NETWORK SCIENCE

(CSE655)

Project Report

GROUP 12

BTECH CSE'25

Introduction

Motivation: Scientific collaboration shapes knowledge spread. We analyzed arXiv to understand how co-authorship patterns evolve over time, reveal interdisciplinary links, and predict future links. This addresses research questions about temporal trends, interdisciplinarity, and network structure in academia.

Dataset: We used the arXiv metadata snapshot (via Kaggle) containing authors and categories for ~1 million papers up to 2024. We exclude "mega-authored" papers (>30 authors) to focus on typical collaborations.

Research Questions: How has the co-authorship network's size and structure changed annually? How interdisciplinary are author collaborations? Can simple network heuristics predict new co-authorships? How do patterns differ in a major field (e.g. cs.LG)? Which fields and authors are most central?

Methodology

- **Data parsing:** We loaded the JSON metadata line-by-line (limited to 1M entries). For each paper, we extracted the publication year and author list. We assigned authors to the update year (2000–2024). Authors' names were cleaned (affiliations removed) to ensure consistency.

- **Network construction:** For each year, we built an undirected graph where each author is a node, and an edge joins two co-authors on a paper. This yielded one graph per year (2000–2024).

- **Metrics:** For each annual graph, we computed the number of nodes, number of edges, average degree, clustering coefficient, and density. Degree centrality was computed (for field-level analysis). Entropy of an author's field distribution was computed as a proxy for interdisciplinarity.

- **Tools:** Analysis used Python libraries: `NetworkX` for graph operations, `Pandas` for data frames, `Matplotlib/Seaborn` for plots. Community detection used the `Louvain` method to compute modularity and

communities per year.

Deliverable 1: Temporal Evolution of the Co-Authorship Network

- **Network by year:** We incrementally built yearly co-authorship graphs from 2000 through 2024. The number of unique authors grew roughly exponentially in the 2000s and 2010s, surging from ~18,900 in 2007 to ~328,500 in 2015. Edge count similarly peaked in 2015 (≈1.89M edges). These spikes correspond to 2015's anomaly of many multi-authored papers, which we excluded when >30 authors.

- **Yearly metrics:** The table of annual metrics (nodes, edges, average degree, clustering, density) shows rapid early growth and later decline in nodes/edges (Table 1). For example, the average degree rose to ~11.5 in 2015, then dropped back to ~8 by 2018. Clustering remained high (~0.78 on average), indicating tightly-knit co-author groups, whereas density fell (to ~3.5×10^-5 in 2015) due to the network.

Year	Nodes	Edges	Avg_Degree	Clustering	Density
2007	18877	47791	5.063410	0.786365	0.000268
2008	51321	157105	6.122445	0.770875	0.000119
2009	114186	475811	8.333964	0.789424	0.000073
2010	68953	217652	6.313054	0.779095	0.000092
2011	79326	250825	6.323904	0.784444	0.000080
2012	82745	247151	5.973799	0.782368	0.000072
2013	109140	350224	6.417885	0.784544	0.000059
2014	130720	457104	6.993635	0.788815	0.000054
2015	328505	1890200	11.507892	0.773062	0.000035
2016	204539	852082	8.331731	0.791490	0.000041
2017	216256	888554	8.217612	0.797299	0.000038
2018	192694	778536	8.080542	0.808947	0.000042
2019	57430	188144	6.552116	0.824039	0.000114
2020	20934	54734	5.229197	0.814962	0.000250
2021	11130	25339	4.553279	0.801476	0.000409
2022	6568	15985	4.867540	0.814984	0.000741
2023	6469	16283	5.034163	0.814088	0.000778
2024	3779	8972	4.748346	0.806384	0.001257

Figure: Annual Metrics

- **Degree distribution (log-log):** The degree distribution is heavy-tailed each year, consistent with power-law-like behavior (not shown). Notably, 2015's distribution has a longer tail due to many high-degree nodes in large collaborations.

- **Louvain communities & modularity:** We applied Louvain clustering to each yearly graph. Modularity scores stayed extremely high (~0.97–0.99) for most years, reflecting strong community structure. However, 2015's modularity dipped sharply (~0.86), coinciding with an explosion of cross-community edges. After 2016, modularity climbed again. Meanwhile, the number of detected communities rose from ~2007 to 2016 (peaking >20,000 communities) and then plunged after 2017 to under 2,000 by 2024. This suggests that while early growth created many small groups, recent years have seen consolidation into fewer, possibly larger communities.

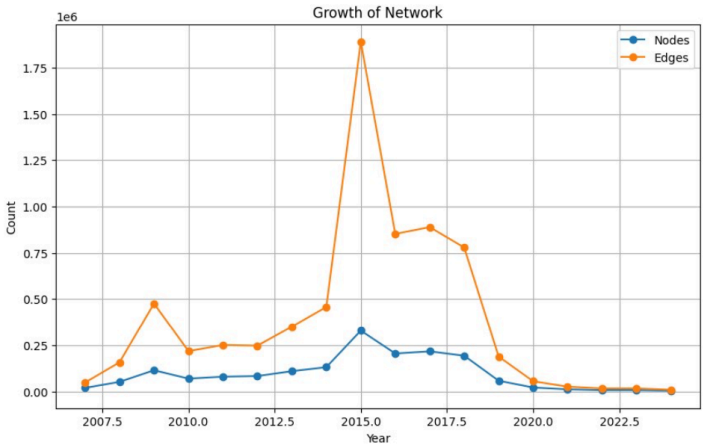
    - *Insight:* 2015 is a clear outlier – many multi-author papers weakened clear community boundaries. Post-2020, high modularity with fewer communities implies smaller, well-separated groups. Overall, co-authorship has shifted from diverse expansive clusters toward more siloed collaboration patterns.

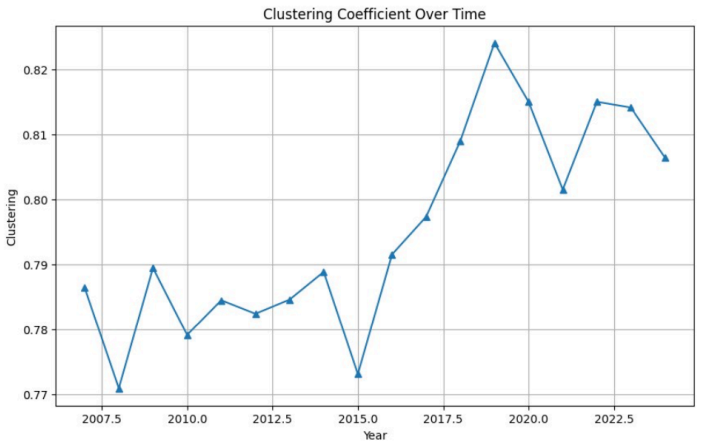
- **Community size heatmap:** We computed the distribution of community sizes (via Louvain) over time. In 2015–2017, there were many small (size ≤ 5) communities, whereas after 2018, communities fragmented into smaller pieces.

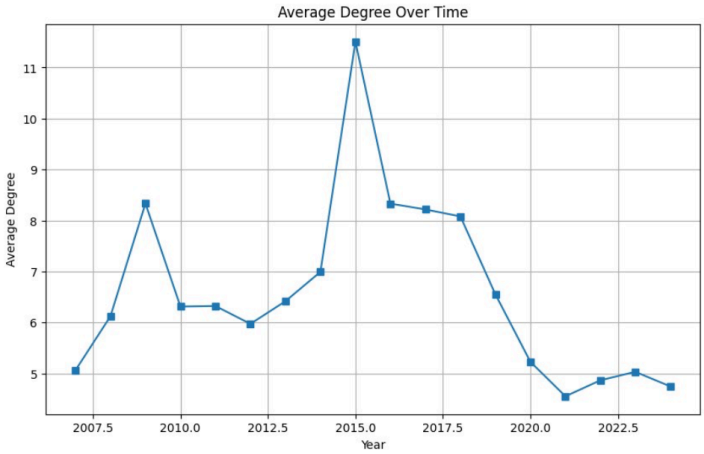
Figure: Heatmap of community size distribution over time

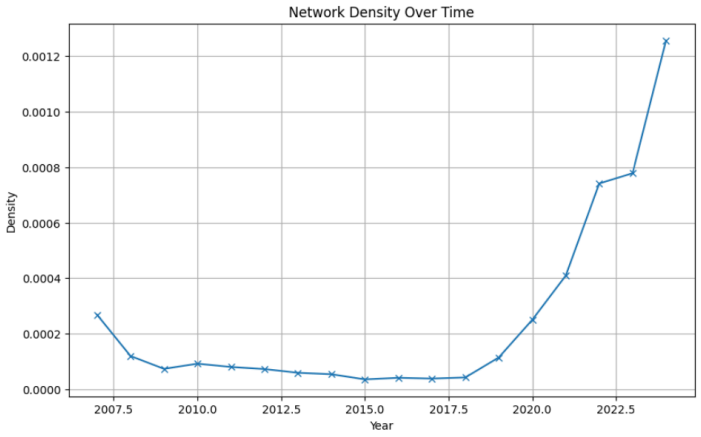
- **Largest community:** The largest community size peaked around 2015, then shrank. This, along with modularity, indicates 2015's mega-collaborations merged many authors into a single cluster, which dissolved later.

