Digital Transformation in the Shipping Industry: a Network-Based Systematic Review

Andreas Pittas[a], Yannes Filippopoulos[a], Zoran Lajic[b], Luca Ferrarini[a,*]

[a] *Department of Information Technologies, University of Limassol, Limassol, Cyprus*
[b] *Department of Energy Efficiency, Angelicoussis Group, Athens, Greece*

Abstract

The shipping industry is undergoing a profound digital transformation, driven by advancements in automation, artificial intelligence, blockchain, and the Internet of Things (IoT). These technologies enhance operational efficiency, optimize supply chain management, and improve sustainability by reducing emissions and fuel consumption. However, navigating this digital revolution requires a structured understanding of emerging trends, challenges, and opportunities. A network-based systematic review serves as a crucial methodological approach for researchers, enabling them to synthesize existing knowledge, identify research gaps, and develop informed strategies to leverage digital transformation effectively. By critically analyzing co-citation and co-authorship networks, modeling topics over time, and performing trend analysis, we gain insights on the current status of digital transformation within the shipping industry, ultimately guiding industry stakeholders and researchers ....XXXXXXXXXXXXXx Our results show that XXXXXXXXXXXXXXXXXXXXXX

Keywords: digital transformation, shipping industry, systematic literature review, complex networks

1. Introduction

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

*Corresponding author

2. Literature Review

3. Methodology

In this section we describe the methodology we followed for the data collection and analysis. Fig. [XXX] shows the overall methodology discussed in this section. Results and implications are discussed in further sections.

3.1. *Keyword identification and data collection*

We asked experts in the shipping industry to identify the most relevant keywords related to the industry itself and to digital technologies and digital transformation. Their analysis resulted in 35 keywords, listed in Table 1.

[Table 1 about here.]

Data was collected from three research engines: EBSCO (Vaughan, 2011), ProQuest (Cooke, 2017), and IEEE eXplore (Wilde, 2016). The search was performed on October the 22nd 2024. For each engine, we retrieved scientific articles containing any of the digital transformation related keywords and any of the shipping industry related keywords, in either their title or abstract. The exact query for each engine are available on request. We limited our results using the following criteria: *a* only English literature, and *b* only scientific contributions published in peer-reviewed journals. Table 2 shows the results.

[Table 2 about here.]

All search engines provided the digital object identifier for the articles. This allowed us to screen the resulting set and identify 2324 unique articles for the subsequent analysis. One challenge of using different data engines is the variety of attributes they return for each article. In order to have the same information for each article, we queried a fourth search engine for all the 2324 articles. We chose OpenAlex (Priem et al., 2022), which has been shown to be suitable for bibliometric analysis (Alperin et al., 2024). Our final result set comprised 2293 scientific publications.

3.2. Descriptive statistics

We started our analysis evaluating descriptive statistics across our article set. More specifically, we calculated:

1. the distribution of the number of publications per year;
2. the distribution of publications across authors, identifying the most prolific authors;
3. the distribution of publications across institutions, identifying the research centers with the highest number of publications;
4. the distribution of publications across countries.

3.3. Co-authorship network analysis

As a second step, we built and analyzed the network of co-authorship. Network analysis was performed in Python, using the NetworkX package (Hagberg et al., 2008). We identified 7723 distinct authors. We built the network using authors as nodes, and setting bi-directional links between them if there existed at least one publication that they co-authored. For each link, we stored within the graph object information about the authors institutes and countries for further analysis.

To determine which distribution best fit the data, we run statistical tests comparing the likelihood of power-law distribution against the exponential distribution, the log-normal distribution, and the truncated power-law distribution.

Next, we focused on the largest connected component of the network, made of 883 authors and 2753 links between them. The choice of focusing on the largest component was dictated mostly by computational limitations.

Working on the largest component, we applied the Louvein community (Blondel et al., 2008) algorithm to identify the major communities of authors and investigated the distribution of institutions and countries across communities.

To conclude, we analyzed the network for small-world behavior. More specifically, we calculated both the clustering coefficient and the average path length and compared them to random networks of equivalent size.

3.4. Co-citation network analysis

We built a co-citation network of nodes (i.e., articles) and links (i.e. co-citation between two articles). The resulting graph had 1298 nodes. The degree distribution was tested for power-law characteristics against other plausible distributions (exponential, log-normal, and truncated power-law).

Next, we identified the most influential articles (i.e., the top 10 in terms of received citations). Our goal was to check if the most cited articles were literature reviews. As presented in the following section, this turned out not to be the case, allowing us to draw relevant considerations over the demand of SLRs at the conjunction of digital transformation and shipping industry.

We then moved our attention to the top 20% cited papers and analyzed their topics. To achieve this, we create a sub-network using only the top 20% cited papers and applied the Louvein community algorithm (Blondel et al., 2008). Next, for each community collected the titles and applied natural language processing (NLP) to model their topics (BERTTopic (Paul et al.)).

To conclude, we applied different centrality measures to the top 20% graph to identify the 5 most relevant articles. These were analyzed more in details in terms of covered research area, as a preliminary trend analysis, further developed in our next and last analysis section.

3.5. Thematic analysis

Working on the entire set of articles (2290) we performed a thematic analysis to identify the major topic of research. We pre-processed the titles with the following steps:

1. lemmatization to transform words into their root forms;
2. removal of stop-words;
3. removal of non alpha-numeric text.

Next, we applied tokenization and embedded each title using BERT (Devlin et al., 2018). The resulting vectors were analyzed for unsupervised clustering. More specifically, we adopted two methods to identify the ideal number of clusters: the Calinski-Harabasz index (Caliński and Harabasz, 1974), and the Davies-Bouldin index (Davies and Bouldin, 1979). Having identified the best number of cluster, we applied the unsupervised K-means algorithm and calculated the centroid for each cluster. Next, we identified for each cluster the 10 articles closest to the corresponding centroid and applied BERTTopic to extract the common themes.

We concluded our thematic analysis by building two word clouds. Using both titles and abstracts from all articles, we applied the TF-IDF algorithm to each word and use it as weight when building the clouds. The first cloud was built over the entire set of words in titles and abstracts, while the second cloud was built after removing all shipping related terms (hence focusing on the digital technologies only).

4. Results

In this section we present the results of our analysis. We then discuss them in the next section.

4.1. Descriptive statistics

Figure 1 shows the distribution of articles across years. Altough the first publications are dates as back as the 1960s, only from the year 2005 we witness an increasing interest in the effects of digital transformation within the shipping and maritime industry. The number of publication increased minimally and not steadily between 2005 and 2015. From 2015 onwards, we witness an exponential increase in the number of publications. After reaching a peak in 2023, the trend seem to have stabilized. Considering that our data was collected at the end of October 2024, we can reasonably argue that the year 2024 has not witnessed a signficant increase of publication, compared to the previous year.

[Figure 1 about here.]

Figures 2 show the top 20 authors, the top 20 institutes, and the top 20 countries in terms of number of publications. Considering the authors, we note how the 0.03% of all authors in our cohort (20 out of 7723) cover over 2.9% of the total publications, suggesting a skewed distriution of publications across authors. When looking at the top insititutes, we see they cover over 21% of the total publications (see Table 3), while the top 5 countries cover up to 50% of total publications (see Table 4). Looking deeper into the top insitute, one can notice how many of those Universityies have strong historical bindings with the sea. Consider, as examples, the Dalian Maritime University, the Shangai Maritime Univerity, and the Delft Technical University. Similarly, looking at the most rapresentative countries one can see they all have strong maritime industry and economy.

[Figure 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

4.2. Co-authorship network analysis

The degree distribution of the co-authorship network seems to follow a power-law curve (see Fig. 3. However, several distributions may present similar curves. To establish which is the best fitting model we run statistical tests. We run statistical tests, calculating the log-likelihood and p-value between different pairs. The power-law distribution was significantly more accurate fit than the exponential one ($p < 0.01$). However, the comparison between power-law and truncated power-law distributions, as well as the one between power-law and log-normal distributions, did not lead to significantly different results ($p=0.32$ and $p=0.39$ respectively). The results confirm the heavy-tail characteristic of the degree distribution (which holds true for both log-normal and power-law), but without further indicate the possible nature of such heavy tail (Mitzenmacher, 2004; Higaki et al., 2020; Liu et al., 2021; Smith, 2021).

[Figure 3 about here.]

As a second step in our co-authorship network analysis, we identified the largest component of the network (made of 2753 authors), and identified its main communities, using the Louvain algorithm (Blondel et al., 2008). We identified 28 communities (see Fig. 4), and map on them the distribution of institutions and countries linked to the authors. Results highlight a high level of international collaborations within each community, as well as a high level of national collaborations (within the same country). This can be seen in Figure 5, where we show the number of different countries and institutions per community.

Furthermore, our network analysis does not show and closed cluster of collaborations. Communities are all well inter-connected, suggesting that the niche nature of this field (i.e., digital transformation in shipping) leads global actors to collaborate extensively in advancing research. In Figure [XXX] and Figure [XXX] we show the chord charts for both country and institution mapping on the co-authorship communities.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

Lastly, we evaluated the small-world properties of the co-authorship network. To do so, we calculated both clustering coefficient and average path length, and compare them with equivalent random networks. Our results show a higher clustering coefficient (0.83 vs 0.007) and a higher average path (7.1 vs. 3.9). In a proper small-world topology, one would expect high clustering coefficient and small average path. Our results, instead, suggest that communities are strongly locally organized, but somehow lack efficiency in cross community communication.

