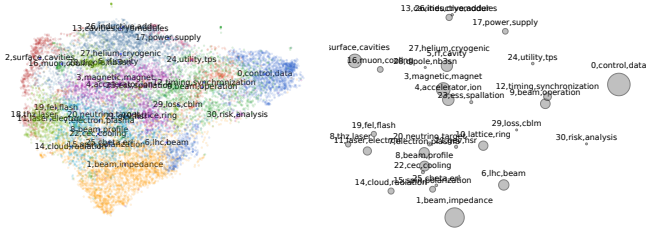


Analyses

1 Search & Embedding



2 Unsupervised topic modeling with `BERTopic`



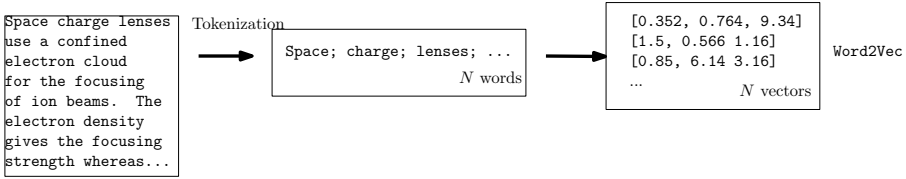
3 Citation graph analysis

4 Document-word bipartite graph analysis

Search

`Word2Vec`

- Each word (token) is represented by one vector.
- Similar **words** have high cosine similarity. (*easy* vs *simple* vs *trivial*)
- One vector per token.
- Keyword search.



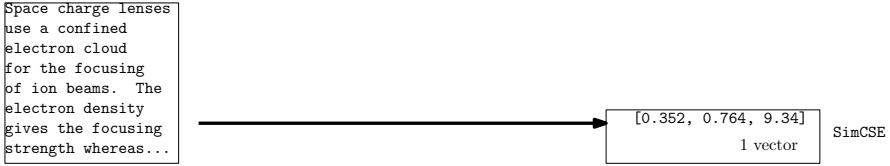
Search

`Word2Vec`

- Each word (token) is represented by one vector.
- Similar **words** have high cosine similarity. (*easy* vs *simple* vs *trivial*)
- One vector per token.
- Keyword search.

`SimCSE`

- The entire text block is one vector.
- Similar **texts** have high cosine similarity.
- One vector per text block (abstract, paragraph).
- Block search (abstract to abstract).



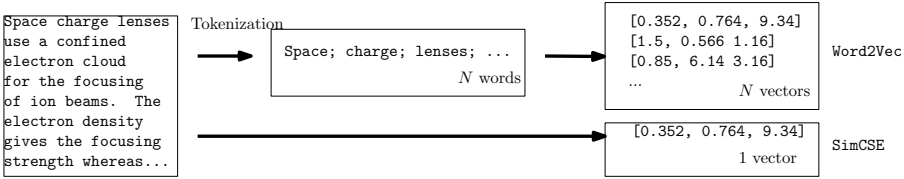
Search

`Word2Vec`

- Each word (token) is represented by one vector.
- Similar **words** have high cosine similarity. (*easy* vs *simple* vs *trivial*)
- One vector per token.
- Keyword search.

`SimCSE`

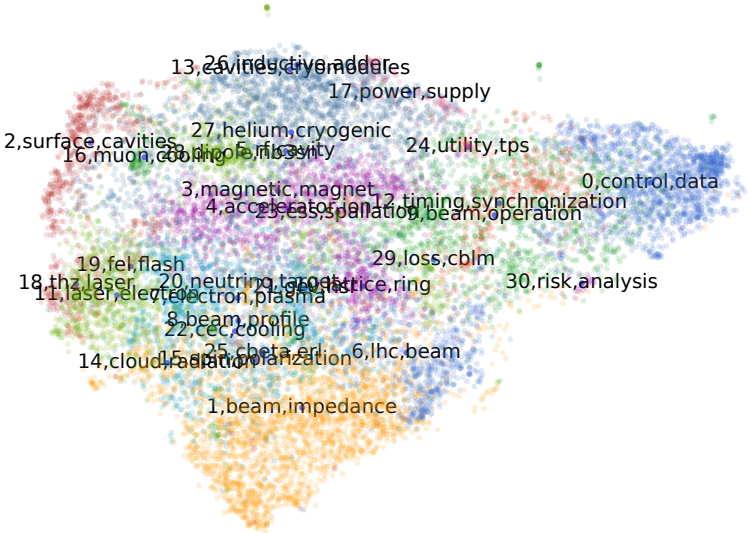
- The entire text block is one vector.
- Similar **texts** have high cosine similarity.
- One vector per text block (abstract, paragraph).
- Block search (abstract to abstract).