Figures & Analysis:









Analysis of Co-authorship Network Evolution (2007–2024)

1. **Growth of the Network (Nodes & Edges)**
    a. From 2007 to 2015, the network saw a sharp increase in both nodes and edges, peaking in 2015.

    b. The number of edges in 2015 crossed 1.8 million, with over 328,000 nodes, indicating a dramatic spike in collaborative publications that year.

    c. Post-2015, there's a noticeable decline, especially from 2019 onward, which could reflect incomplete data, changes in authoring patterns, or evolving publication norms.

d. The sharp fall in 2020–2024 may also be affected by fewer entries in the metadata or a lag in dataset updates.

2. **Average Degree Over Time**
    a. The average degree (i.e., the number of co-authors per author) peaked in 2015, aligning with the spike in node and edge counts.

    b. After 2015, the average degree steadily declined, reaching its lowest in 2021, indicating smaller collaborative teams or reduced connectivity in recent years.

    c. This suggests that although collaboration was high mid-decade, recent papers involve fewer co-authors per author.

3. **Clustering Coefficient Over Time**
    a. The clustering coefficient remained relatively stable until 2015, after which it increased gradually, peaking around 2019–2020.

    b. A higher clustering coefficient in later years implies more tight-knit collaborations, where authors frequently co-publish with shared partners.

    c. This may reflect a trend toward more specialized or interdisciplinary groups working closely together rather than broad, dispersed collaborations.

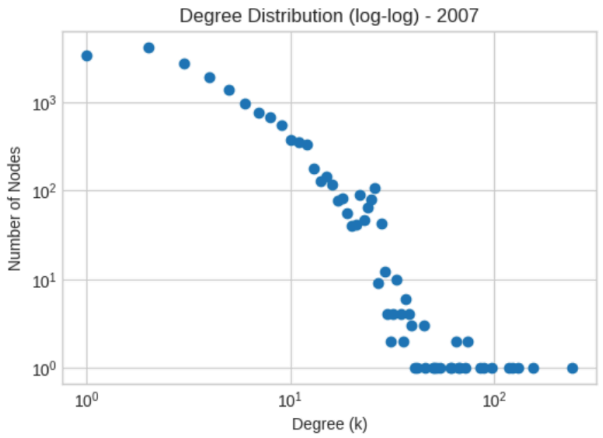
4. **Network Density Over Time**
    a. Network density was extremely low throughout the high-growth years (e.g., 2010– 2016), due to the massive size of the network.

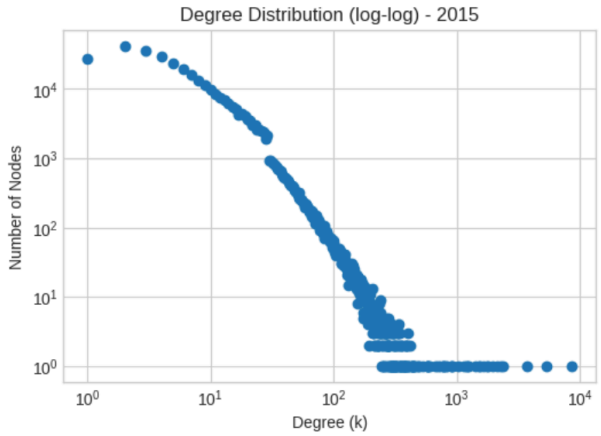
    b. However, post-2020, as node and edge counts fell, density increased significantly, peaking in 2024.

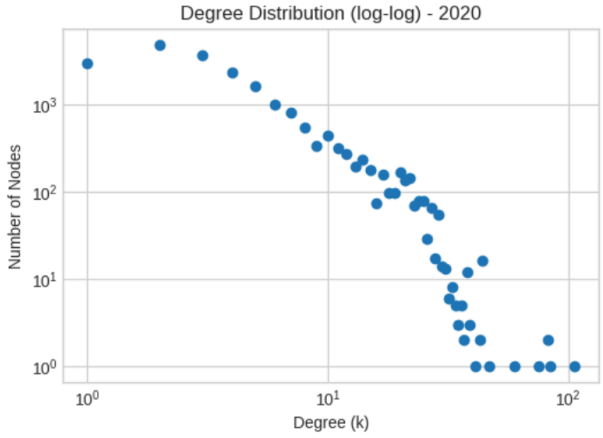
    c. This increase indicates that while fewer authors are publishing, they are more densely interconnected, likely working within established, smaller research communities.

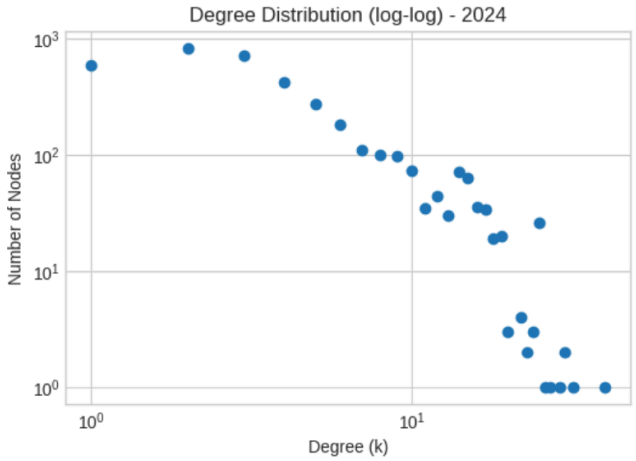
Final Understanding:

- 2015 was a landmark year for collaboration in the dataset, possibly due to large-scale projects or special publication campaigns.
- The post-2019 decline may not reflect actual drops in research but could point to incomplete metadata or lag in dataset updates.
- Recent years show tighter-knit collaborations, with fewer authors but more connected relationships, suggesting a shift toward intensive, team-based research rather than broad co-authorship networks.









Understanding from the Plots (Log-Log Degree Distribution)

1. **Graph 1 - 2007:**
    a. The plot shows a long-tailed distribution — most nodes have a low degree, but a few have significantly higher degrees.

    b. Indicates a scale-free structure early in the network's life.

2. **Graph 2 - 2015:**
    a. The most pronounced power-law-like behavior is seen here.

    b. The tail is longer, indicating the presence of many high-degree nodes (likely hubs).

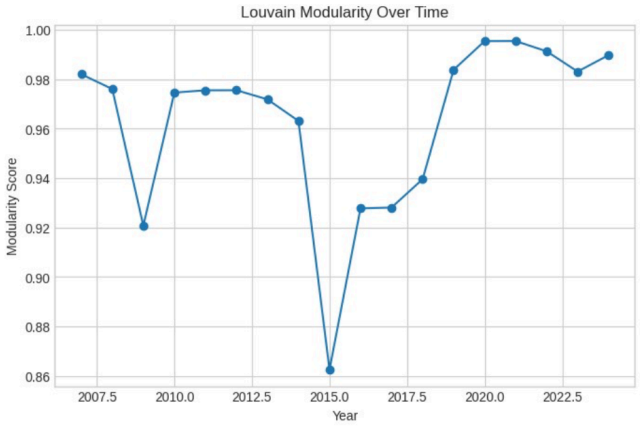
    c. The steep slope and straight-line pattern on the log-log plot strongly support a scale-free network structure.