4.3. Co-citation network analysis

The co-citation network was analyzed for its largest connected component (made of 1298 articles). The results on the degree distribution are similar to those we obtained for the co-authorship network. More specifically, the statistical comparison between degree distribution excluded an exponential distribution ($p < 0.05$), and did not favor a power-law distribution against log-normal or truncated power-law distribution ($p=0.06$ and $p=0.9$ respectively). Figure 8 shows the degree distribution.

[Figure 8 about here.]

Using the degree distribution, we identified the most influential articles (i.e. top 10 articles with the highest number of co-citation). Table 5 shows such influential works. One can see that amongh the most influential works we find SLRs and bibliometric studies, supporting the relevance of such publicaitons within the industry.

[Table 5 about here.]

Next, we created a sub-graph considering only the 20% most cited articles (257 nodes). We identified the communities using the Louvain algorithm and perform a topic analysis on the titles of the articles per community. Table 6 reports the main topics for each of the 7 communities we identified, while Figure 9 show the color-mapped communities.

[Figure 9 about here.]

[Table 6 about here.]

Finally, we adopted several centrality measures to identify the most relevant articles. More specifically, we identified the 5 top articles for five different centrality measures: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and page rank. We union the results and identified 10 relevant articles for further analysis. We then looked more in details to the theme covered in these articles to extrapolate relevant research areas the field. Results are shown in Table 7.

[Table 7 about here.]

4.4. Thematic analysis

For the thematic analysis, we considered all titles from the 2290 articles. After having pre-processed, tokenized, and vectorized all titles, we used clustering methods to identify the best number of clusters for the vectors. More specifically, we used the Calinski-Harabasz index and the Davies-Bouldin index. The curves are shown in Figure 10. All indexes pointed to 8 ideal clusters. For each cluster, we calculated the centroid and then selected the 10 closest vectors (i.e. articles) to each centroid. Focusing on their titles, we highlighted the main themes for each cluster. Results are shown in Table 8.

[Figure 10 about here.]

[Table 8 about here.]

Next, we built words cloud for our articles. In this case, we used both title and abstract words. A first word cloud was built using all words after pre-processing them. The second word cloud was built after remoing shipping related key words. This allowed us to focus on technical key words for the second word cloud. Both word clouds were based on TF-IDF analysis and are shown in Figure 11 and Figure 12.

[Figure 11 about here.]

[Figure 12 about here.]

As a last analysis, we focused on the concept tags reported by OpenAlex for each paper. OpenAlex organized topics in a tree-like structure which is a modification of the one produced by (Shen et al., 2018). Specifically, there are 19 root-level concepts: engineering, computer science, business, economics, philosophy, mathematics, political science, medicine, psychology, environmental science, geology, geography, chemistry, physics, biology, sociology, art, history, and materials science. We have collected the root-level concepts related to the publications in out study and plot the number of corresponding papers over time (see Figure 13).

[Figure 13 about here.]

The predominant top-level concept over time are engineering, computer science, and business. We next focused on the second and third level of concepts of Open-Alex, limiting our search to those having as parents either engineering, computer science, or business. Among those, we selected the 10 most relevant and showed their evolution over time in a heatmap (see Figure 14).

[Figure 14 about here.]

5. Discussion

6. Conclusion

Summarize key findings and future work.

References

Alperin, J.P., Portenoy, J., Demes, K., Larivière, V., Haustein, S., 2024. An analysis of the suitability of openalex for bibliometric analyses. arXiv preprint arXiv:2404.17663 .

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008, P10008.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3, 1–27.

Cooke, R., 2017. Proquest ebook central. The Charleston Advisor 19, 39–43.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence , 224–227.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Hagberg, A., Swart, P.J., Schult, D.A., 2008. Exploring network structure, dynamics, and function using NetworkX. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

Higaki, A., Uetani, T., Ikeda, S., Yamaguchi, O., 2020. Co-authorship network analysis in cardiovascular research utilizing machine learning (2009–2019). International Journal of Medical Informatics 143, 104274.

Liu, P., Jiang, L., Liu, H., 2021. Structural and evolutionary analysis of coauthorship network in scientometrics, in: 2021 International Conference on Education, Information Management and Service Science (EIMSS), IEEE. pp. 514–519.

Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions. Internet mathematics 1, 226–251.

Paul, P.C., Rahman, M., Begum, A., Ahmed, M.T., Chakraborty, D., Rahman, M.S., . Combining bert with lda: Improved topic modeling in bengali language .

Priem, J., Piwowar, H., Orr, R., 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 .

Shen, Z., Ma, H., Wang, K., 2018. A web-scale system for scientific knowledge exploration. arXiv preprint arXiv:1805.12216 .

Smith, K.M., 2021. Explaining the emergence of complex networks through log-normal fitness in a euclidean node similarity space. Scientific Reports 11, 1976.

Vaughan, J., 2011. Ebsco discovery services. Library Technology Reports 47, 30–38.

Wilde, M., 2016. Ieee xplore digital library. The Charleston Advisor 17, 24–30.

