Menu

Go to

- PART1
- PART2
- PART3

Community Detection

Louvian Algorithm

	Community	Size	Top Authors
0	Community 1	116129	Xing Zhou, Tao Fang, Lap-Pui Chau, Cyrus Shahabi, Shu-Yuen Didi Yao, Roger Zimmermann, Huaguang Zhang, Zhenwei Liu, Guang-Bin Huang, Zhanshan Wang...

	Community	Size	Top Authors
1	Community 2	66702	Genevi eve Paquin, Laurent Vuillon, Simonetta Balsamo, Gian–Luca Dei Rossi, Andrea Marin, Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, Hector Garcia-Molina...
2	Community 3	45662	Archana K. Singh, Hideki Asoh, Jocelyn Y. K. Aulin, Djordje Jeremic, Antoine Bossard, Keiichi Kaneko, Shietung Peng, Elon Rimon, Hiro Takahashi, Takeshi Kobayashi...
3	Community 4	34310	Efraim Laksman, Håkan Lennerstad, Pietro Andreani, Ahmed Bader, Eylem Ekici, Daniele D. Giusto, Maurizio Murroni, Giulio Soro, Emilie Bosc, Patrick Le Callet...
4	Community 5	29796	Olof Olsson, Philippe Hanhart, Touradj Ebrahimi, Elisabeth André, Peter Ingwersen, Chu-Ren Huang, Daniel Ferrés, Horacio Rodríguez, Hossein Sameti, Hisami Suzuki...

Conclusion

Louvain community detection algorithm has identified clusters of authors based on their co-authorship relationships.Each "Community" represents a group of authors who are more(densely connected to each other than to the rest of the network.

Spectral clustering groups authors into distinct research communities based on their collaborations.