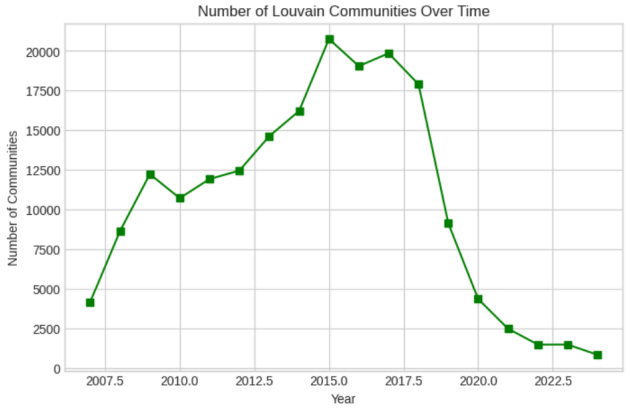
3. **Graph 3 - 2020:**
    a. Still follows a power-law, but the range of degrees has compressed due to a drop in network size (pandemic-era dip).
    b. Fewer high-degree nodes, possibly indicating fewer collaborations or fewer papers during this time.

4. **Graph 4 - 2024:**
    a. Very few nodes and degrees due to recency.• Still shows the long-tail pattern but is sparser, and deviation from the ideal power-law shape is visible due to insufficient data.

Final Understanding

These log-log degree distribution plots validate that the collaboration network across years generally exhibits scale-free properties, especially in its most active periods (e.g., 2015). Such networks are resilient to random failures but vulnerable to targeted attacks on hubs, a common trait in real-world complex systems like citation and collaboration networks.



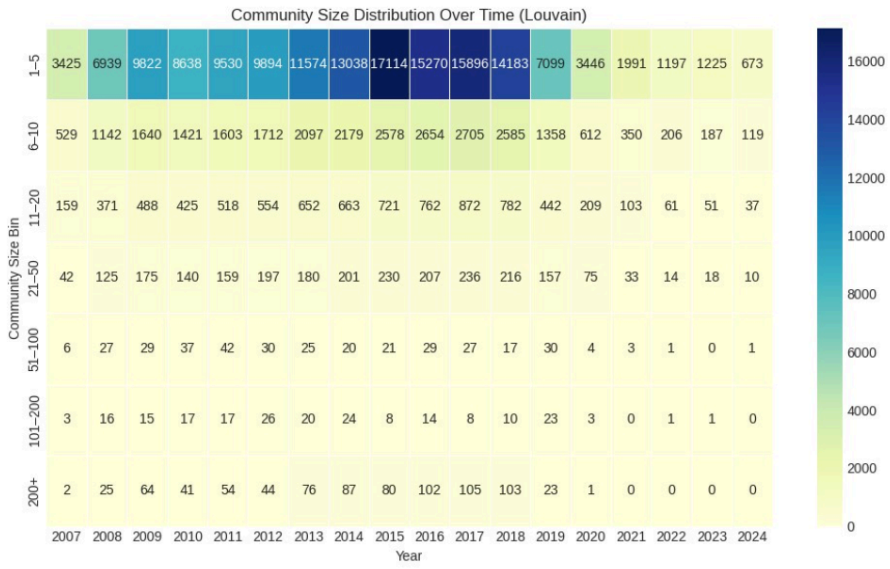


Final Understanding

- 2015 is a clear anomaly, seen in both modularity and community count — likely caused by a surge in publications with many co-authors, weakening modular boundaries.

- Post-2020, a high modularity despite fewer communities suggests smaller but well-separated collaboration groups.

- The evolving structure of co-authorship reflects changes in how research is conducted, from diverse, expansive collaborations toward possibly tighter or more siloed communities.

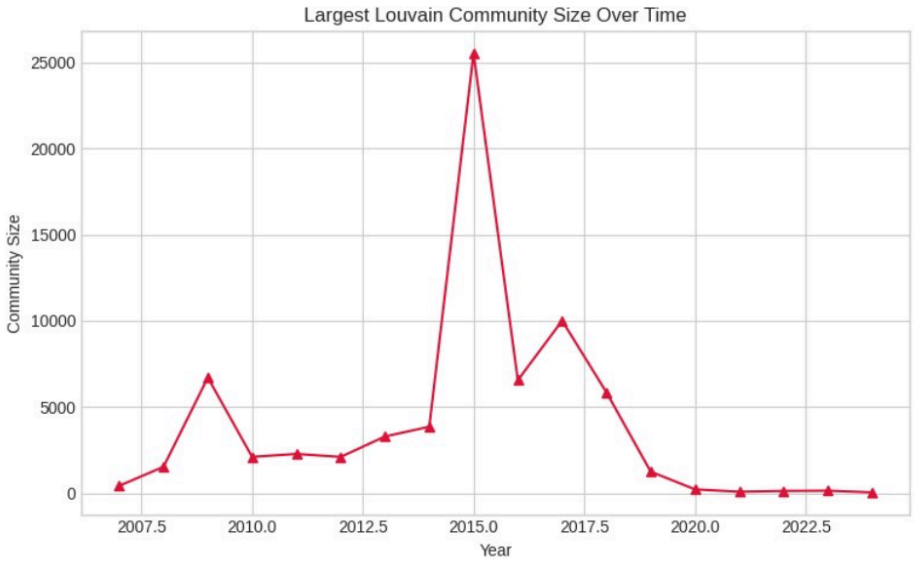


Final Understanding

- The collaboration structure of academic publishing (as captured via co-authorship networks) was highly modular with many small, tight-knit communities, especially around 2015–2017.

- Post-2018, there's a noticeable fragmentation and shrinking of community sizes, signaling a change in the collaboration dynamics—likely due to fewer active authors, tighter niche groups, or external disruptions (e.g., COVID-19).

- Overall, the network evolves from large, dense clusters to sparse, fragmented groups in recent years.



Final Understanding

- 2015 likely represents a critical structural transition in the network, possibly due to merging of communities or dataset anomalies (e.g., bulk upload, change in collaboration structure).

- Post-2018, the fragmentation of the network is clear — no dominant clusters, lower cohesion, and decentralization of research communities.

- This matches the trend observed in modularity (dip in 2015) and density/average degree (fall after 2015).

Deliverable 2: Interdisciplinarity in Co-Authorship Patterns

- **Entropy per author:** For each author, we computed the Shannon entropy of their set of arXiv fields. This measures interdisciplinarity: higher entropy means the author publishes in many different fields. Most authors have low-to-moderate entropy (specialized or bi-disciplinary); very high interdisciplinarity is rare. The entropy distribution is right-skewed, confirming a mix of specialists and generalists.

- **Top interdisciplinary authors:** Ranking by entropy, we identified the top researchers spanning diverse fields. A histogram (Figure 2) lists the top 15 authors; these "generalists" collaborate across numerous domains. For example, authors at the top have worked in fields spanning AI, physics, and finance, making them bridges in the network.

- **Field co-occurrence heatmap:** We built a co-occurrence matrix of fields (counting how often authors publish in pairs of fields) and plotted it as a heatmap. The diagonal dominates (authors tend to stay in one primary field), but bright off-diagonal spots appear for related fields. Notably, cs. AI, cs.LG, and stat.ML strongly co-occur, reflecting machine learning's interdisciplinary nature. Other hotspots include overlaps among physics subfields (e.g., cond-mat, hep-th). Sparse off-diagonal regions mark siloed disciplines.

    - *Insight:* The heatmap reveals clusters of fields that naturally collaborate, highlighting candidates for joint research. For example, the tight CS-AI-ML cluster suggests strong integration of those areas. In contrast, isolated fields have little overlap. Overall, fields that frequently co-occur may act as "knowledge bridges" fostering innovation across domains.

- **Bridge authors & cross-field edges:** In 2015's network, we identified "bridge authors" by counting the number of distinct neighbor fields each

author has. The top 10 (e.g. Wei Li, Jian Wang) each co-authored with peers from ~40–46 different fields. These authors not only publish prolifically (degree 100–400) but also in highly diverse areas, marking them as interdisciplinary connectors.