List of Figures

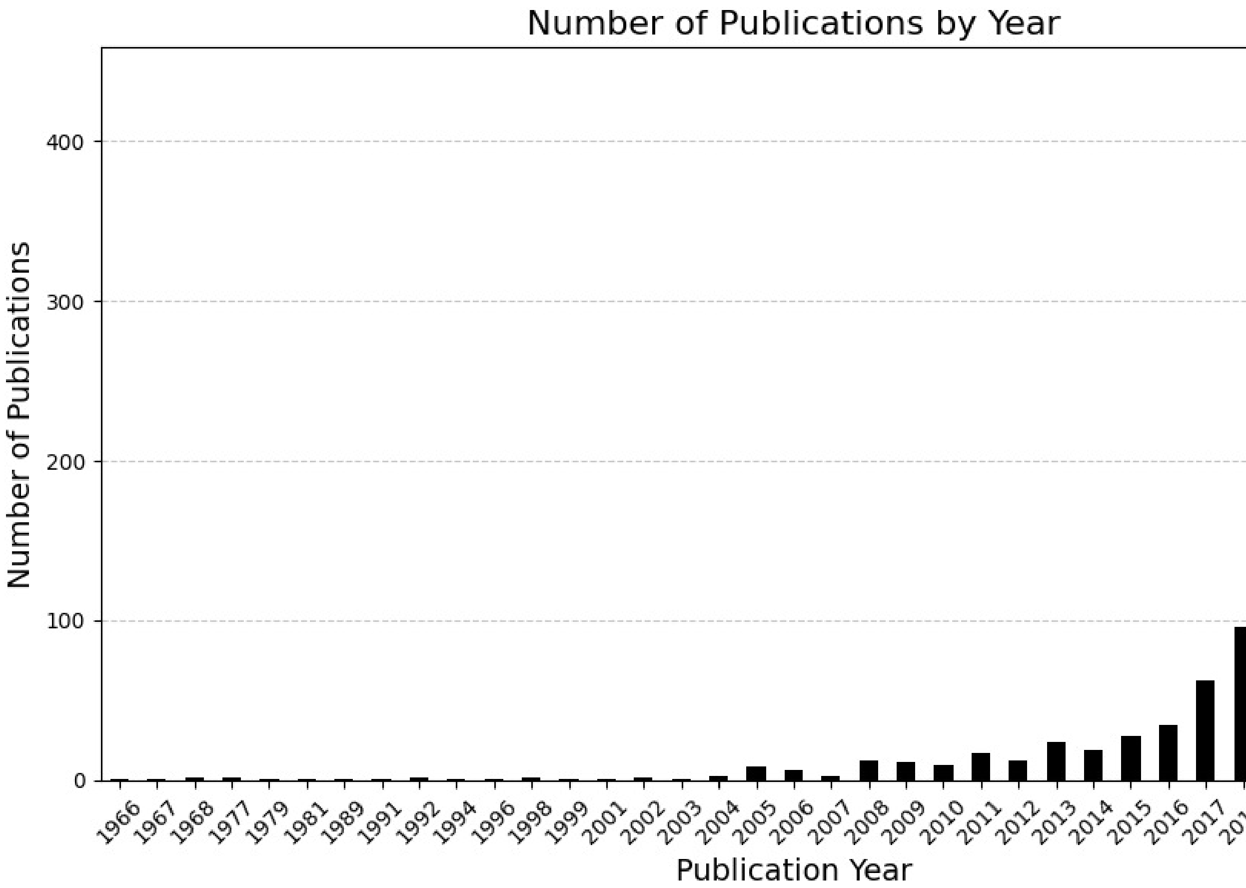
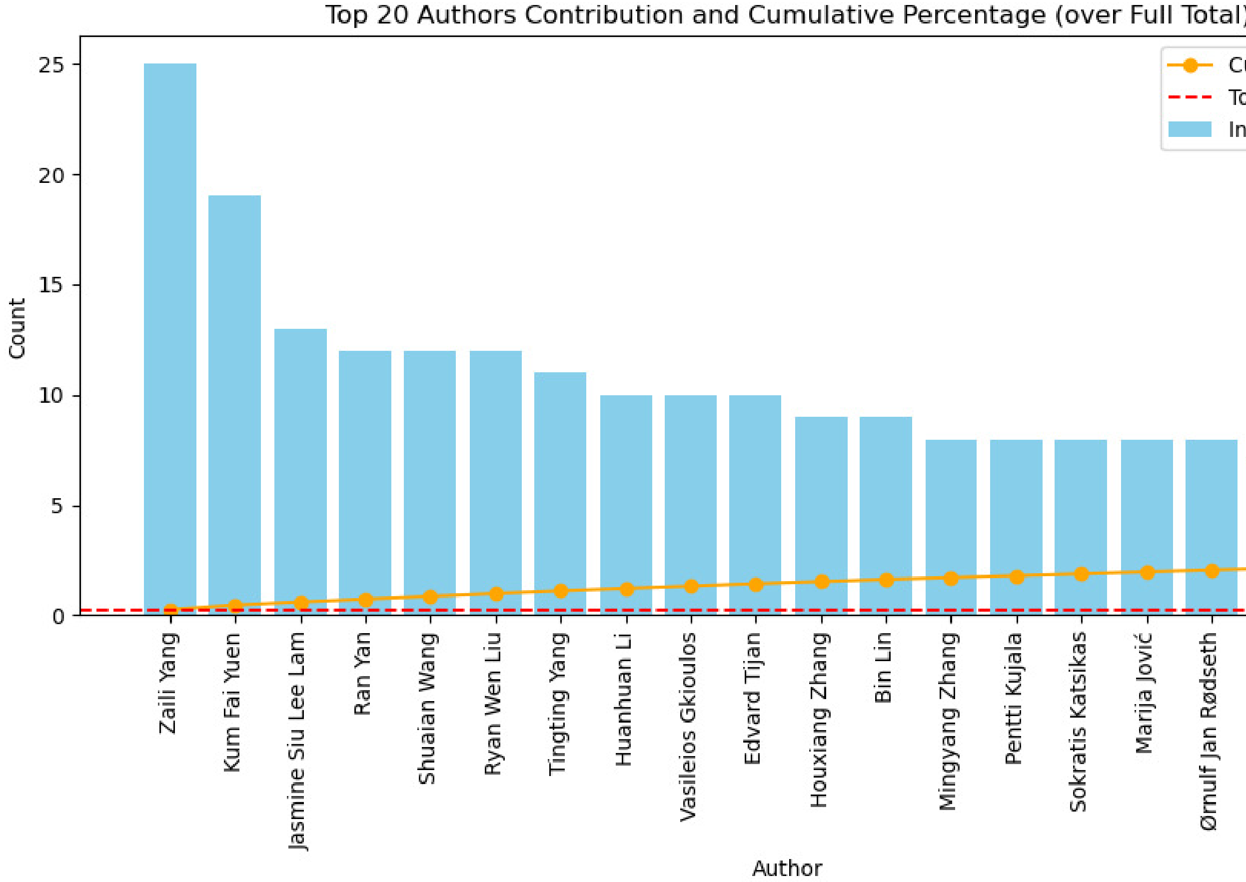
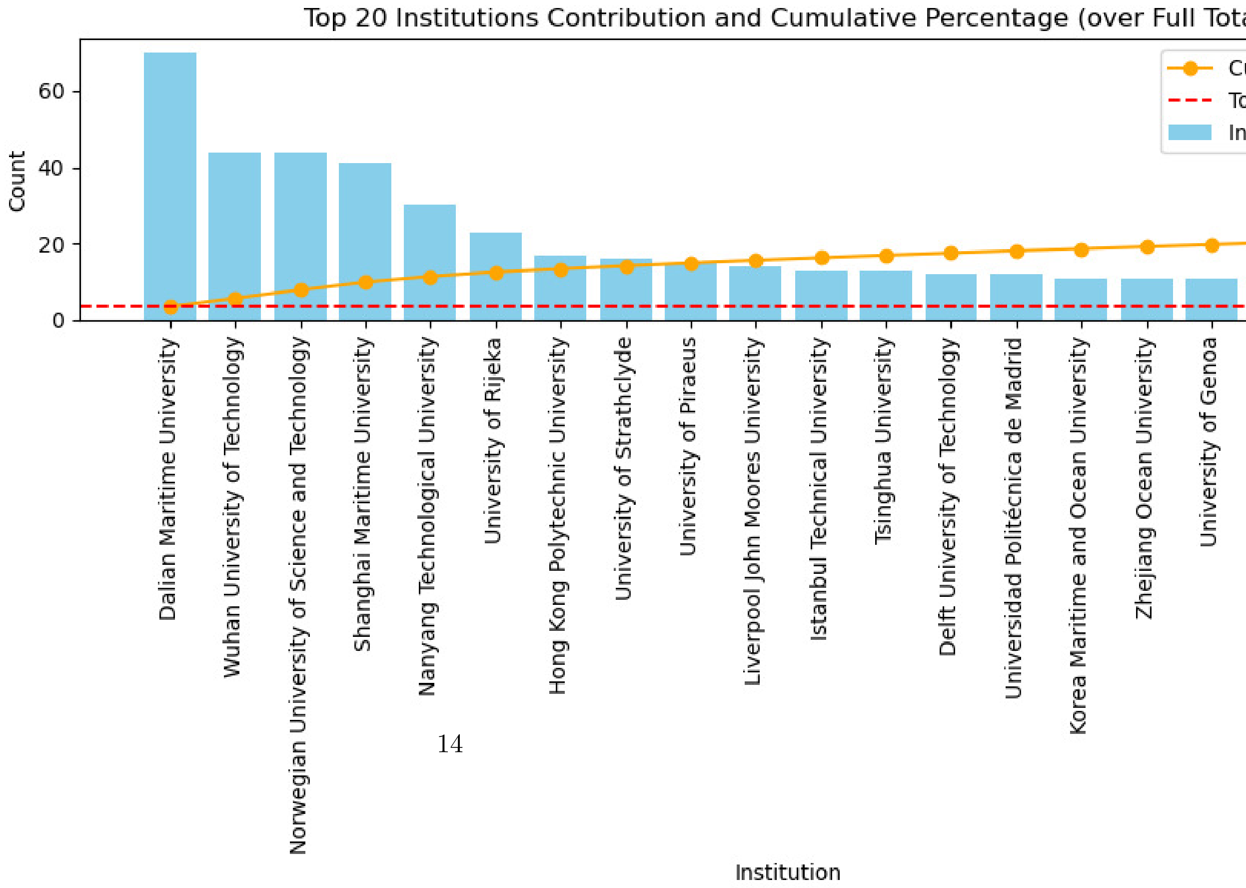


Figure 1: Distribution of publications across years


Top 20 Authors Contribution and Cumulative Percentage (over Full Total)
Count
25
20
15
10
5
0
Zaili Yang
Kum Fai Yuen
Jasmine Siu Lee Lam
Ran Yan
Shuaian Wang
Ryan Wen Liu
Tingting Yang
Huanhuan Li
Vasileios Gkioulos
Edvard Tijan
Houxiang Zhang
Bin Lin
Mingyang Zhang
Pentti Kujala
Sokratis Katsikas
Marija Jović
Ørmulf Jan Rødseth
Author


Top 20 Institutions Contribution and Cumulative Percentage (over Full Tota
Count
60
40
20
0
Dalian Maritime University
Wuhan University of Technology
Norwegian University of Science and Technology
Shanghai Maritime University
Nanyang Technological University
University of Rijeka
Hong Kong Polytechnic University
University of Strathclyde
University of Piraeus
Liverpool John Moores University
Istanbul Technical University
Tsinghua University
Delft University of Technology
Universidad Politécnica de Madrid
Korea Maritime and Ocean University
Zhejiang Ocean University
University of Genoa
Institution