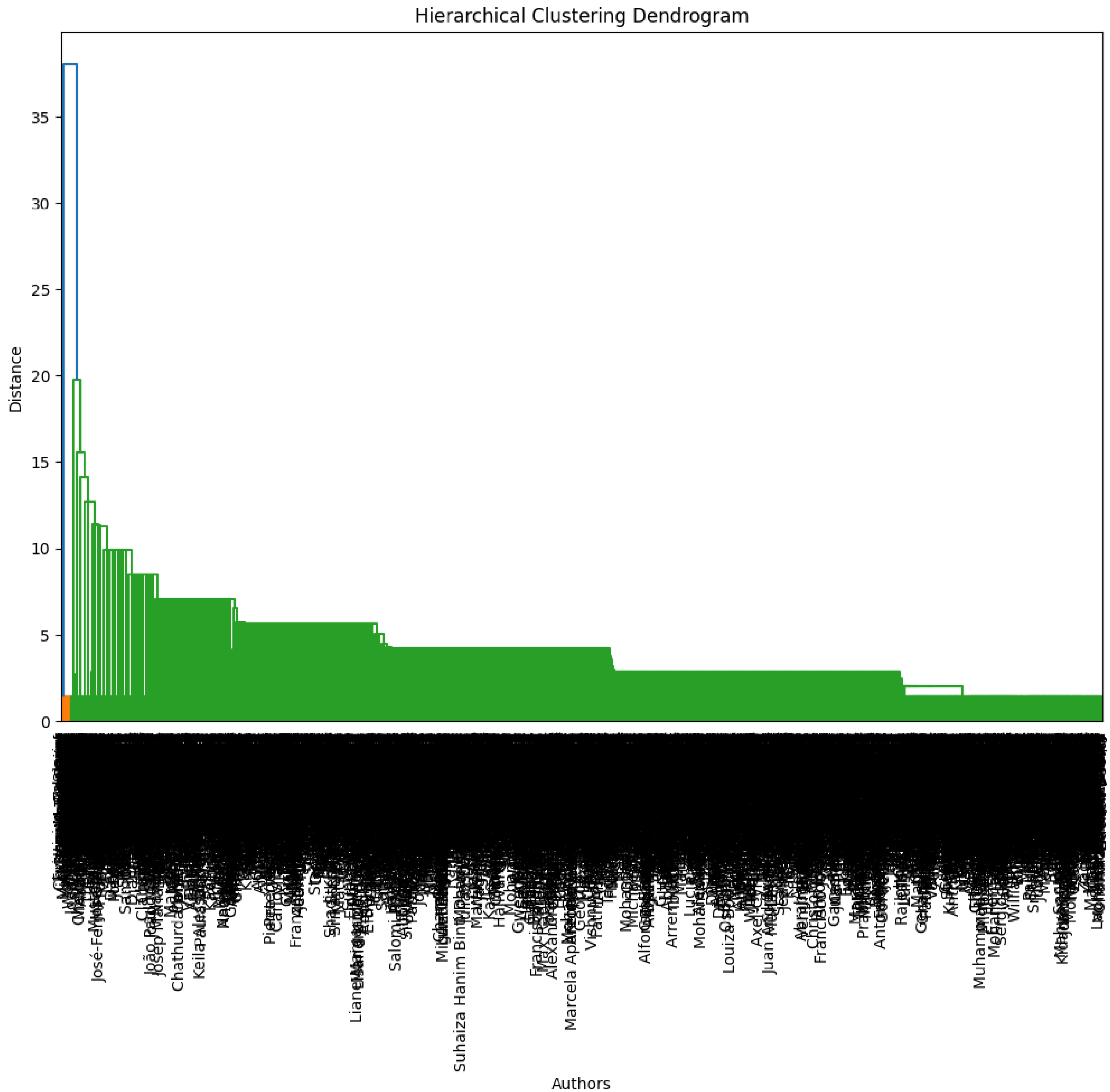
Spectral Clustering Communities

	Community	Size	Top Authors
0	Community 1	11081	Maria G. Koziri, Panos Papadopoulos, Nikos Tziritas, Antonios N. Dadaliaris, Thanasis Loukopoulos...
1	Community 2	920	Artur Zawadzki, Marek Gorgon, Chia-Yen Chen, Radim Sara, Chih-Lin Chi...
2	Community 3	921	Arber Murturi, Burak Kantarci, Sema Oktug, Nicholas R. Howe, Andreas Fischer...
3	Community 4	1013	Jiancong Luo, Ishfaq Ahmad, Yu Sun, Rajesh M. Krishnaswamy, Kumar N. Sivarajan...
4	Community 5	885	J. Jesu Vedha Nayahi, V. Kavitha, Deepika Saini, Sanjeev Kumar, Fawang Liu...

Hierarchical Clustering Communities

	Community	Size	Top Authors
0	Community 1	2938	Maria G. Koziri, Panos Papadopoulos, Nikos Tziritas, Antonios N. Dadaliaris, Thanasis Loukopoulos...
1	Community 2	15	Cheryl H. Porter, Chris Villalobos, Dean P. Holzworth, Roger Nelson, Jeffrey W. White...
2	Community 3	11	Seydou Traoré, Abdrahamane Anne, Aly Khalifa, S. Bosomprah, F. Caroline...

	Community	Size	Top Authors
3	Community 4	28	Rolf Apweiler, Terri K. Attwood, Amos Marc Bairoch, Alex Bateman, Ewan Birney...
4	Community 5	12	Lambert Schaelicke, John B. Carter, Wilson C. Hsieh, Mark R. Swanson, Lixin Zhang...


Hierarchical Clustering Dendrogram
Distance
0
5
10
15
20
25
30
35
Authors

dendogram

Louvain Clustering Coefficient: 0.6703

Spectral Clustering Coefficient: 0.5786

Hierarchical Clustering Coefficient: 0.9597

Clustering Algorithm Comparison:

- Louvain: 0.6703
- Spectral: 0.5786
- Hierarchical: 0.9597

Best Clustering Algorithm Based on Clustering Coefficient: Hierarchical

Louvain clustering is optimized for modularity and often performs well in large-scale networks.

A 0.6703 coefficient suggests that the detected communities have a moderate level of internal connectivity.

Spectral clustering operates by leveraging the eigenvalues of the adjacency matrix.

A coefficient of 0.5786 is lower than Louvain, indicating that the detected clusters are less tightly connected than those found by Louvain.

Hierarchical clustering groups nodes based on similarity in a recursive manner.