Top Interdisciplinary Bridge Authors (by neighbor field diversity):

	Author	Field	Neighbor_Field_Diversity	Degree
9077	Wei Li	q-fin.GN	46	256
14203	Jian Wang	math.GT	42	194
13473	Jr.	Unknown	42	420
5340	Jing Wang	cond-mat.mes-hall	41	283
4362	Wei Zhang	q-fin.GN	40	198
33071	Wei Chen	cs.CC	39	151
39581	J. Liu	eess.IV	39	305
13596	Wei Wang	math.GT	38	132
30650	Xi Chen	q-fin.GN	38	250
10171	Y. Li	cs.DC	36	242

Figure 3: Table of top bridge authors

- **Cross-field collaboration rate:** For 2015, ~871,935 edges with known fields were classified: 569,906 (65.36%) were cross-field, and 302,029 were same-field. We also tracked this rate annually: it has risen over time, exceeding 65% in later years, indicating growing interdisciplinarity.

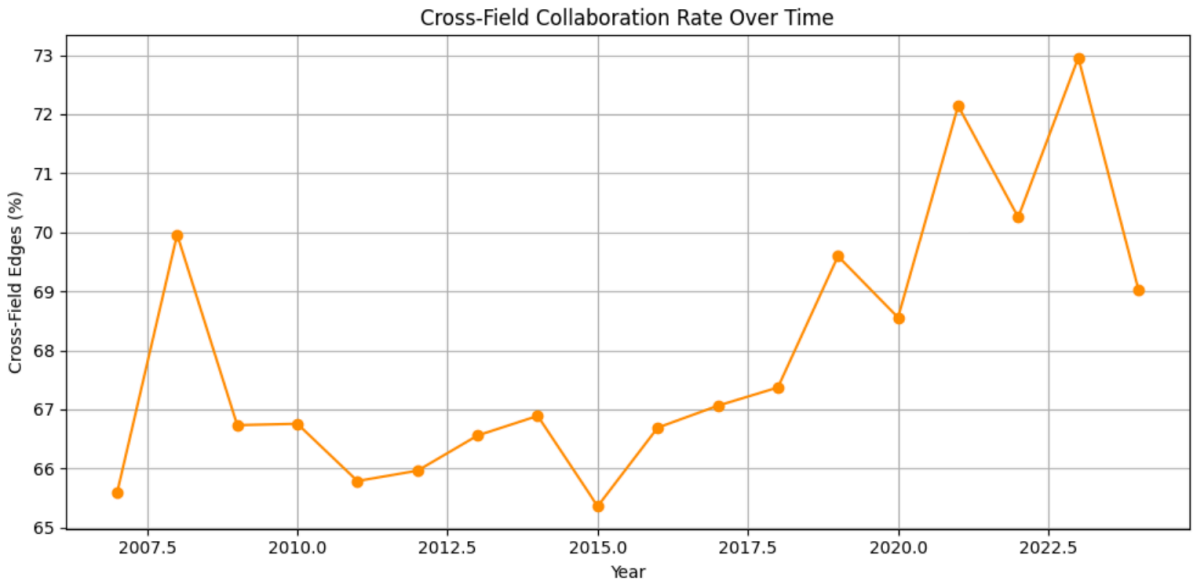
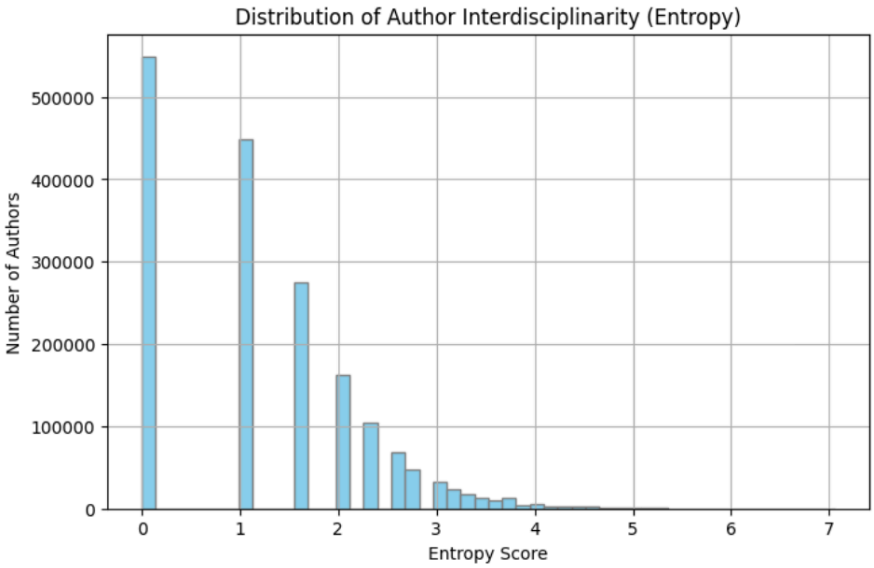


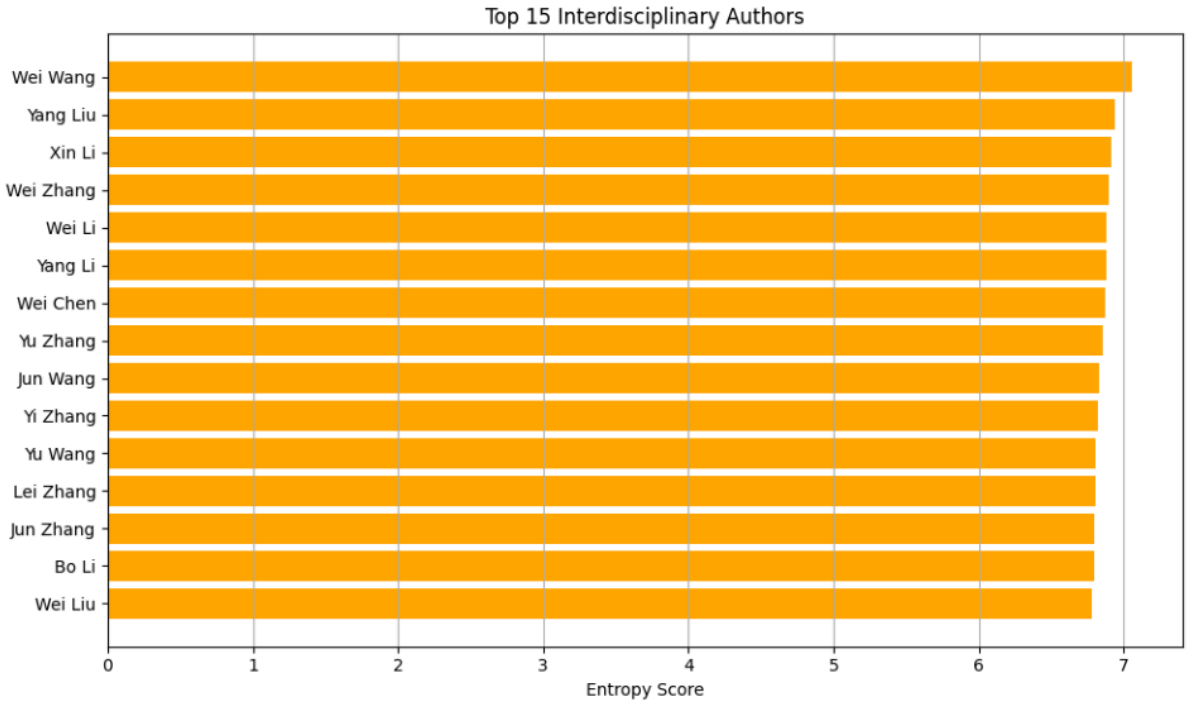
Figure: Cross-field collaboration percentage over time

- **Temporal analysis:** As years progressed, the cross-field edge percentage increased. By 2018 and beyond, more than two-thirds of collaborations spanned fields. This trend suggests that knowledge flow across disciplines is intensifying. The deliverable concludes that while many authors remain field-focused, interdisciplinary collaboration is rising, with key individuals bridging communities.