Top 20 Countries Contribution and Cumulative Percentage (over Full Tota
500

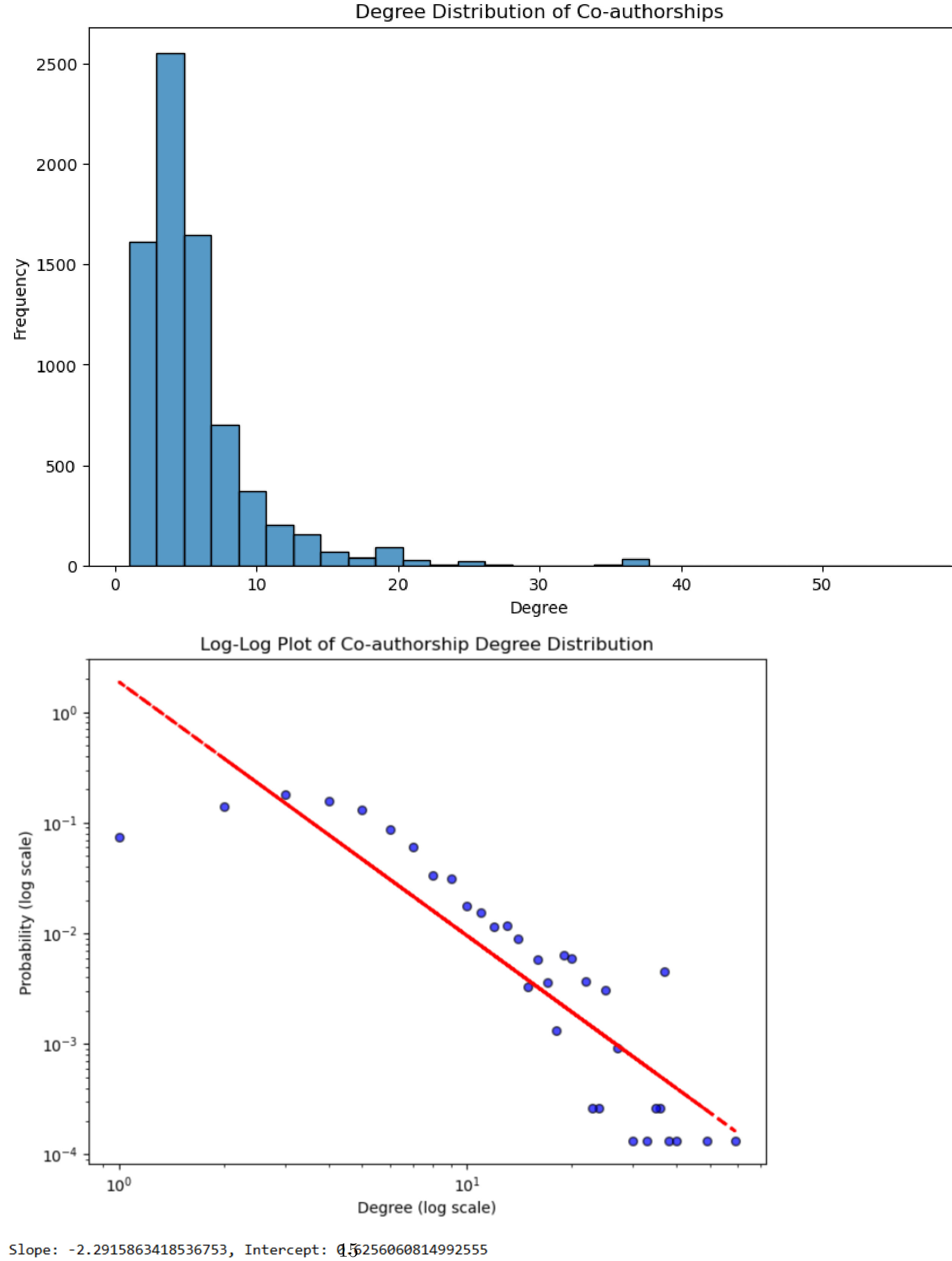


Figure 3: Co-authorship degree distribution (*top*), and corresponding log-log chart (*bottom*)