A 0.9597 coefficient suggests that the communities detected are very strongly connected, meaning the hierarchical structure effectively captures natural clusters.

If the dataset size increases significantly, Louvain might still be preferred due to its scalability, while Hierarchical Clustering could become computationally expensive.

Naming the Communities

These keywords have been automatically extracted using BERT-based embeddings, meaning they are not just random words but contextually important terms found in the paper.The extracted keywords successfully summarize each paper’s topic.They provide a quick way to understand what a paper is about without reading the full abstract. The keywords highlight core research areas, making it easier to categorize, cluster, or search for similar papers.

Combined Keywords

	ID	Title	Combined Keywords
0	987231	Slice-based parallelization in HEVC encoding...	alternative hevc, hevc, hevc using, paralleli...
1	79954	MIDAS: A Middleware for Information Systems wi...	middleware information, qos concerns, isolati...
2	567130	Automatically controlled pan-tilt smart camera...	analysis camera, smart camera, camera device,...
3	500891	Particle Filter for Visual Tracking Using Mult...	particle filter, tracking using, visual track...

Community Keywords

	Community	Combined Keywords
0	Community 0	matrix multiplication, hevc using, matrix powers, adaptation distributed, overhead hevc, parallel algorithms, scalable parallel, parallelization hevc, based parallelization, video coding...
1	Community 1	middleware information, qos concerns, isolation concurrency, middleware, midas middleware, storage engine, systems qos, database servers, minimizes transactions, concurrency...
2	Community 2	analysis camera, smart camera, camera fpga, camera device, fpga based, tracking moving, tilt smart, controlled camera, position camera, orientation camera...
3	Community 3	particle filter, tracking using, visual tracking, tracking, bayesian filtering, visual track, track ing, cameras overlapping, multiple cameras...
4	Community 4	crowdsourcing service, crowdsourcing development, service, workflow crowdsourcing, mobile crowdsourcing, reference model, crowdsourcing services, crowdsourcing, model crowdsourcing...

Associated Community

	ID	Paper ID	Title	Community
0	987231	3e5ef05b-bb89-4717-bfc3-74a467529ded	Slice-based parallelization in HEVC encoding...	0
1	79954	653894e3-b581-412c-ae3f-71d267b0ea9d	MIDAS: A Middleware for Information Systems wi...	1
2	567130	0ba8b19f-9659-453f-8deb-50b87e26e41f	Automatically controlled pan-tilt smart camera...	2

Community Keywords

	ID	Paper ID	Title	Combined Keywords
0	987231	3e5ef05b-bb89-4717-bfc3-74a467529ded	Slice-based parallelization in HEVC encoding...	matrix multiplication, hevc using, matrix pow...
1	79954	653894e3-b581-412c-ae3f-71d267b0ea9d	MIDAS: A Middleware for Information Systems wi...	middleware information, qos concerns, isolati...
2	567130	0ba8b19f-9659-453f-8deb-50b87e26e41f	Automatically controlled pan-tilt smart camera...	analysis camera, smart camera, camera fpga...
3	500891	f12c84ba-82a9-4835-b5fa-5e6715b1b6b3	Particle Filter for Visual Tracking Using Mult...	particle filter, tracking using, visual track...
4	55399	5d197014-1fbb-40d8-8223-378dea0d4a30	A reference model for crowdsourcing as a service	crowdsourcing service, crowdsourcing developm...
5	135049	7790685d-525b-4f64-bd58-12d8cea683c8	A Landweber algorithm for 3D confocal microsco...	microscopy, deconvolution introduced, landweb...
6	733378	08f4b403-7400-48d6-981b-6e76424ac967	Computing group cardinality constraint solutio...	minimization logistic, constraint, cardinalit...

Aggregation:

The goal is to assign relevant keywords to each community, which can be used to understand the main research topics within those communities.

Each research paper is associated with one or more communities based on its authors. Since authors are clustered into research communities, a paper written by multiple authors might belong to multiple communities. To keep track of this relationship, we create a mapping where each paper ID is linked to the communities it belongs to.

For example:

- Paper A is written by authors from Community 0 and Community 1.
- Paper B is written by authors from Community 1 and Community 2.