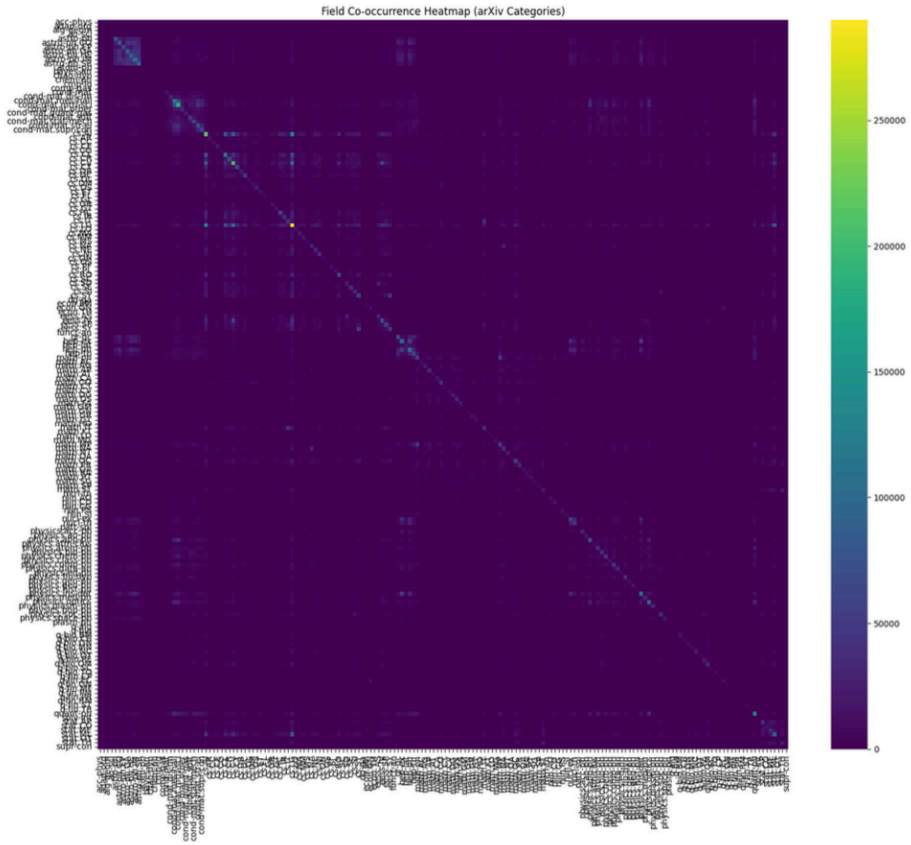
Figures & Analysis:



This histogram visualizes the distribution of interdisciplinarity scores across authors. Most authors have low to moderate entropy, indicating that while many collaborate across fields, true interdisciplinarity (high entropy) is relatively rare.



This reflects a mix of specialists and generalists in the academic ecosystem. This block ranks authors by their entropy scores to identify the most interdisciplinary researchers. These individuals likely collaborate across diverse domains, playing key roles as bridges in the academic network.

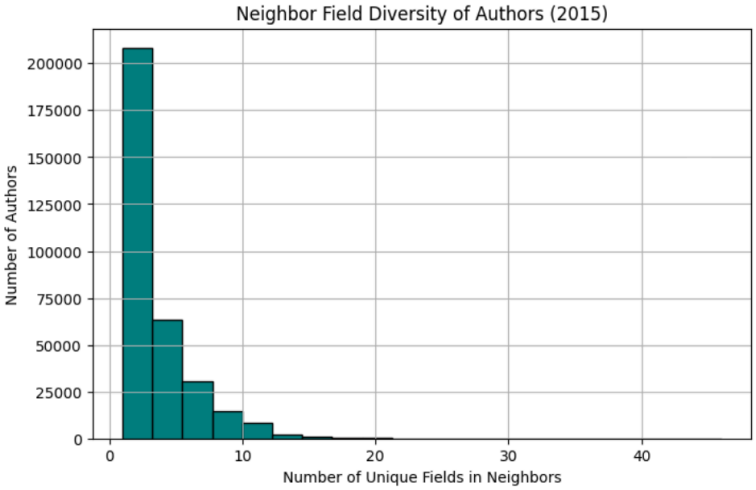


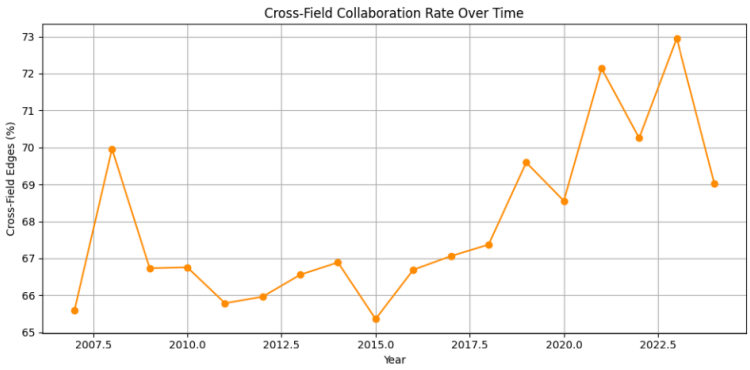
Interpretation

- The bright diagonal indicates that most authors consistently publish within their primary field — an expected result.

- Visible off-diagonal hotspots (brighter patches off the diagonal) show strong co-occurrence between:
    - cs.AI, cs.LG, and stat.ML — indicative of tight integration in machine learning and artificial intelligence research.
    - Certain combinations within physics subfields (cond-mat, hep-th, gr-qc) also show strong internal overlap.

- Sparse regions represent siloed disciplines, where authors rarely publish across boundaries.





This deliverable demonstrates that interdisciplinarity in academia is both quantifiable and visibly rising. While most authors remain field-specific, structural metrics show that:

- Cross-field collaboration is growing,
- Key authors act as bridges,
- Certain fields consistently interact more than others.

These insights are foundational for understanding how knowledge flows across disciplines, shaping the future of collaborative science.

Deliverable 3: Link Prediction

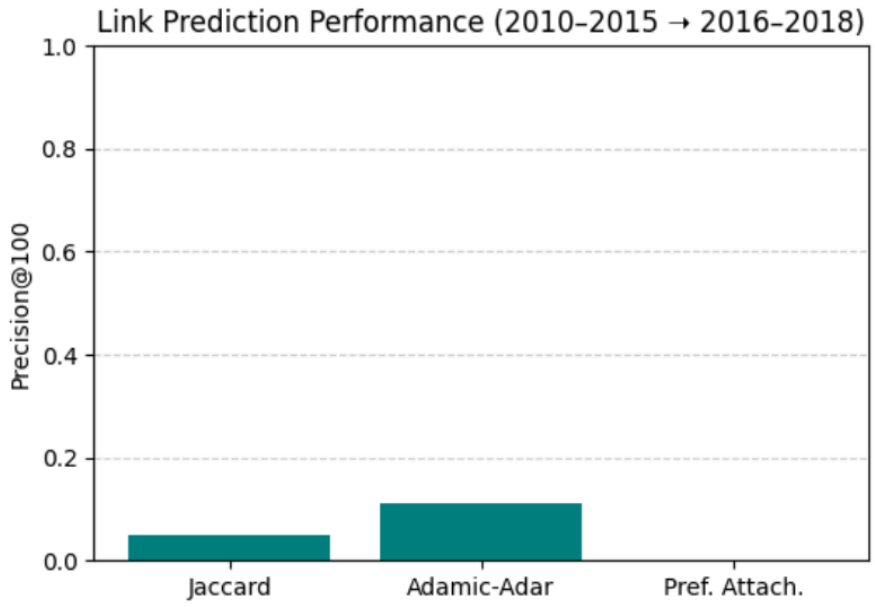
- **Train/test setup:** We built a training graph from 2010–2015 papers (3,109,445 edges) and a test set of future edges from 2016–2018 (553,103 potential new links). Only edges among nodes present in the training graph were considered as candidates.

- **Heuristics used:** Three classic link-prediction scores were computed on all candidate node pairs: Jaccard coefficient, Adamic–Adar, and Preferential Attachment. These use only graph structure: common neighbors for Jaccard/AA, and degree product for PA.

- **Evaluation (Precision@100):** We ranked predicted edges by each score and measured how many of the top 100 truly appeared in 2016–2018. Adamic–Adar achieved Precision@100 = 0.110 (11 correct predictions), Jaccard 0.050 (5 correct), while Preferential Attachment yielded 0.000 (no correct links).

- **Results:** Adamic–Adar outperformed the other heuristics, confirming that shared neighbors matter in predicting collaborations. Jaccard was modestly effective. The failure of Preferential Attachment suggests that simply linking the highest-degree nodes does not capture academic collaboration patterns.