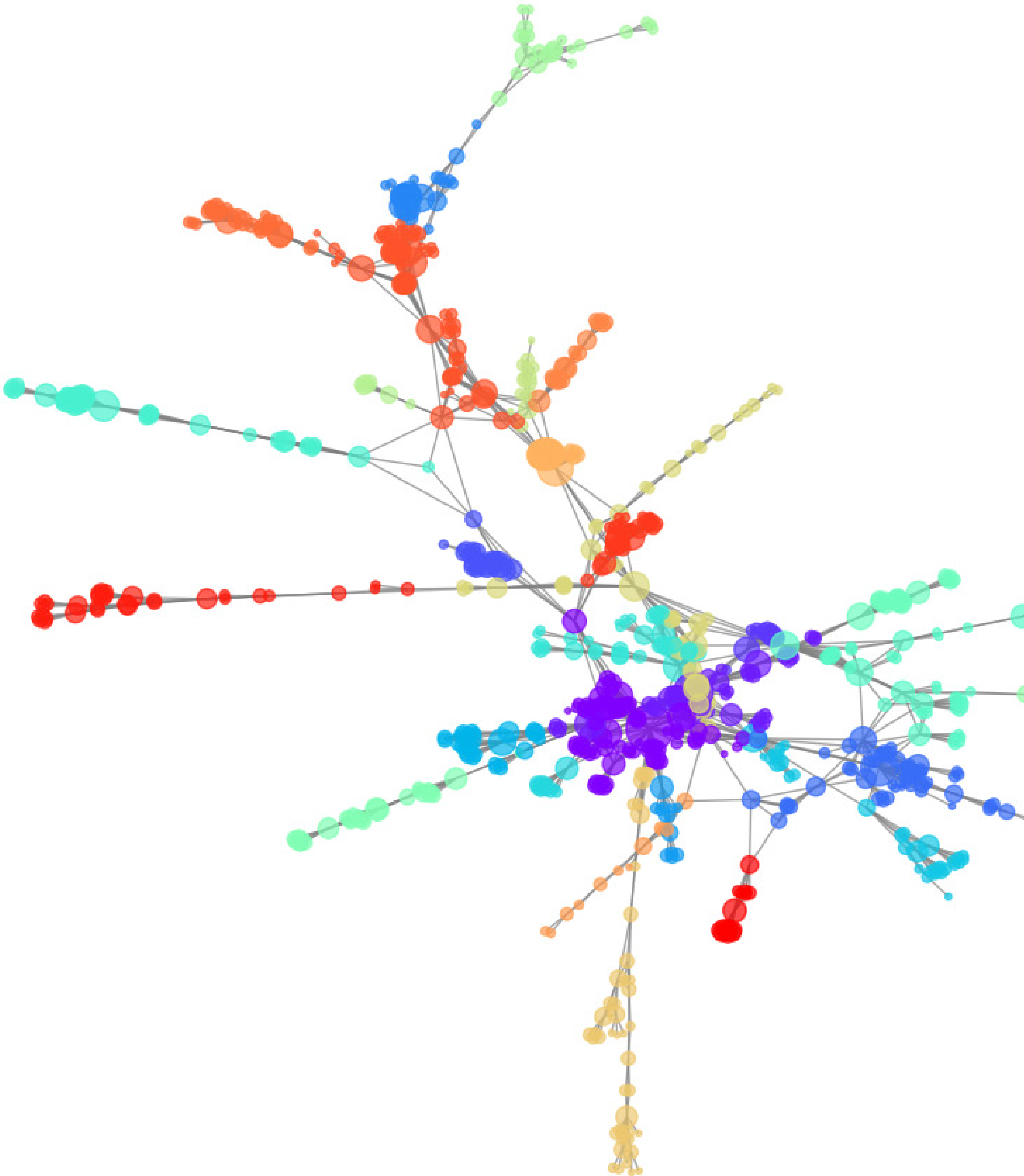


Figure 4: Co-authorship network with communities

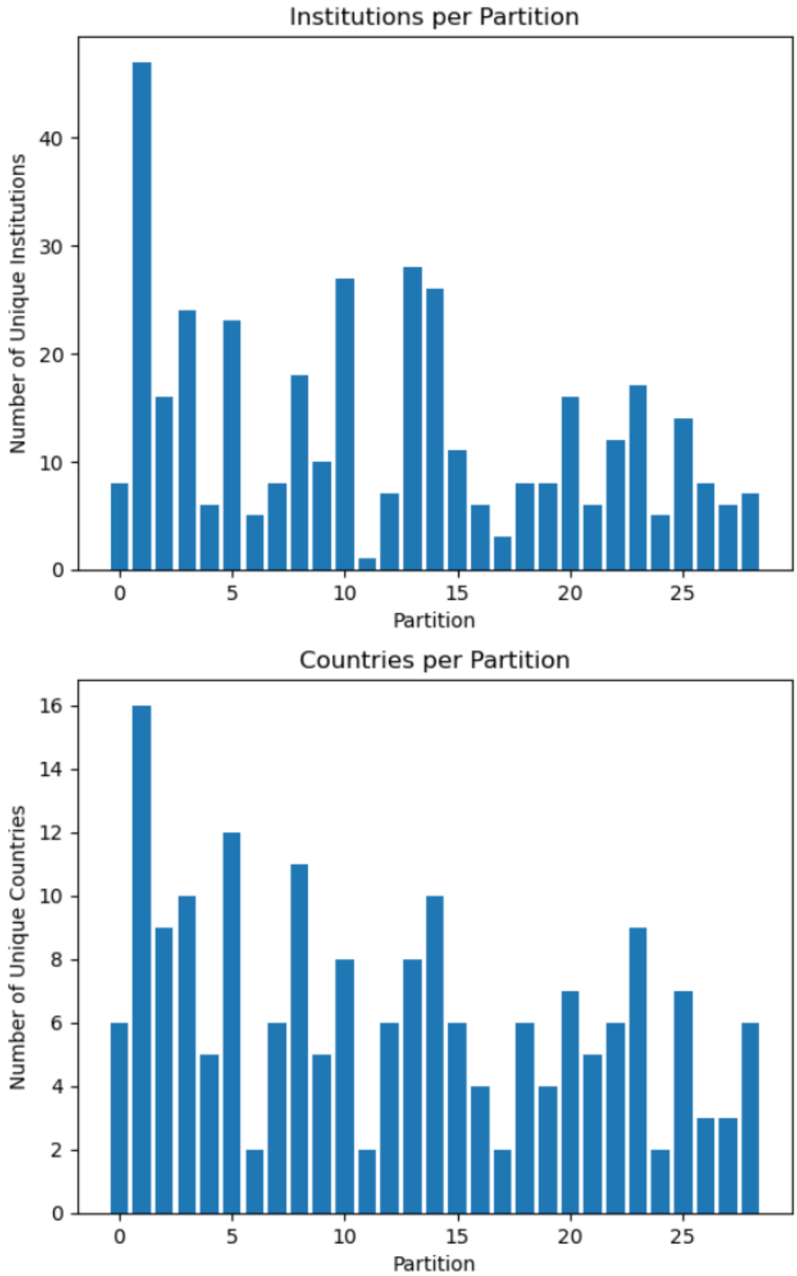
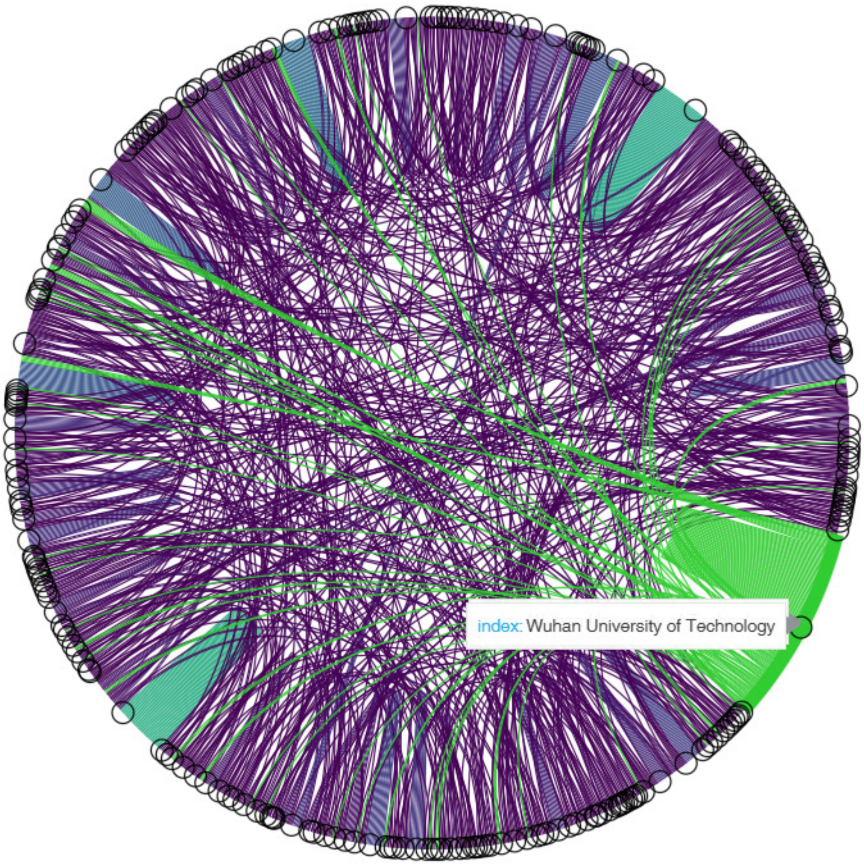
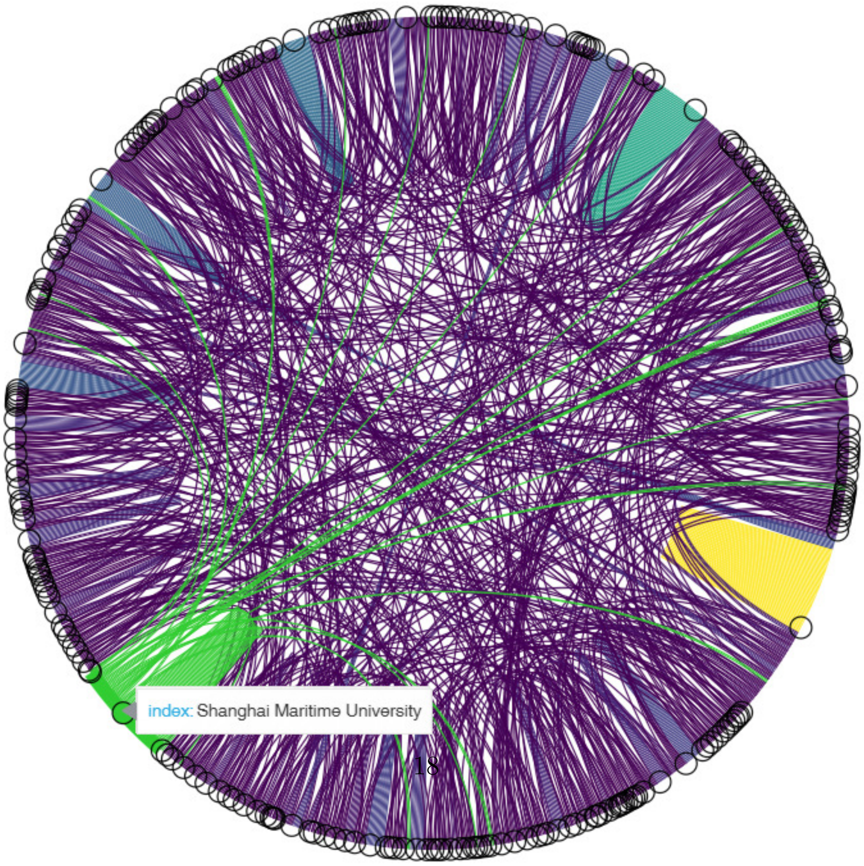


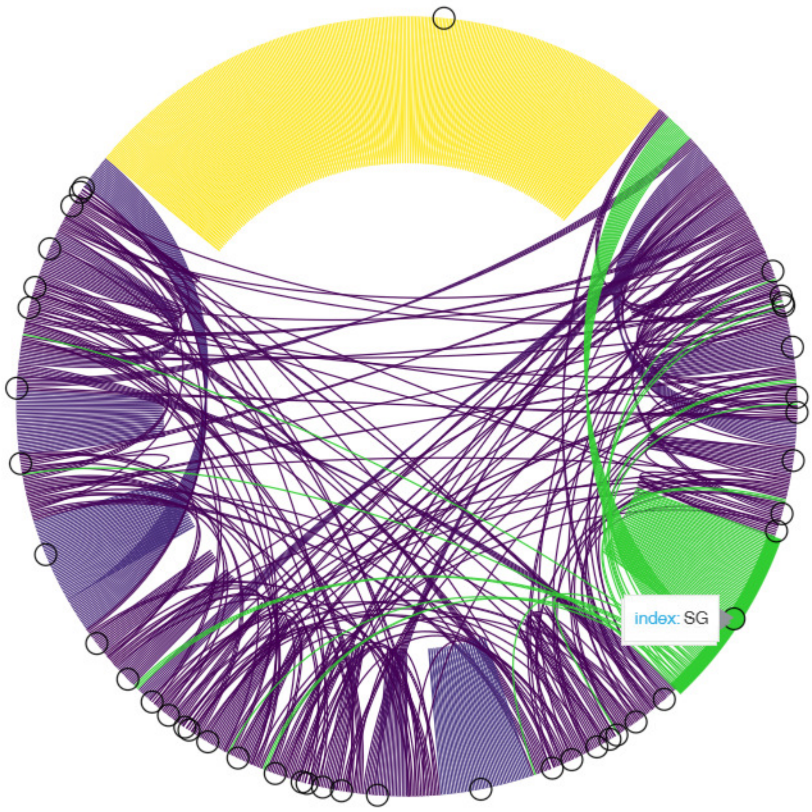
Figure 5: Co-authorship distribution of insitutions (*top*) and countries (*bottom*) across partitions