This mapping ensures that each paper contributes to the research themes of all its associated communities.

Each paper contains a title and an abstract, which are rich sources of information about its topic. To extract meaningful keywords from these text fields, we use KeyBERT, a model that identifies the most important words and phrases in a given text.

For example, a paper titled 'Advances in Deep Learning for Natural Language Processing' might have extracted keywords like:

- Deep Learning
- Natural Language Processing
- Neural Networks

These keywords represent the core research areas of the paper. Once the keywords for a paper are extracted, they are distributed to all the communities the paper is associated with. This ensures that each community gets insights into the topics covered by the papers written by its authors.

For instance:

- Paper A has keywords: 'Deep Learning', 'AI', 'Neural Networks'
- Paper A belongs to Community 0 and Community 1.
- These keywords are added to both Community 0 and Community 1.

Since multiple papers in the same community might have overlapping keywords, we use a set to store keywords instead of a list. A set automatically removes duplicate entries, ensuring that each keyword appears only once per community.

For example:

If Community 1 receives keywords from multiple papers:

```
{"Deep Learning", "AI", "Neural Networks", "AI", "Deep Learning"}
```

The final set for Community 1 will be:

```
{"Deep Learning", "AI", "Neural Networks"}
```

	Community	Keywords
0	Community 0	matrix multiplication, hevc using, matrix powers, adaptation distributed, overhead hevc, parallel algorithms, scalable parallel, parallelization hevc, based parallelization, video coding...
1	Community 1	middleware information, qos concerns, isolation concurrency, middleware, midas middleware, storage engine, systems qos, database servers, minimizes transactions, concurrency...

	Community	Keywords
2	Community 2	analysis camera, smart camera, camera fpga, camera device, fpga based, tracking moving, tilt smart, controlled camera, position camera, orientation camera...
3	Community 3	particle filter, tracking using, visual tracking, tracking, bayesian filtering, visual track, track ing, cameras overlapping, multiple cameras...
4	Community 4	crowdsourcing service, crowdsourcing development, service, workflow crowdsourcing, mobile crowdsourcing, reference model, crowdsourcing services, crowdsourcing, model crowdsourcing...
5	Community 5	microscopy, deconvolution introduced, landweber algorithm, 3d confocal, deconvolution algorithm, microscopy deconvolution, deconvolution, iterative deconvolution, microscopy restoration, confocal microscopy...
6	Community 6	minimization logistic, constraint, cardinality constraint, features weighted, restoring contrast, contrast restoration, degraded images, fog degraded, constraint weighted, fog...
7	Community 7	auditory cortex, cortex deep, speech reconstruction, human auditory, deep neural...

Assign a name

Each community has a set of unique keywords extracted from the papers authored by its members.