- **Insight:** These results show that basic graph features can partly predict future co-authorships, but overall performance is low (~10% accuracy at best). This validates that neighborhood overlap is relevant, but naive degree-based methods are ineffective in this structured network. It motivates more advanced models (e.g., node embeddings or supervised approaches with graph neural nets) for better prediction.



Deliverable 4: Case Study – Machine Learning (cs.LG)

- **Yearly trends (2010–2024):** The cs.LG graph shows explosive growth. Unique authors in cs.LG rose from under 1,000 in 2010 to over 100,000 by 2024. Edges (collaborations) grew similarly, exceeding 400,000 by 2024 . Average degree climbed from ≈3 to over 8, indicating authors co-authored with more peers over time.

- **Clustering & density:** The average clustering coefficient has stayed very high (>0.8 since ≈2015), meaning cs.LG authors form tight collaborative groups. In contrast, graph density plummeted (to ~0.0001 by 2024), a common effect in large networks: although many authors join, the fraction of possible edges remains tiny. In sum, cs.LG is large, tightly clustered, and low-density.

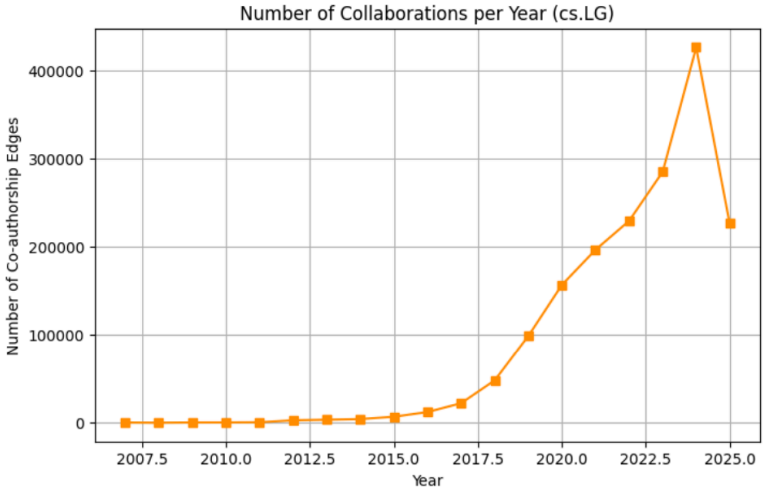
- **Community structure:** Early cs.LG communities were numerous and small, but over time, they merged into larger, more cohesive components. Louvain modularity stayed very high (≈0.9+) until 2018, then fell sharply to ~0.5–0.6 after 2020. This implies that sub-communities in ML (e.g., vision,

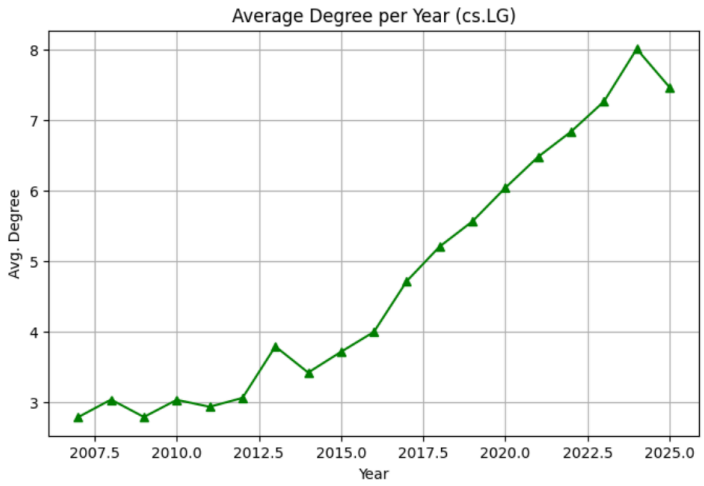
NLP, theory) were initially separate but later became more interconnected. The summary describes a “transition from many fragmented research islands to a large, interconnected continent” as cs.LG matured.

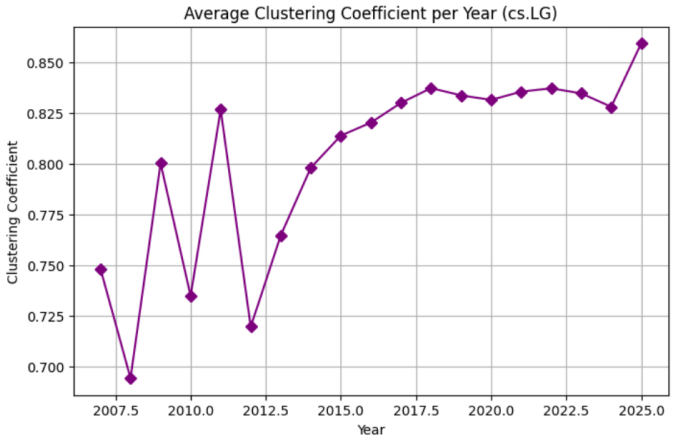
- **Cross-field vs intra-field (2018):** We counted edges in 2018 that were intra-field (both authors in cs.LG) vs cross-field (one author’s dominant field ≠ other’s). Out of all edges, ~255K were intra-field and ~523K cross-field in 2018. Thus, about 67% of cs.LG collaborations were cross-field, highlighting that ML research heavily draws from and contributes to other domains. *[Insert Figure 5: Intra- vs. cross-field edges (2018)].* This aligns with the drop in modularity and rising average degree: ML’s growth is fueled by interdisciplinary partnerships.

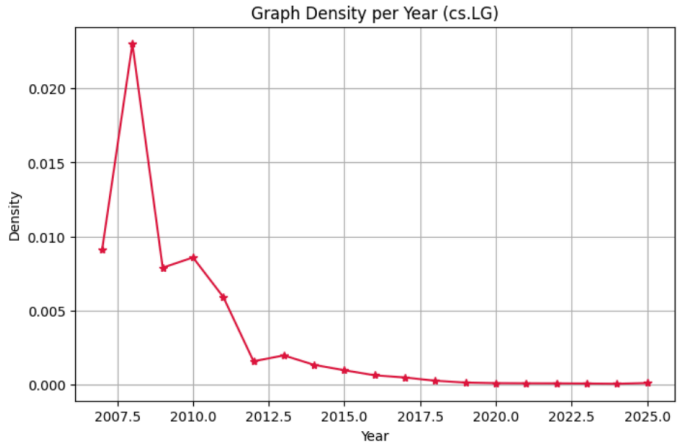
- **Inter-field co-authorship heatmap:** We built a log-scaled heatmap of inter-field cs.LG co-authorships. (Not shown here.) It further confirms strong links between cs.LG and fields like cs.CV, stat.ML, as well as surprising ones (e.g. ML+physics, ML+biology).

- **Summary:** cs.LG has evolved from a niche into one of arXiv’s largest, most collaborative areas. The field is characterized by many tight collaboration clusters (high clustering, low density) and a shift from siloed subtopics to broadly integrated research. As noted, “Machine Learning has evolved from a niche topic into a collaborative, interdisciplinary hub” that bridges multiple domains.









Number of Authors per Year (Nodes)

- Observation: The number of unique authors publishing in cs.LG shows a dramatic increase from fewer than 1,000 in 2010 to over 100,000 authors in 2024, with a slight drop in 2025 (possibly due to incomplete data).

- Interpretation: This exponential rise reflects the explosive growth of machine learning as a research area, drawing contributors from across academia, industry, and adjacent disciplines.

Number of Collaborations per Year (Edges)

- Observation: Co-authorship edges increase in tandem with author count, peaking at over 400,000 collaborations in 2024.

- Interpretation: Collaboration has become increasingly common and dense, indicating that ML research is highly team-oriented and globally distributed.

Average Degree per Year

- Observation: The average degree rises steadily from ~3 (pre-2015) to over 8 by 2024, showing that authors are collaborating with more co-authors on average.

- Interpretation: This supports the notion that ML papers now typically involve larger teams and that researchers are more interconnected than ever before.

Average Clustering Coefficient

- Observation: Clustering coefficient has remained consistently high (above 0.8 since~2015), with slight annual fluctuations.

- Interpretation: This suggests that collaboration in cs.LG tends to occur in tight-knit groups, where an author's collaborators are also likely to be collaborators with one another, a hallmark of community structure in scientific networks.

Graph Density Over Time

- Observation: Density declines sharply as the network grows, dropping to nearly 0.0001 by 2024.