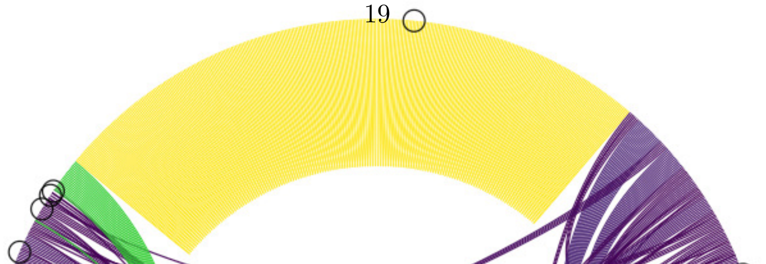

index: Wuhan University of Technology

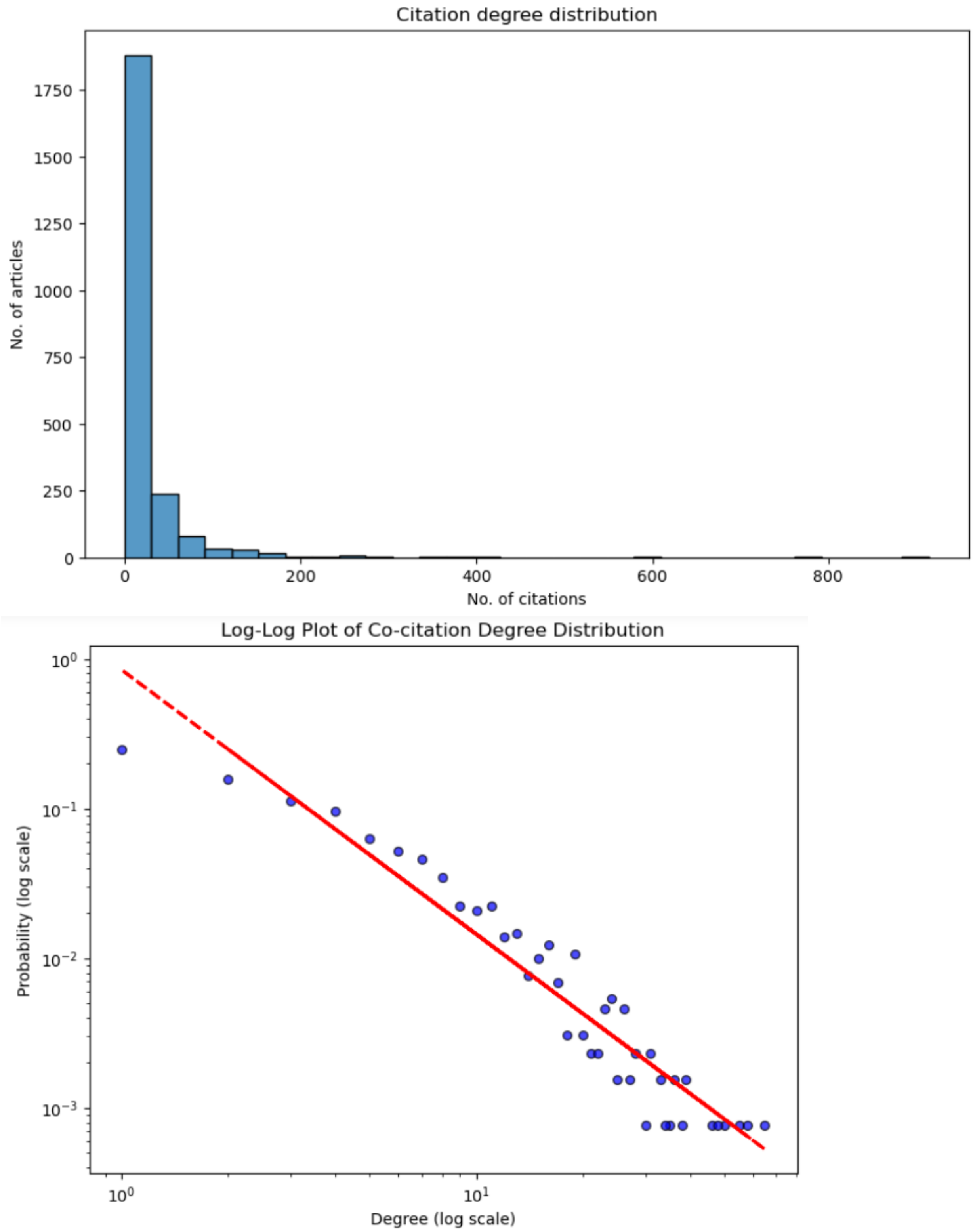

index: Shanghai Maritime University


index: Dalian Maritime University


index: CN


index: SG






```
Slope: -1.7682132421361874, Intercept: -0.16382794519668037
```

Figure 8: Co-citation degree distribution and log-log chart

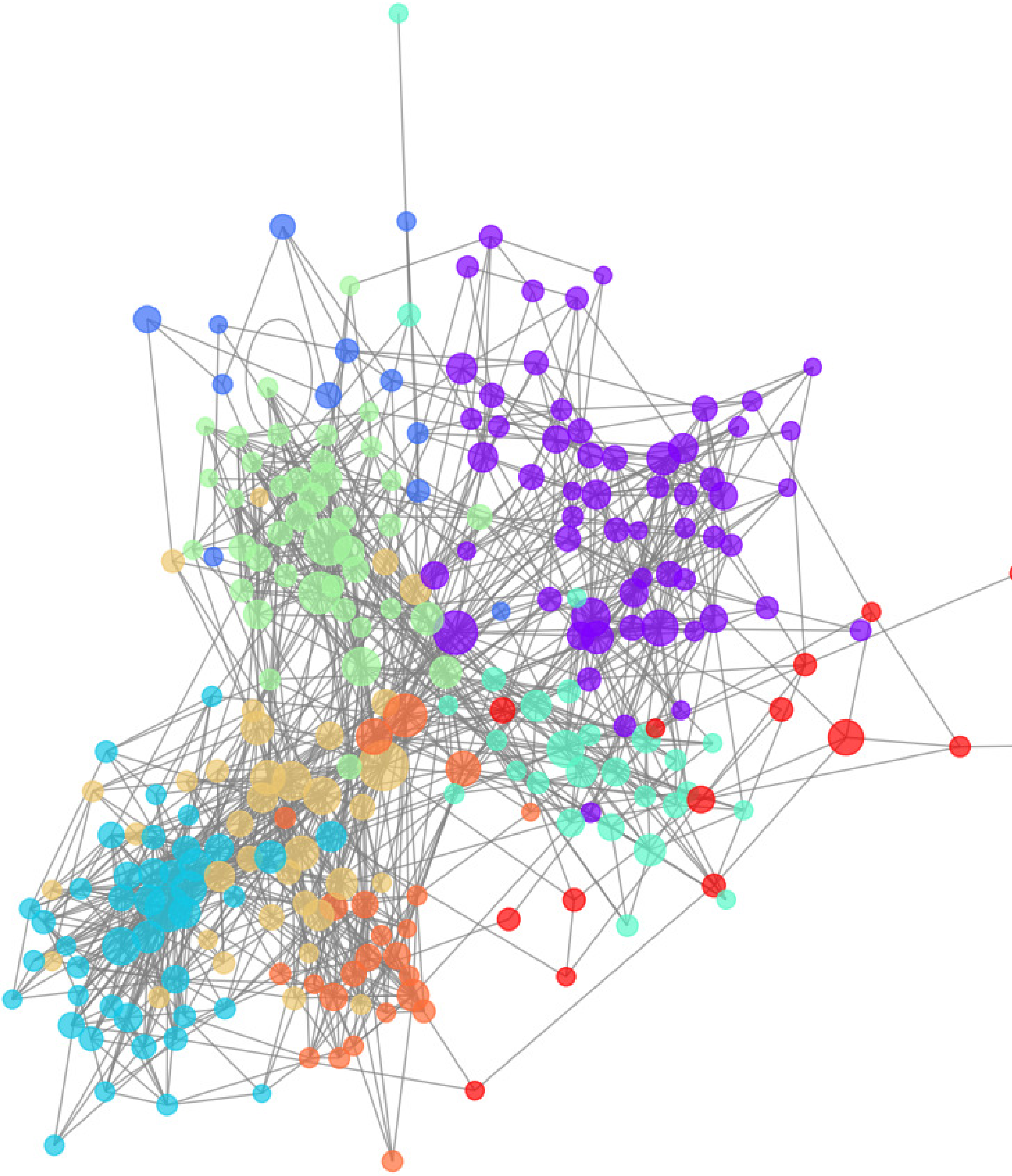


Figure 9: Co-citation network with communities

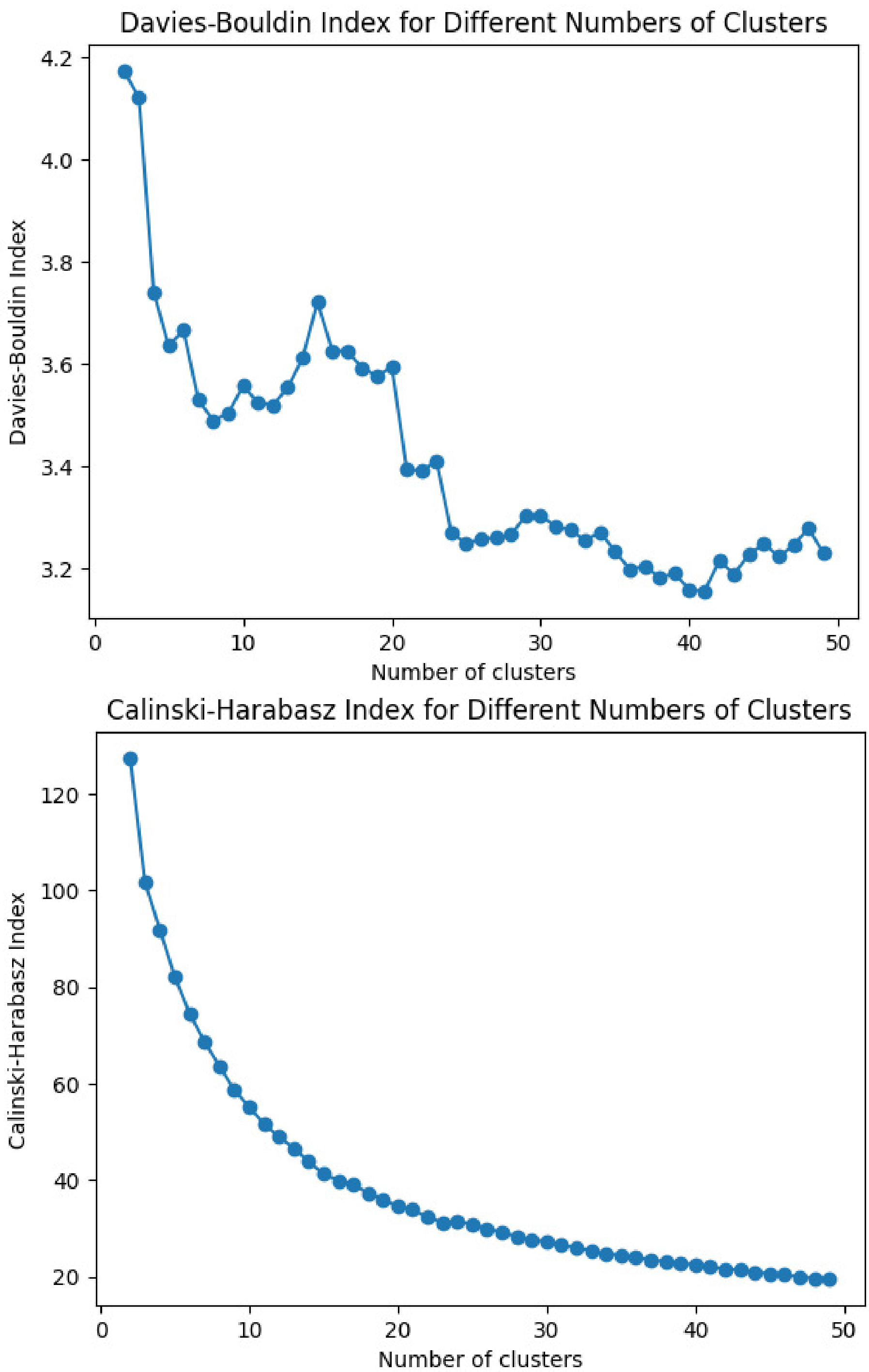


Figure 10: Davies-Bouldin (top) and Calinski-Harabasz (*bottom*) indexes.

