Full Name: Jawwad Shadman Siddique

Email: siddiquj@oregonstate.edu

Major: Computer Science

Course Name: Introduction to Applied Data Science

Course Number: CS 332

Project Title: **Mapping the Landscape of Academic Research: Identifying Communities of Scholarly Collaboration and Similarity in Computer Science Publications**

TABLE OF CONTENTS

1. INTRODUCTION

Topic Overview

Academic research in computer science has become increasingly specialized and interdisciplinary, creating a complex web of scholarly communities that span institutions, disciplines, and geographic boundaries. Understanding how researchers cluster into communities based on their research interests, publication patterns, and citation metrics is crucial for multiple stakeholders in the academic ecosystem. This project aims to identify and analyze communities of similar research by examining patterns in professors' and researchers' publication records from two premier computer science organizations: the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE).

The identification of research communities has significant implications for university administrators seeking to build interdisciplinary teams, funding agencies looking to identify emerging research areas, early-career researchers trying to find mentors and collaborators, and academic institutions aiming to assess their research strengths and gaps. By applying unsupervised learning techniques (clustering) to academic publication data from ACM and IEEE; two of the most prestigious venues for computer science research, we can uncover hidden patterns of intellectual similarity that transcend traditional departmental boundaries.

This dataset provides a unique opportunity to examine publication metrics including citation counts, h-index values, authorship patterns, and publication venues. Furthermore, understanding which factors predict a researcher's community membership (through supervised learning) can help identify the key drivers of scholarly collaboration and intellectual affinity. This analysis affects hundreds of thousands of computer science researchers worldwide, influences billions of dollars in research funding allocation, and shapes the direction of technological innovation and scientific discovery.

Why This Topic is Relevant to Me

I chose this project because understanding research communities is essential for navigating the modern academic landscape in computer science and making informed decisions about collaboration, specialization, and career development. As someone interested in research and network science, I recognize that identifying where my interests align with existing research communities can help me find potential mentors, collaborators, and opportunities for impactful work.

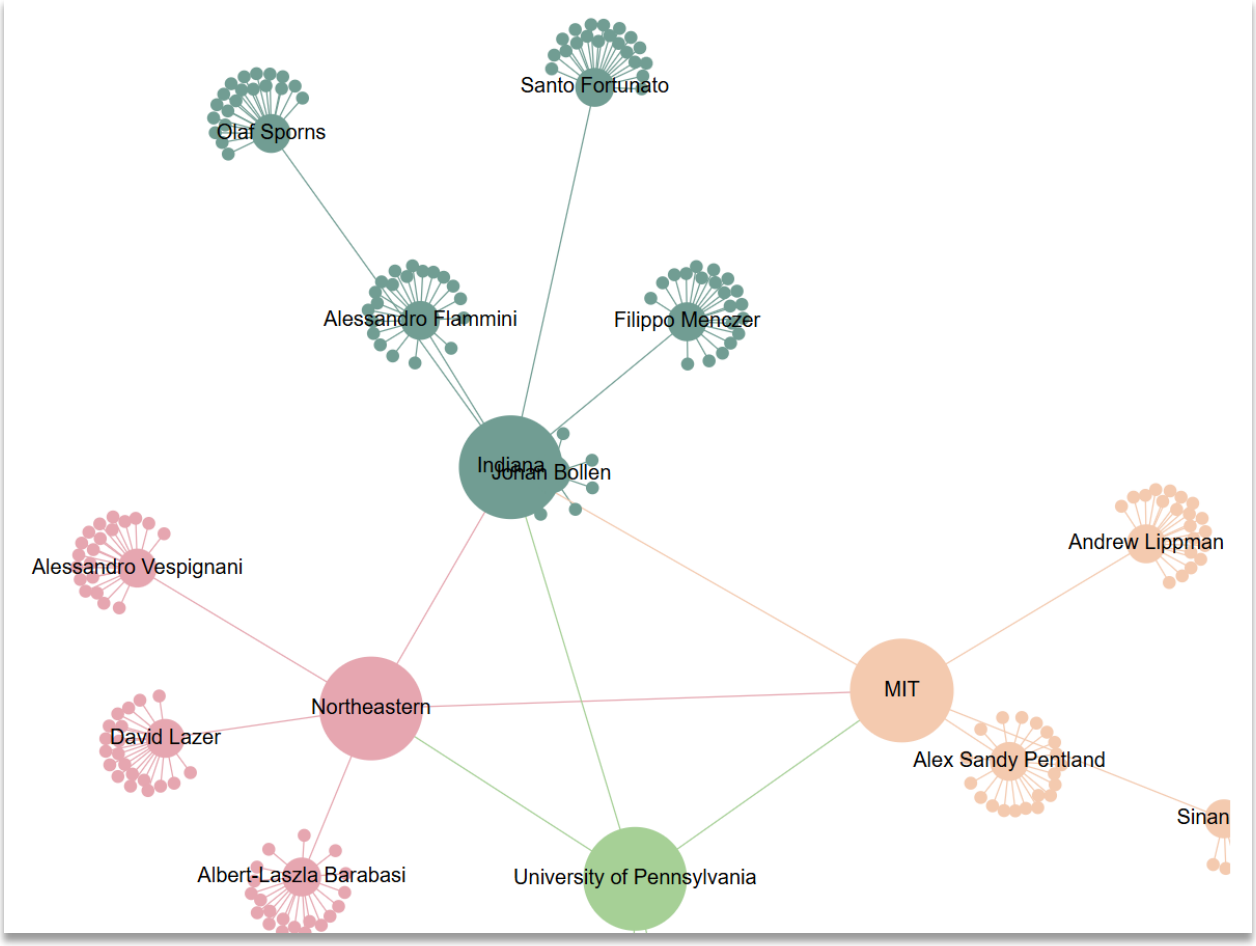
The ACM and IEEE represent the pinnacle of computer science publishing, and analyzing their publication data provides insights into the structure of the entire field. Additionally, the problem of community detection is a fascinating application of data science that combines network analysis, bibliometrics, and machine learning; skills that are valuable across many industries beyond academia.