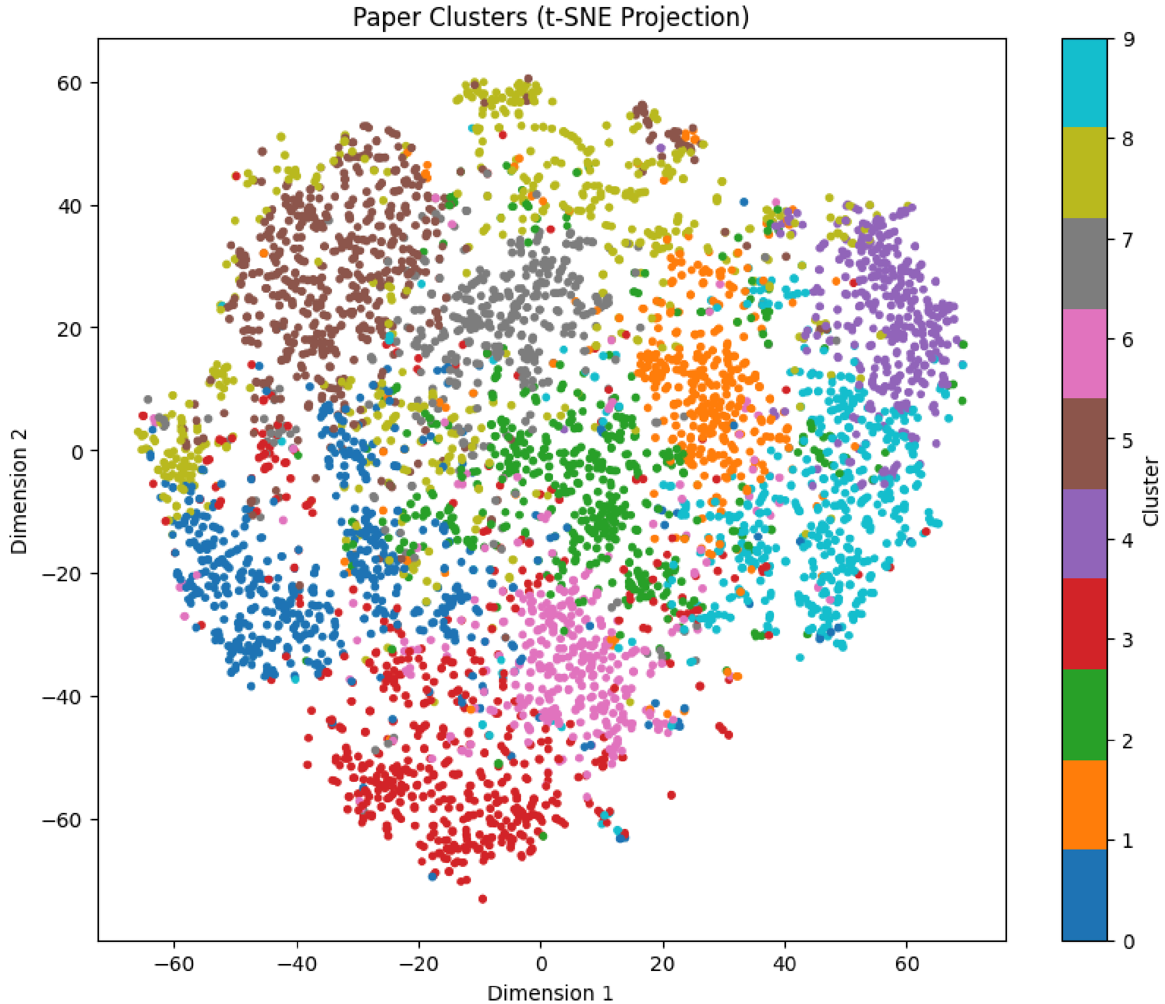
The most frequently occurring key terms represent the central research themes of the community.

The keywords are grouped into broader categories such as Machine Learning, NLP, Computer Vision, etc.

These topics are aligned with known research domains.

	Paper ID	Keywords
0	987231	matrix multiplication
1	79954	middleware information
2	567130	analysis camera
3	500891	particle filter
4	55399	crowdsourcing service
5	135049	microscopy
6	733378	minimization logistic
7	732057	auditory cortex

	Paper ID	Keywords
8	51333	infection



clustring

I used the K-Means clustering algorithm for clustering the paper embeddings.

Each point represents a research paper.

Spatial proximity between points indicates semantic similarity (papers that are closer together are more related).

Clusters are well-formed, indicating a meaningful separation of research topics.

Some overlapping regions suggest interdisciplinary research areas or potential improvements in clustering methodology.

Larger clusters may correspond to broad topics, while smaller clusters could indicate specialized fields.

Conclusion:

DBI measures cluster compactness and separation.

Lower values indicate better clustering because it means that clusters are well-separated and compact.

A DBI score of 4.2225 is quite high, meaning:

- The clusters might be overlapping too much.
- The clusters might not be well-separated.

Silhouette Score measures how similar points are within their cluster vs. other clusters.

Range:

- +1 → Perfect clusters (tight & well-separated)
- 0 → Overlapping clusters
- -1 → Bad clustering (points assigned to the wrong cluster)

A Silhouette Score of 0.0314 is extremely low, indicating:

- The clusters overlap significantly.
- K-Means might not be the best algorithm for your data.

Jaccard Similarity measures how similar two sets are (intersection over union).-0.0117 : Extremely low overlap, meaning does not align well with venues.