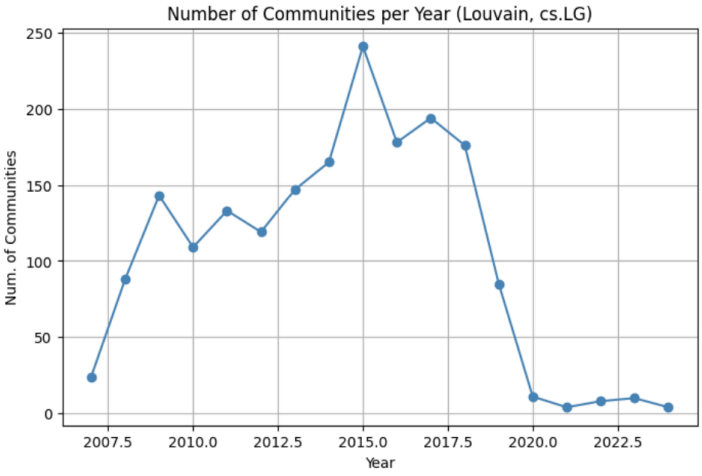
- Interpretation: This is expected in large graphs: while the number of authors and edges increases, the number of possible edges grows quadratically, leading to sparser networks. It reflects that while collaboration grows, authors do not form edges with everyone, the field still has subfield boundaries.

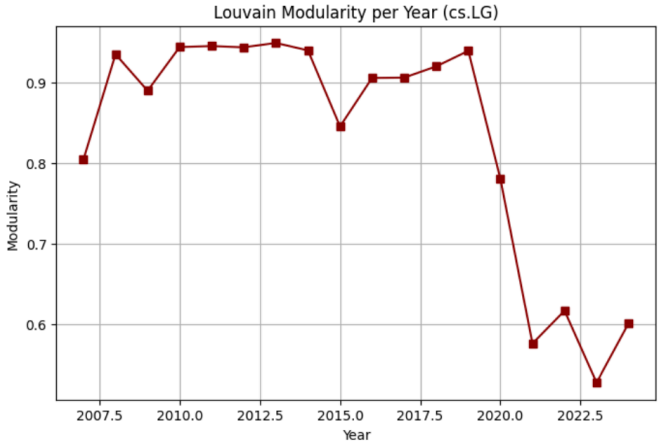
Summary

- cs.LG has evolved from a niche research category into one of the largest and most collaborative domains on arXiv.

- The field exhibits a high clustering, low-density structure, indicating localized, but tight collaboration clusters.

- The sharp rise in average degree and clustering implies that interdisciplinary, multi-author projects are now the norm in ML research.





Number of Communities per Year

- Observation: The number of communities peaked in 2015 at around 240 clusters, but sharply dropped after 2019, reaching single-digit clusters by 2024.

- Interpretation: Initially, the field exhibited fragmented structure with many small clusters. But over time, communities merged, leading to larger, more cohesive components, possibly due to increasing interdisciplinary collaboration and convergence of subfields in ML.

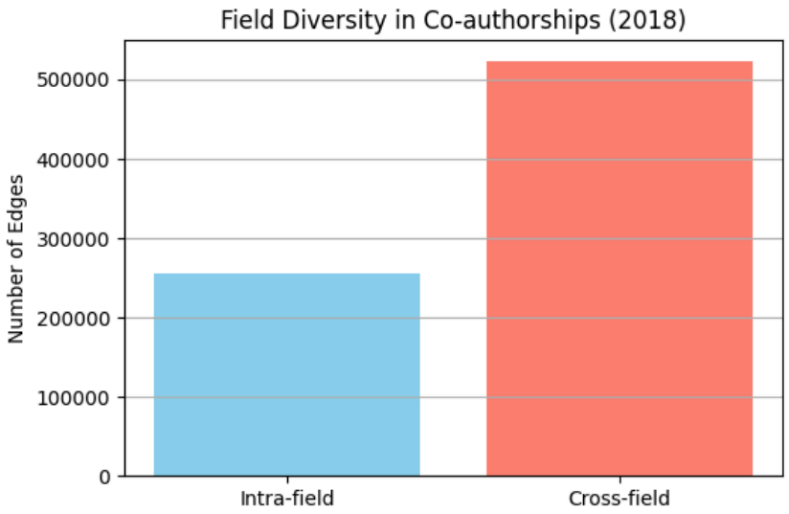
Louvain Modularity per Year

- Observation: Modularity remained high (above 0.9) until ~2018, indicating well-separated topical clusters. It then declined sharply post-2020, dropping to 0.5–0.6, suggesting more inter-community mixing.

- Interpretation: This suggests a structural flattening of the community landscape, as machine learning matures, subfields interconnect more, reducing modularity. This may reflect broader trends like:
    - Growth of integrative ML topics (e.g., multimodal learning, ML+biology)
    - Increased cross-disciplinary team composition

Summary

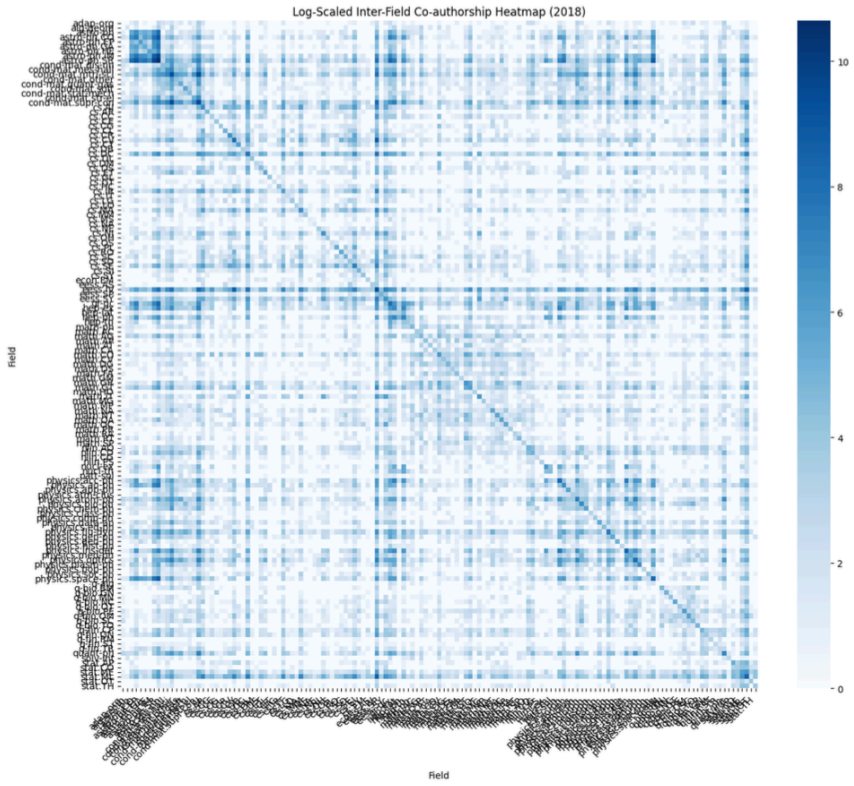
The evolution of cs.LG reveals a transition from many fragmented research islands to a large, interconnected scientific continent. The drop in modularity highlights the collapse of strict subfield silos in favor of integrated, cross-domain collaboration.



- ~67% of collaborations in 2018 were cross-field, indicating a strong interdisciplinary nature in the research ecosystem.

- The cross-field collaboration rate significantly outweighs same-field partnerships, which aligns with earlier findings on modularity decline and increasing average degree.

The machine learning field (cs.LG) is not only growing in volume but also in conceptual diversity, acting as a hub for cross-domain research involving computer vision, statistics, physics, and even biology.



- Diagonal intensity indicates strong intra-field collaboration — most pronounced in high-volume categories like cs.LG, cond-mat, and hep-th.

- Off-diagonal brightness highlights frequent cross-field collaboration, especially among:
    - cs.LG and stat.ML (machine learning ↔ statistics)
    - cond-mat and quant-ph (physics domains)
    - cs.AI, cs.CV, and eess.IV — evidence of ML bridging computer vision and embedded systems

The ML domain acts as a cross-disciplinary glue, connecting computational sciences (CS, statistics) with domains in physics and engineering. The dense

inter-field connections validate earlier metrics showing high cross-field collaboration rates.