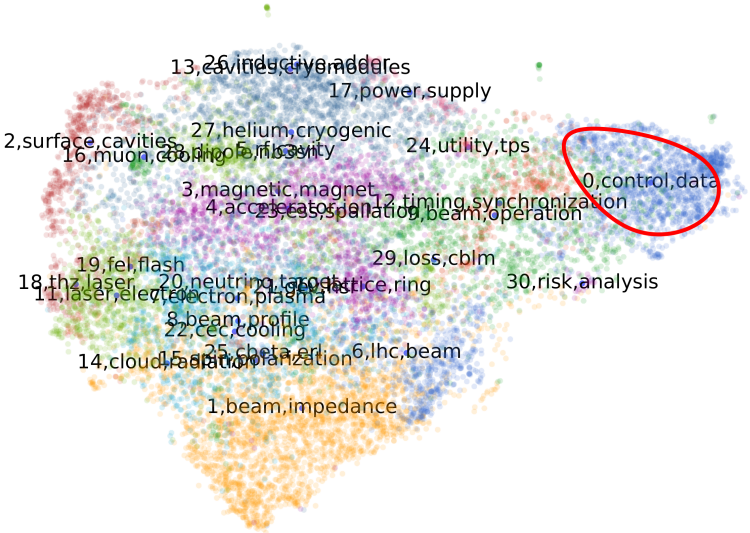
SimCSE Search Example

```
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('TEXT_ICALEPCS/simcse')
texts = ["Biennial International Conference on. Accelerator and Large
Experimental Physics Control Systems", "ICALEPCS"]
e = model.encode(texts)
```