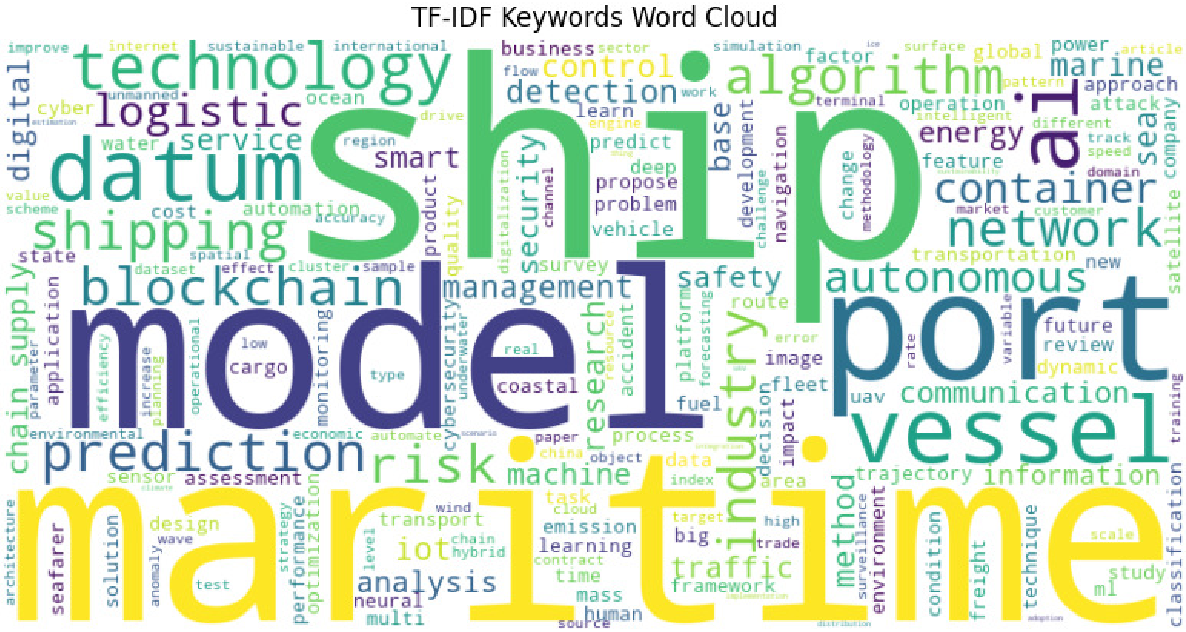


Figure 11: Word cloud based on TF-IDF

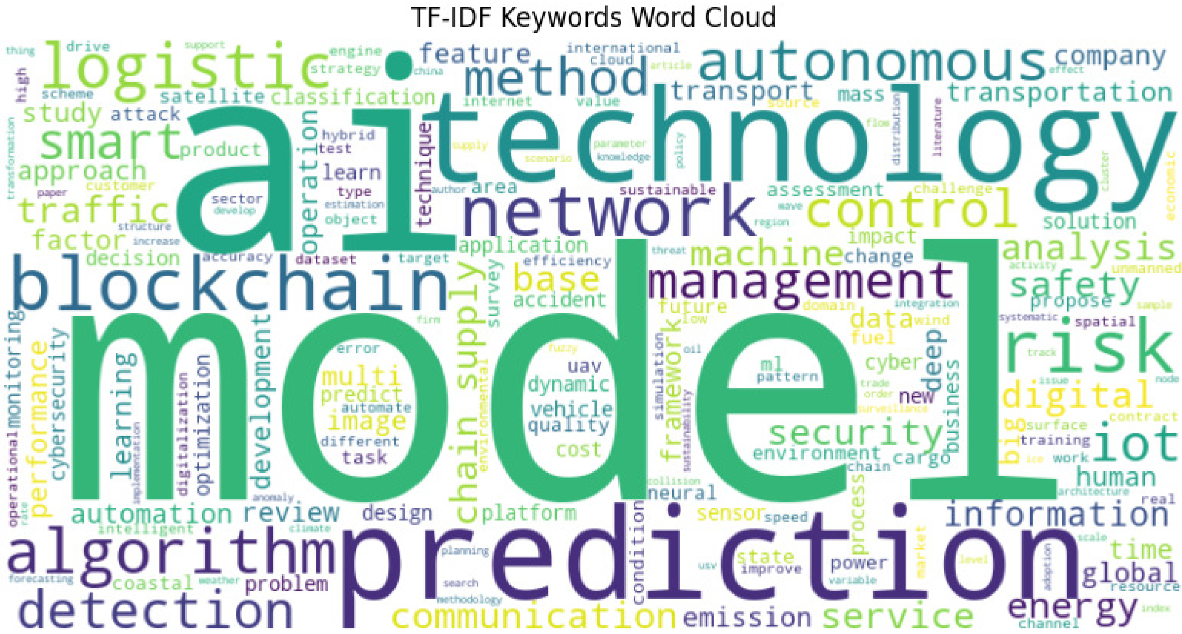


Figure 12: Word cloud focused on technological terms

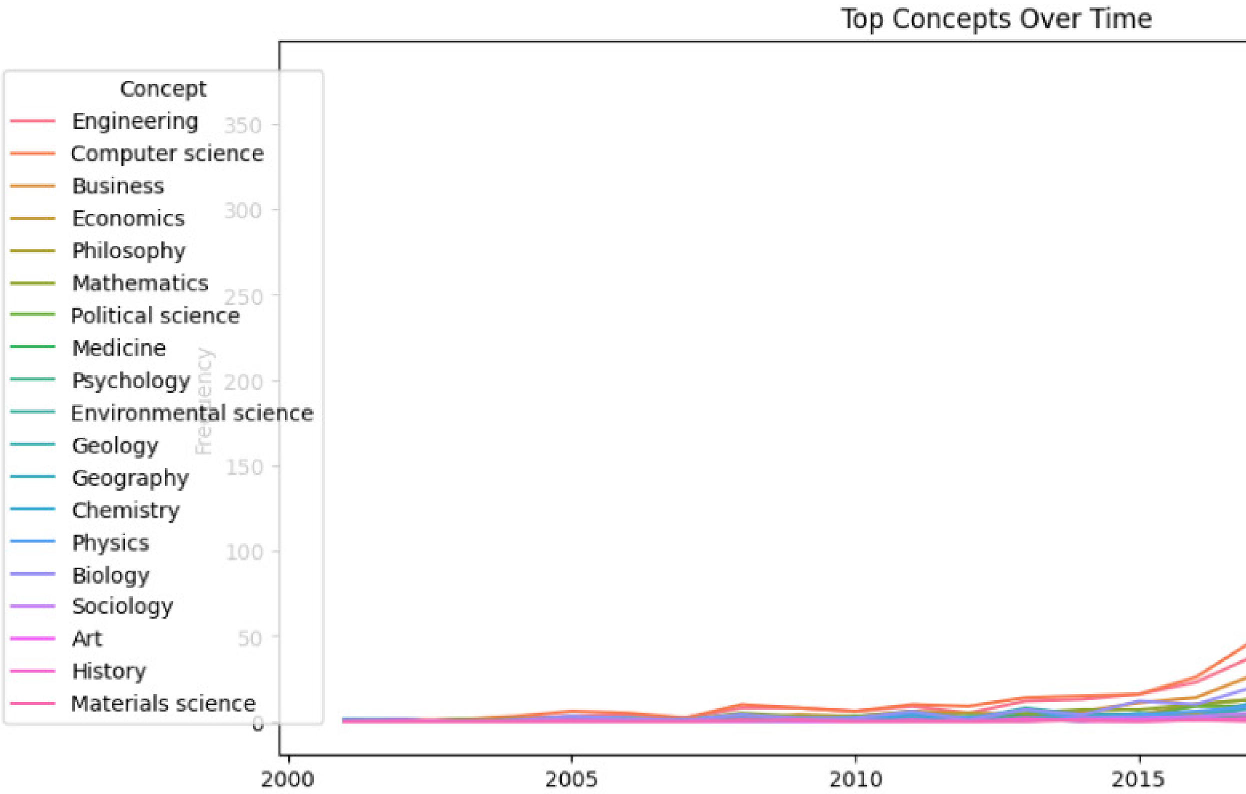


Figure 13: Top-level OpenAlex concept distribution over time terms

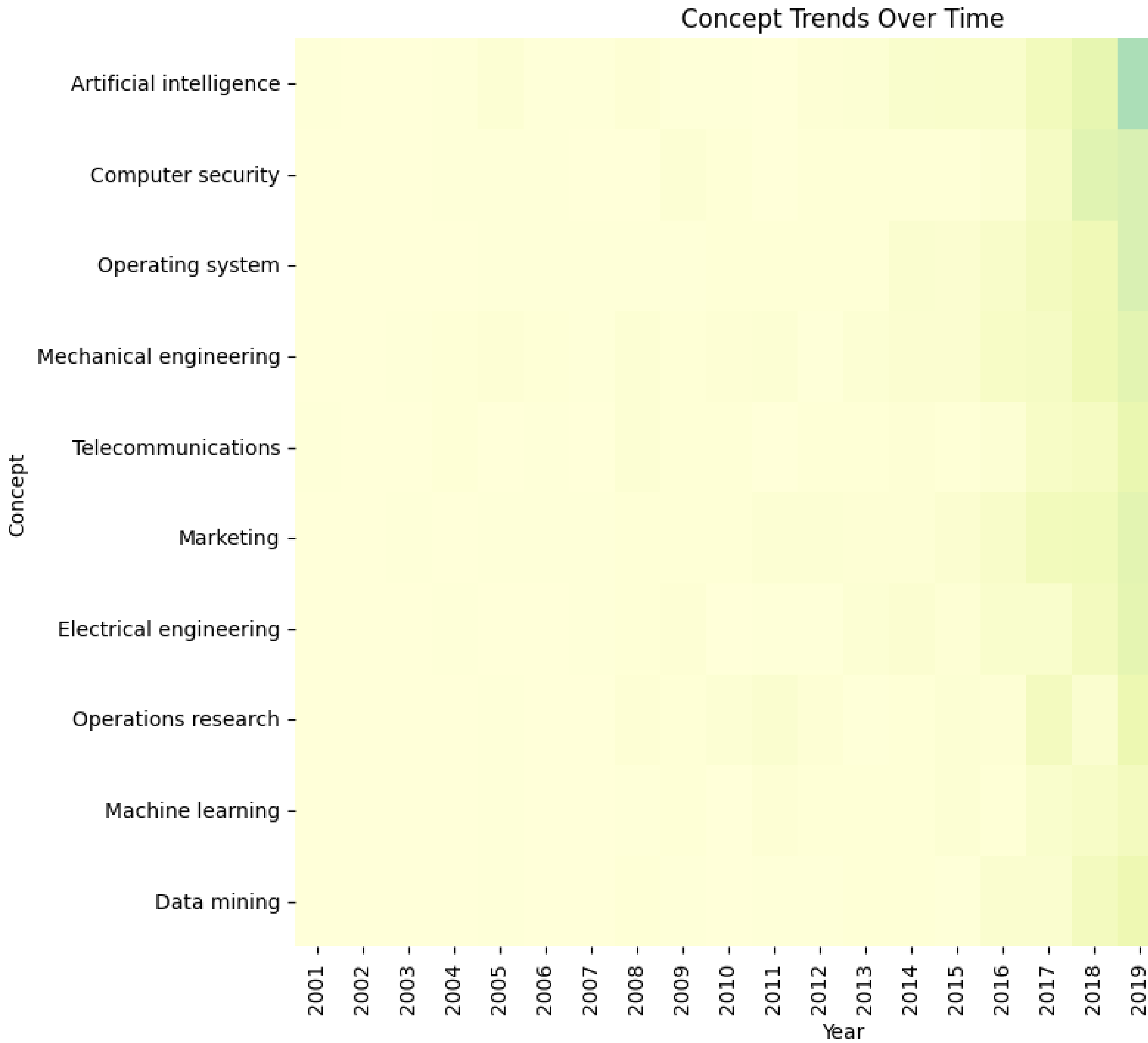


Figure 14: Detail-level OpenAlex concept distribution over time for the 10 most relevant terms

List of Tables

Table 1: List of keywords identified by experts.

Keyword	Type (Digit. Trans. or Shipping)
Digital transformation	Digit. Trans.
Digital innovation	Digit. Trans.
Digital ecosystems	Digit. Trans.
Digitization	Digit. Trans.
Digitalization	Digit. Trans.
Digital platforms	Digit. Trans.
Industry 4.0	Digit. Trans.
Smart technologies	Digit. Trans.
Data-driven transformation	Digit. Trans.
Automation	Digit. Trans.
Internet of Things	Digit. Trans.
Blockchain	Digit. Trans.
Data analysis	Digit. Trans.
Artificial intelligence	Digit. Trans.
Machine learning	Digit. Trans.
Big data	Digit. Trans.
Cloud computing	Digit. Trans.
Cyber-physical systems	Digit. Trans.
Digital twins	Digit. Trans.
Edge computing	Digit. Trans.
5G networks	Digit. Trans.
Predictive analytics	Digit. Trans.
Cybersecurity	Digit. Trans.
Supply chain integration	Digit. Trans.
shipping	Shipping
maritime	Shipping
Sea freight	Shipping
Smart ports	Shipping
Autonomous ships	Shipping
Fleet management	Shipping
Cargo tracking	Shipping
Digital shipyards	Shipping
Port digitalization	Shipping
Port automation	Shipping
Vessel performance	Shipping

Table 2: Number of retrieved articles per research engine.

Engine	No. of scientific articles
EBSCO	1904
ProQuest	2011
IEEE eXplore	300

Table 3: Distribution of publications across institutions.

Institution	No. of publications	Cumulative %	% Of total
Dalian Maritime University	70	3.48	3.48
Wuhan University of Technology	44	5.67	2.19
Norwegian University of Science and Technology	44	7.86	2.19
Shanghai Maritime University	41	9.90	2.04
Nanyang Technological University	30	11.39	1.50
University of Rijeka	23	12.53	1.14
Hong Kong Polytechnic University	17	13.38	0.85
University of Strathclyde	16	14.17	0.80
University of Piraeus	15	14.92	0.75
Liverpool John Moores University	14	15.61	0.70
Istanbul Technical University	13	16.26	0.65
Tsinghua University	13	16.91	0.65
Delft University of Technology	12	17.50	0.60
Universidad Politécnica de Madrid	12	18.10	0.60
Korea Maritime and Ocean University	11	18.65	0.55
Zhejiang Ocean University	11	19.19	0.55
University of Genoa	11	19.74	0.55
National Technical University of Athens	11	20.29	0.55
University of South-Eastern Norway	10	20.79	0.50
Aalto University	10	21.28	0.50

Table 4: Distribution of publications across countries.

Country ISO	No. of publications	Cumulative %	% Of total
CN	487	24.29	24.29
US	138	31.17	6.88
GB	107	36.51	5.34
KR	96	41.30	4.79
NO	95	46.03	4.73
GR	66	49.33	3.29
IN	59	52.27	2.94
ES	58	55.16	2.89
IT	54	57.86	2.69
SG	54	60.55	2.69
DE	50	63.04	2.49
CA	47	65.39	2.34
ID	42	67.48	2.09
AU	40	69.48	2.00
HR	39	71.42	1.95
PL	38	73.32	1.90
TR	36	75.11	1.80
SE	30	76.61	1.50