This topic also resonates with me because it addresses a real challenge in modern computer science research: the difficulty of discovering relevant work and potential collaborators in an era of

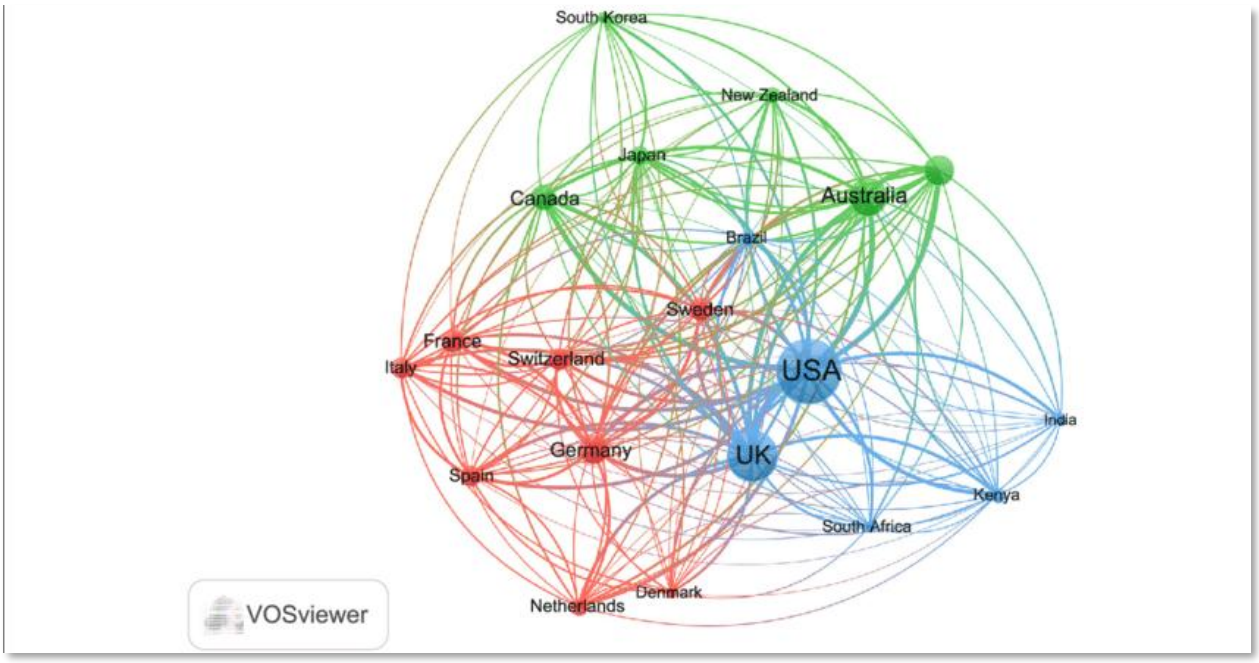
information overload. With tens of thousands of papers published in ACM and IEEE venues annually, systematic approaches to identifying communities of similar research can help researchers cut through the noise and find their intellectual homes. By completing this project, I hope to develop practical skills in unsupervised learning while contributing insights that could help researchers, including myself, make better-informed decisions about their academic trajectories and understand what distinguishes highly-cited, influential research from less impactful work.