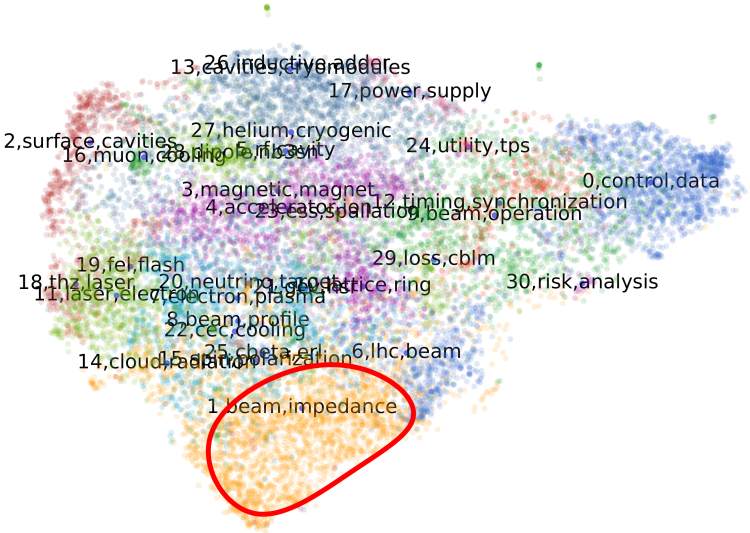
Topic Modeling



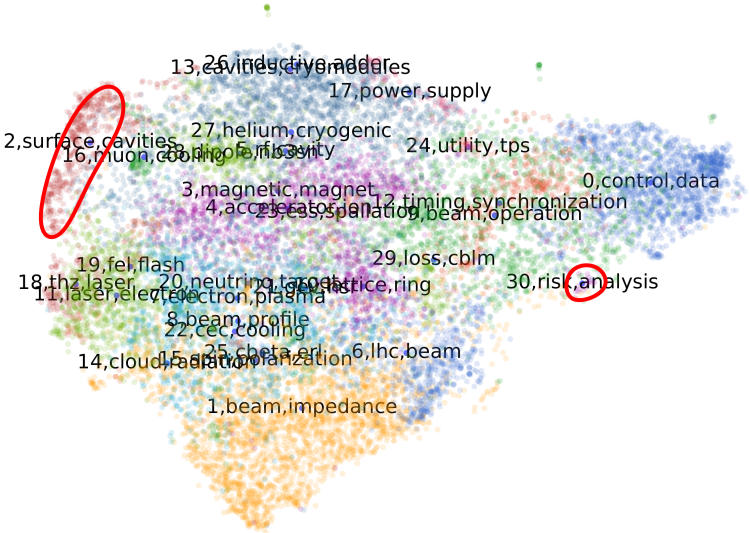
Topic Modeling



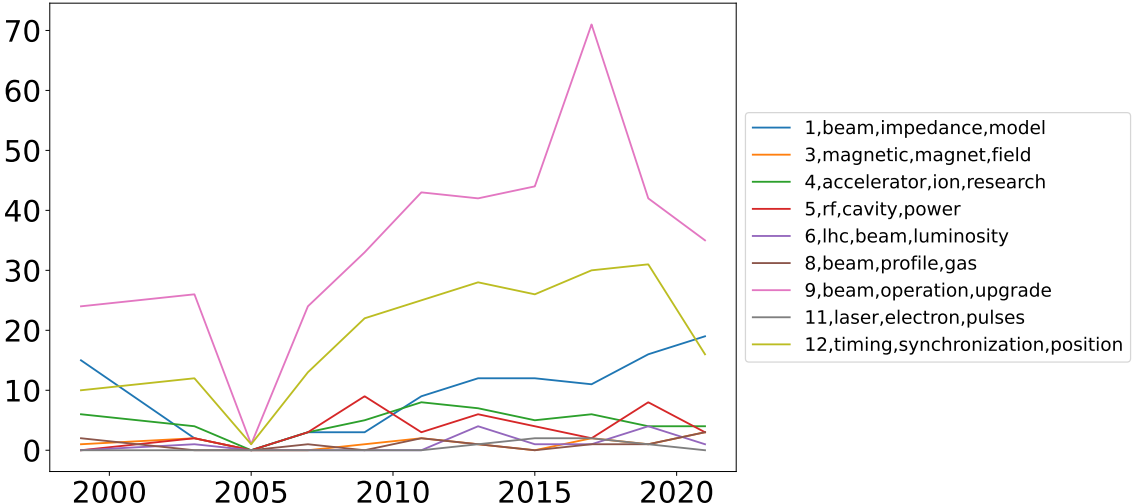
Topic Modeling



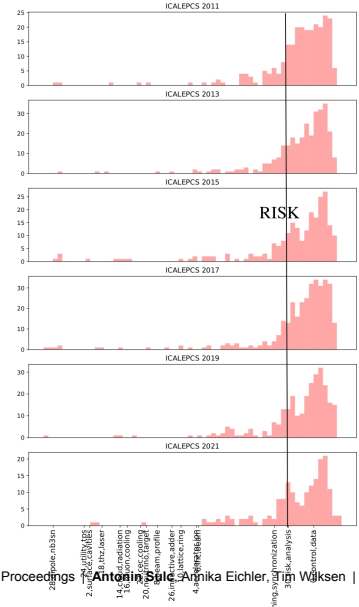
Topic Modeling



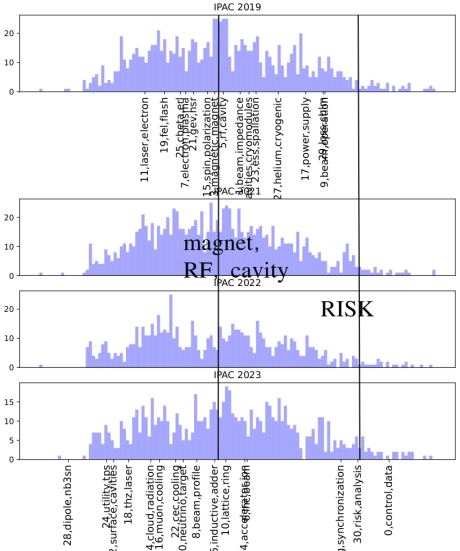
Topic Modeling - ICALEPCS



Topics over Time - ICALEPCS



Topics over Time - IPAC



Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14
3. B. Keil et al., "*The European XFEL Beam Position Monitor System*", IPAC'10

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14
3. B. Keil et al., "*The European XFEL Beam Position Monitor System*", IPAC'10

Betweenness centrality (high overall influence):

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14
3. B. Keil et al., "*The European XFEL Beam Position Monitor System*", IPAC'10

Betweenness centrality (high overall influence):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14
3. B. Keil et al., "*The European XFEL Beam Position Monitor System*", IPAC'10