FI	29	78.05	1.45
JP	26	79.35	1.30

Table 5: Top 10 influential papers based on citations.

DOI	Title
https://doi.org/10.1080/03088839.2020.1788731	Big data and artificial intelligence in the maritime indus
https://doi.org/10.3390/jmse12060919	Comprehensive Analysis of Maritime Cybersecurity Lan
https://doi.org/10.1016/j.tre.2019.09.020	Maritime shipping digitalization: Blockchain-based techn
https://doi.org/10.1080/01441647.2019.1649315	How big data enriches maritime research–a critical revie
https://doi.org/10.3390/app14145994	Harnessing AI for sustainable shipping and green ports:
https://doi.org/10.1016/j.ijcip.2022.100571	Developments and research directions in maritime cyber
https://doi.org/10.1109/tits.2019.2908191	Traffic pattern mining and forecasting technologies in m
https://doi.org/10.3390/s19040926	Toward Digitalization of Maritime Transport?
https://doi.org/10.3390/info13010022	Cyber security in the maritime industry: A systematic s
https://doi.org/10.3390/jmse10040486	Digitalization in maritime transport and seaports: biblio

Table 6: Community-based topic analysis

Community	Main theme
1	Optimization and prediction of fuel consumption, energy efficiency, and environmental impa
2	Maritime safety, risk management, and the application of machine learning techniques to pr
3	Machine learning, artificial intelligence, and big data applications in the maritime domain.
4	Integration of Internet of Things (IoT), mobile edge computing, communication networks, a
5	Digital transformation and technological advancements within the maritime sector, particula
6	Development and optimization of smart ports, focusing on the integration of Industry 4.0 te
7	Maritime cybersecurity, with an emphasis on cyber risks associated with the digital transfor
8	Adoption and application of blockchain technology in the maritime industry, specifically in a

Table 7: Centrality-based topic analysis.

OpenAlex ID	Top 5 in N centrality measures	Main topic
W3041382323	5	Big data and artificial intelligence
W4400493457	4	Artificial intelligence, sustainable shipping, and green po
W2964482263	4	Big data
W2978644098	3	Blockchains
W4386245296	2	Data and IoT
W4225993858	2	Digitalization
W4399283331	2	Cybersecurity
W3090216936	1	Blockchain conceptual framework
W4205557186	1	Cybersecurity
W3213918042	1	Blockchain

Table 8: Results of thematic analysis.

Topic 1: Advancements in AI and Machine Learning for Maritime Applications. This topic ex
Topic 2: Spatial and Environmental Analysis in Maritime and Inland Waters. This topic focus
Topic 3: Predictive Modeling and Risk Assessment in Maritime Operations. This topic delves
Topic 4: Data-Driven Optimization in Maritime Engineering and Traffic Management. This t
Topic 5: Blockchain and Edge Computing in Maritime Operations. This topic investigates the a
Topic 6: Innovations and Reviews in Autonomous and Smart Shipping. This topic reviews the s
Topic 7: Digital Transformation and Sustainability in Maritime Logistics. This topic focuses on
Topic 8: Maritime Industry's Adaptation to Digital and Sustainable Practices. This topic expl