Supporting Image

This interactive visualization is developed by me using Javascript.



The following image is taken from the given link: https://www.researchgate.net/figure/Network-visualization-map-of-international-research-collaboration-among-top-20-active_fig3_341254154



Research Questions

1. What distinct communities of researchers emerge when clustering based on publication metrics and citation patterns?

Approach: Use unsupervised clustering algorithms on features including citation counts, h-index, publication counts, and author collaboration patterns.

2. How do citation metrics (h-index, total citations, papers published) vary across different research communities?

Approach: Perform exploratory data analysis with visualizations (box plots, violin plots, histograms) comparing citation distributions across identified clusters. Calculate summary statistics (mean, median, quartiles) for each community to identify high-impact vs. emerging research areas.

3. Can we predict which research community a researcher belongs to based on their publication and citation characteristics?

Approach: Build a decision classifier using features like number of publications, h-index, citation counts, author position patterns, and publication venue distribution. Identify most important features for classification.

4. What distinguishes highly-cited researchers from less-cited researchers within the ACM/IEEE ecosystem?

Approach: Compare publication patterns, authorship characteristics, and collaboration networks between high-citation and low-citation groups. Use statistical tests to identify significant differences in key metrics.

5. Are there patterns in authorship (solo vs. collaborative work) across different research communities?

Approach: Analyze author count distributions within each cluster. Calculate collaboration indices and compare across communities. Visualize with bar charts and network graphs showing co-authorship density.

6. How do h-index values correlate with total publication counts and citation counts across communities?

Approach: Create correlation matrices and scatter plots with regression lines. Perform correlation analysis to quantify relationships. Compare correlation strengths across different clusters.

7. Can we identify "bridge researchers" who connect multiple research communities through diverse publication patterns?

Approach: Identify researchers whose publication metrics place them at cluster boundaries or who show characteristics of multiple communities. Analyze their publication diversity and impact across different research areas.

8. Which features (publication count, h-index, citation count, author position, etc.) are most important for determining research community membership?

Approach: Create visualizations (bar charts, heatmaps) showing relative importance of different features in clustering and classification.

9. Do ACM and IEEE publications show different patterns in citation impact and author collaboration?

Approach: If publication venue data distinguishes between ACM and IEEE, perform comparative analysis of citation metrics, h-indices, and collaboration patterns between the two organizations. Use statistical tests to identify significant differences.

10. Can we identify emerging vs. established research communities based on publication and citation distributions?

Approach: Analyze cluster characteristics looking for patterns that suggest maturity (high h-index, stable citation counts) vs. emergence (rapidly growing citations, newer publication patterns). Create profiles for each community type.

11. What is the relationship between collaboration intensity and research impact across different communities?

Approach: Calculate collaboration intensity metrics (average co-authors per paper, percentage of multi-author papers) for each researcher and cluster. Create scatter plots showing the relationship between collaboration intensity and impact metrics (h-index, citations per paper).

12. Can we predict a researcher's h-index category (low/medium/high) using publication patterns and collaboration characteristics?

Approach: Create categorical labels by dividing researchers into h-index tertiles or quartiles (e.g., emerging: h-index 0-5, established: 6-15, elite: 16+). Build a decision tree classifier using features such as total publications, citation count, average co-authors, first-author publications, and years active.

13. Are there distinct authorship position patterns (first, middle, last author) that characterize different research communities or career stages?

Approach: Analyze the distribution of authorship positions within each identified cluster. Calculate metrics such as percentage of first-author papers, last-author papers (often indicating senior/PI status), and middle-author papers. Create stacked bar charts or heatmaps showing authorship position profiles for each community.