Betweenness centrality (high overall influence):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. A. Grudiev and W. Wuensch, "*A New Local Field Quantity Describing the High Gradient Limit of Accelerating Structures*", LINAC'08

Citation Graphs - Larger View

IPAC & ICALEPCS citations are too sparse, so we extended sources.
Closeness Centrality (nodes that most efficiently spread information):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. D. Noelle, "*Status of the Standard Diagnostic Systems of the European XFEL*", IBIC'14
3. B. Keil et al., "*The European XFEL Beam Position Monitor System*", IPAC'10

Betweenness centrality (high overall influence):

1. W. Decking et al , "*Status of the European XFEL*", IPAC'13
2. A. Grudiev and W. Wuensch, "*A New Local Field Quantity Describing the High Gradient Limit of Accelerating Structures*", LINAC'08
3. K. Ko and A. E. Candel, "*Advances in Parallel Electromagnetic Codes for Accelerator Science and Development*", LINAC'10

Bipartite Graphs

It's about (maximal) coverage.

Degree Centrality (most connected nodes):

1. L. Medina et al , "*Cavity Control Modelling for SPS-to-LHC Beam Transfer Studies*", IPAC'21
2. M. Chung et al , "*Transient Beam Loading Effects in Gas-filled RF Cavities for a Muon Collider*", IPAC'13
3. K. Yonehara et al , "*R&D of a Gas-Filled RF Beam Profile Monitor for Intense Neutrino Beam Experiments*", IPAC'17

Degree centralities: `beam` (0.07), `cavity` (0.05), `bunch` (0.05), `rf` (0.05), `control` (0.04), `lhc` (0.04), `laser` (0.04), `cavities` (0.04), `fel` (0.04), `electron` (0.04)

Future Work

- We would like to **collect more sources** (arXiv, Inspire HEP).
- Improve the **pre-processing** (equations, tables).
- Feed the **LLM with data**, developing an AI assistant.
- We are **fine-tuning an AI-assistant** based on our logbook.
- Incorporating logbook data together with the community wisdom for better analysis and decisions for e.g.
  - Root Cause Analysis,
  - FAIR (Findable, Accessible, Interoperable, and Reusable)-ness
- We already have some partial success.

Conclusion

- Semantic search, topic modeling, and graph analysis can extract insights from past conferences.
- Topic modeling revealed the evolution of research priorities and emerging trends for each conference over time.
- Citation graph exposed influential papers based on connections to other papers.
- Bipartite graph highlighted important concepts.

Thank you!

https://github.com/sulcantonin/TEXT_ICALEPCS23/

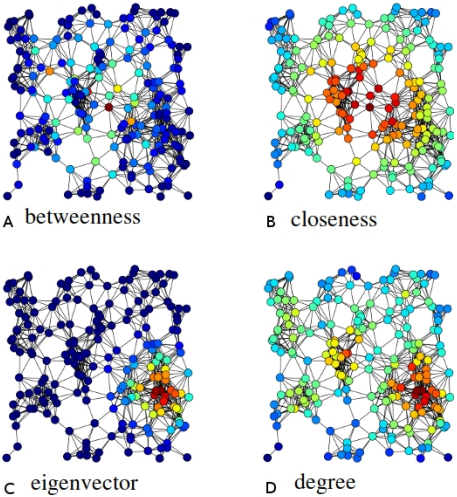
Contact

Deutsches Elektronen-
Synchrotron DESY

Antonin Sulc, Annika Eichler, Tim Wilksen
0000-0001-7767-778X
MCS
antonin.sulc@desy.de

www.desy.de

Extra Slide - Centralities



FAIR

Findablity data should be easy to find for both humans and computers

F**A**IR

Findablity data should be easy to find for both humans and computers
Accesilibity once the user finds the required data, she/he/they need to know how they can be accessed

FAIR

Findablity data should be easy to find for both humans and computers
Accesilibity once the user finds the required data, she/he/they need to know how they can be accessed
Interoperable - data need to interoperate with applications or workflows for analysis, storage, and processing

FAI**R**

Findablity data should be easy to find for both humans and computers
Accesilibity once the user finds the required data, she/he/they need to know how they can be accessed
Interoperable - data need to interoperate with applications or workflows for analysis, storage, and processing
Reusable - data should be well-described so that they can be replicated and/or combined in different settings