14. What are the characteristics of "high-efficiency" researchers who achieve exceptional impact with relatively fewer publications?

Approach: Calculate a research efficiency metric (e.g., citations per paper, h-index divided by publication count) to identify researchers who achieve high impact with fewer publications. Use scatter plots to visualize the efficiency frontier.

2. DATA GATHERING

Dataset Description

Alternate Dataset: Research Citation Network - 5M Papers

- **Dataset Name:** Research Citation Network (5 Million Papers)
- **Source:** Kaggle (provided by Agung Pambudi)
- **URL:** https://www.kaggle.com/datasets/agungpambudi/research-citation-network-5m-papers

- **Primary Files:** Multiple CSV files including paper metadata, citations, and author information
- **Dataset Size:** 5 million research papers with comprehensive citation network data
- **Focus:** Multi-disciplinary academic publications with detailed citation relationships
- **Time Coverage:** Extensive temporal range covering multiple decades of scholarly research

Dataset Characteristics:

This dataset provides an extraordinary collection of research paper metadata combined with citation network information, creating a rich resource for understanding how knowledge flows through academic communities. It includes paper-level data with titles, authors, publication years, venues, and citation relationships that map which papers cite which other papers. The dataset is particularly valuable because it explicitly captures the citation network structure, making it ideal for identifying communities of researchers who build upon each other's work and form intellectual clusters within the broader academic ecosystem.

The data represent published research articles from diverse academic disciplines, complete with citation graphs that reveal the interconnectedness of papers, authors, and research topics. Each paper includes bibliographic metadata along with inbound citations (papers that cite it) and outbound citations (papers it references), creating a directed network that can be analyzed using both traditional bibliometric methods and advanced network analysis techniques. This dataset is exceptionally well-suited for clustering analysis because it contains multiple dimensions of similarity: content-based similarity (through titles and topics), collaboration-based similarity (through co-authorship), and influence-based similarity (through citation patterns).

The citation network structure enables sophisticated community detection algorithms that can identify research clusters based on how papers reference each other, revealing intellectual lineages and schools of thought. With 5 million papers, the dataset provides sufficient scale to detect both large established research communities and small emerging subfields. This dataset's strength lies in its ability to support multiple types of community analysis: co-citation analysis (papers cited together likely belong to the same community), bibliographic coupling (papers sharing references likely address similar topics), and direct citation analysis (highly-cited papers within clusters indicate community thought leaders).

Primary Dataset: Academic Publications Metrics from ACM/IEEE

- **Dataset Name:** Academic Publications Metrics from ACM, IEEE, and Other Sources
- **Source:** Kaggle
- **URL:** https://www.kaggle.com/datasets/thedevastator/academic-publications-metrics-from-acm-ieee-and
- **Primary File:** `final_data_2.csv`
- **Dataset Size:** Multiple CSV files with comprehensive publication and author metrics
- **Focus:** Computer science publications from premier venues (ACM, IEEE)

Dataset Characteristics:

This dataset provides a rich collection of academic publication metrics from two of the most prestigious computer science organizations. It includes individual researcher-level data with publication counts, citation metrics, h-index values, and authorship patterns. The dataset is particularly valuable because it focuses on high-quality computer science research, making it ideal for identifying communities within specific CS subfields.

The data represents published research articles, conference papers, and journal publications, with associated metrics that allow for both bibliometric analysis and machine learning applications. This dataset is particularly well-suited for clustering analysis because it contains multiple quantitative features that can reveal natural groupings of researchers based on their scholarly impact and publication patterns.

We will work with the primary dataset.

Primary Dataset Screenshot

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
		Unnamed:	Document	Authors	Author Affi	Universitie	Publicatior	Abstract	ISSN	ISBNs	DOI	Article Cita	Patent Cita	Publisher	
0	0	0													
1	1	1	A Blockcha	C. Zhang; Y	China	['Hunan Ur	2022	Model migı	2327-4662		10.1109/JIOT.2022.3171926			IEEE	
2	2	2	A Blockcha	C. Zhang; Y	Sweden	['Hunan Ur	2022	Model migı	2327-4662		10.1109/JIOT.2022.3171926			IEEE	
3	3	3	BESIFL: Blc	Y. Xu; Z. Lu	China	['Fudan Un	2021	Federated	2327-4662		10.1109/JI	2		IEEE	
4	4	4	BESIFL: Blc	Y. Xu; Z. Lu	USA	['Fudan Un	2021	Federated	2327-4662		10.1109/JI	2		IEEE	
5	5	5	BESIFL: Blc	Y. Xu; Z. Lu	at	['Fudan Un	2021	Federated	2327-4662		10.1109/JI	2		IEEE	
6	6	6	BPPS:Bloc	K. Park; J. I	India	['Kyungpoo	2022	With the or	1941-0018		10.1109/TI	3		IEEE	
7	7	7	TRUCON: I	M. Yuan; Y	China	['Central S	2022	The Interne	1558-0016		10.1109/TITS.2022.3226500			IEEE	
8	8	8	Conditiona	J. Liu; W. Ji	Japan	['Xidian Un	2022	As an impo	2168-2208		10.1109/JBHI.2022.3183397			IEEE	
9	9	9	Conditiona	J. Liu; W. Ji	China	['Xidian Un	2022	As an impo	2168-2208		10.1109/JBHI.2022.3183397			IEEE	
10	10	10	Secure and	H. Zhang; Y	China	['Beijing Ur	2021	With the ra	2327-4662		10.1109/JIOT.2021.3121482			IEEE	
11	11	11	MSTDB: A I	E. Zhou; Z.	Hong Kong	['Xidian Un	2022	Blockchair	1558-2191		10.1109/TKDE.2022.3220522			IEEE	
12	12	12	MSTDB: A I	E. Zhou; Z.	Norway	['Xidian Un	2022	Blockchair	1558-2191		10.1109/TKDE.2022.3220522			IEEE	
13	13	13	MSTDB: A I	E. Zhou; Z.	China	['Xidian Un	2022	Blockchair	1558-2191		10.1109/TKDE.2022.3220522			IEEE	
14	14	14	MSTDB: A I	E. Zhou; Z.	at	['Xidian Un	2022	Blockchair	1558-2191		10.1109/TKDE.2022.3220522			IEEE	
15	15	15	Blockchair	Z. Zhou; Y.	China	['Guilin Uni	2022	Federated	1941-0050		10.1109/TII.2022.3215192			IEEE	
16	16	16	A Tractable	A. Hafid; A	Morocco	['University	2022	Blockchair	2168-6750		10.1109/TI	1		IEEE	
17	17	17	A Tractable	A. Hafid; A	Canada	['University	2022	Blockchair	2168-6750		10.1109/TI	1		IEEE	
18	18	18	Immutable	W. Pourma	Canada	['Ryerson U	2021	Service Lev	1939-1374		10.1109/TS	1		IEEE	
19	19	19	Immutable	W. Pourma	United Stat	['Ryerson U	2021	Service Lev	1939-1374		10.1109/TS	1		IEEE	
20	20	20	GpDB: A G	Z. Liao; S. (	China	['410114 C	2022	The industı	1941-0050		10.1109/TII.2022.3162201			IEEE	
21	21	21	GpDB: A G	Z. Liao; S. (	United Stat	['410114 C	2022	The industı	1941-0050		10.1109/TII.2022.3162201			IEEE	

Variables in the Dataset

Based on the ACM/IEEE Academic Publications dataset (final_data_2.csv), the key variables include:

Quantitative Variables:

1. **Citation Count:** Total number of citations received by a researcher's publications
2. **h-index:** Measure of researcher productivity and citation impact (h papers with at least h citations each)

3. **Publication Count:** Total number of papers published by the researcher
4. **i10-index:** Number of publications with at least 10 citations (if available)
5. **Author Position Metrics:** Frequency of being first author, last author, or middle author
6. **Collaboration Metrics:** Average number of co-authors per paper
7. **Citations per Paper:** Average citation impact per publication
8. **Publication Venue Counts:** Distribution of publications across different journals/conferences

Qualitative Variables:

1. **Author Name/ID:** Identifier for individual researchers
2. **Affiliation Information:** Institution or organization (if available)
3. **Research Area/Field:** Computer science subfield or specialization (if available)
4. **Publication Venue Names:** Specific journals or conference names
5. **Paper Titles:** Titles of published works (if available)
6. **Keywords/Topics:** Research topic indicators (if available)

Is the Dataset Labeled?

Partially Labeled:

- **Natural groupings exist:** The dataset may contain inherent categories based on research areas or publication venues
- **Labels will be created:** Through the clustering process, we will create labels (cluster assignments) that identify research communities
- **For supervised learning:** After clustering, the cluster assignments become labels for the classification task
- **Binary labels possible:** We can create labels like "High-Impact Researcher" vs. "Emerging Researcher" based on citation thresholds

Does it Contain Both Qualitative and Quantitative Data?

Yes.

Quantitative Data (Primary Focus):

- Citation counts, h-index, publication counts
- Author collaboration metrics
- Citation-per-paper ratios
- Temporal metrics (if publication years available)

Qualitative Data:

- Author names and identifiers
- Affiliation information
- Publication venues

- Research area classifications
- Paper titles and keywords (if available)

This combination makes the dataset ideal for comprehensive analysis using both statistical methods and machine learning algorithms.

Data Augmentation and Enhancement Strategies

To maximize the value of this dataset for community detection, consider these enhancements:

1. Feature Engineering:

- **Research Impact Score:** Combine h-index, citations, and publication count into composite metrics
- **Collaboration Index:** Calculate average co-authors per paper
- **Productivity Metrics:** Publications per year (if temporal data available)
- **Citation Velocity:** Rate of citation accumulation
- **Seniority Indicators:** Proxy measures based on publication history

2. Derived Categorical Variables:

- **Impact Quartiles:** Classify researchers into quartiles based on h-index or citations
- **Researcher Type:** Solo vs. collaborative researcher based on authorship patterns
- **Publication Focus:** Primary venue or research area

Data Quality Considerations

Strengths:

- Focuses on high-quality CS publications (ACM/IEEE)
- Contains standardized bibliometric measures (h-index)
- Includes multiple dimensions for clustering (citations, publications, collaboration)
- Suitable for both unsupervised and supervised learning

Potential Challenges:

- May contain missing values that need handling
- Citation counts may be time-dependent (older papers have more citations)
- May need normalization due to different scales (h-index vs. citation count)
- Name disambiguation may be necessary (same author, different spellings)

Data Cleaning Strategy:

1. Handle missing values (imputation or removal)
2. Remove duplicate entries
3. Normalize/standardize features for clustering

4. Handle outliers (extremely high citation counts)
5. Create consistent author identifiers

Why This Dataset is Ideal for the Project

1. **Perfect for Clustering:** Multiple numerical features enable meaningful community detection
2. **Real-World Relevance:** ACM/IEEE data represents actual computer science research landscape
3. **Sufficient Complexity:** Enough variables for sophisticated analysis without being overwhelming
4. **Actionable Insights:** Results can inform real career and collaboration decisions
5. **Complete Workflow:** Supports both unsupervised (clustering) and supervised (decision trees) learning objectives

REFERENCES

Dataset Sources

Pambudi, A. (2023). *Research Citation Network (5 Million Papers)*. Kaggle. https://www.kaggle.com/datasets/agungpambudi/research-citation-network-5m-papers

TheDevastator. (2023). *Academic Publications Metrics from ACM, IEEE, and Other Sources*. Kaggle. https://www.kaggle.com/datasets/thedevastator/academic-publications-metrics-from-acm-ieee-and

Academic Literature & Methodology

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3-5), 75-174. https://doi.org/10.1016/j.physrep.2009.11.002

Bibliometrics & Citation Analysis

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569-16572. https://doi.org/10.1073/pnas.0507655102

Research Communities & Scholarly Communication

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039. https://doi.org/10.1126/science.1136099

Computer Science Research Landscape

ACM Digital Library. (2024). *Association for Computing Machinery*. https://dl.acm.org/

IEEE Xplore Digital Library. (2024). *Institute of Electrical and Electronics Engineers*. https://ieeexplore.ieee.org/

Image Sources

Siddique, J. S. (2025). *Interactive network visualization of research collaboration* [JavaScript visualization]. Custom development.

Uddin, S., Hossain, L., & Rasmussen, K. (2013). Network visualization map of international research collaboration among top 20 active countries. *ResearchGate*. https://www.researchgate.net/figure/Network-visualization-map-of-international-research-collaboration-among-top-20-active_fig3_341254154