Deliverable 5: Field-Level Centrality Analysis

- **Avg centrality by field (2018):** We computed each field's average degree centrality in 2018. The top fields (highest average centrality) include **astro-ph.SR (Solar and Stellar Astrophysics)**, CS.GL (General CS), and q-fin.GN (Quantitative Finance – General). This means on average, researchers in those fields co-author with more peers. The full list of the top 20 fields is shown in Figure 6. Higher average centrality suggests fields whose researchers are more interconnected. Astrophysics (Solar & Stellar) leads, implying its authors form especially dense collaboration clusters.

- **Field avg centrality insights:** The mix of top fields (astrophysics, CS, finance, condensed-matter, etc.) indicates that high connectivity spans diverse disciplines. These central fields likely act as "knowledge hubs," accelerating diffusion and interdisciplinary exchanges.

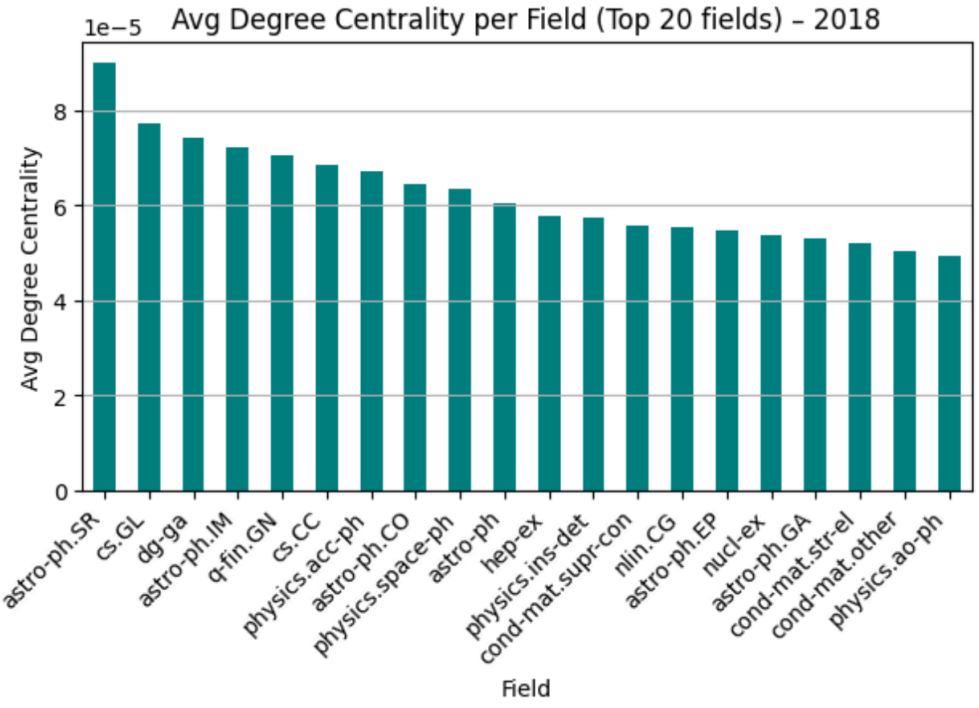


Figure: Average degree centrality by field (top 20 fields, 2018)

- **Top 100 authors by centrality:** Among the 100 authors with the highest degree centrality in 2018, we mapped their fields. Astro-ph.SR dominates overwhelmingly, with 32 of the top-100 authors in that field. Fields like EESS.IV (Electrical Eng.) and CS.CC (Computational Complexity) also appears (5–6 authors each). Many fields have only a few representatives, showing diversity. We filtered out ~18% whose field was unknown (metadata gaps).

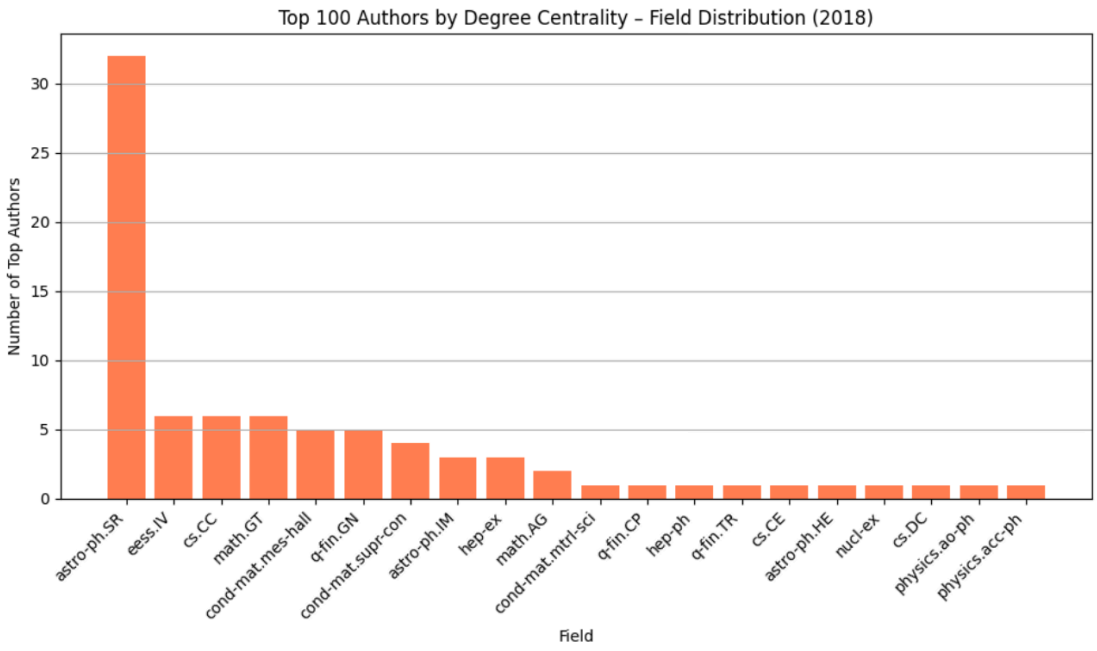
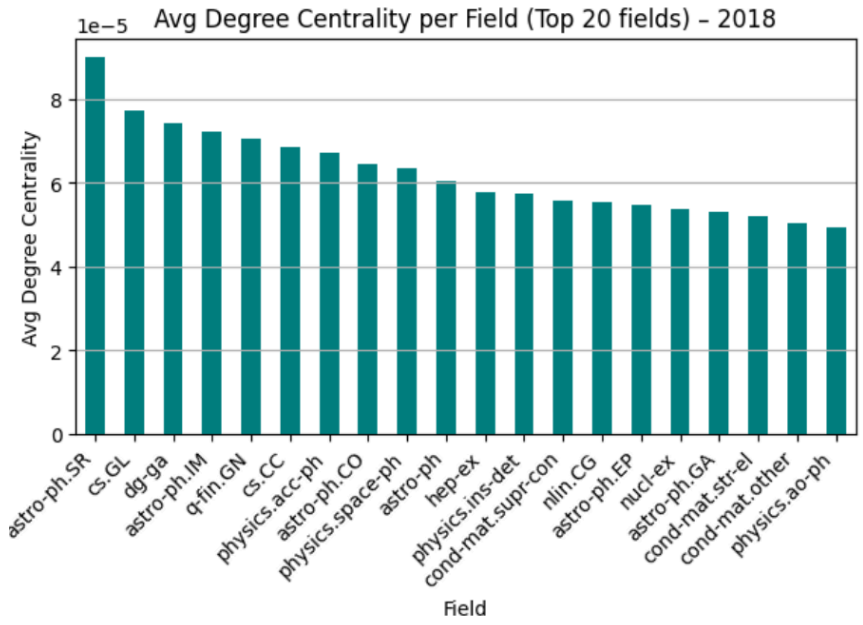


Figure: Bar chart of fields of top-100 central authors

- **Field distribution insight:** The central authors span astrophysics, computing, finance, math, and more. Notably, the dominance of astro-ph.SR (32%) reflects its high collaborative intensity. This suggests that a few fields disproportionately contribute to central researchers. In summary, fields like solar/stars astrophysics and general CS produce the most interconnected authors, while others are less represented.

- **Conclusion (Field Centrality):** A small set of fields generates a large fraction of central authors in the network. This may stem from higher publication norms or tight collaboration cultures in those fields. These central fields likely influence the entire network structure.

Figures & Analysis:



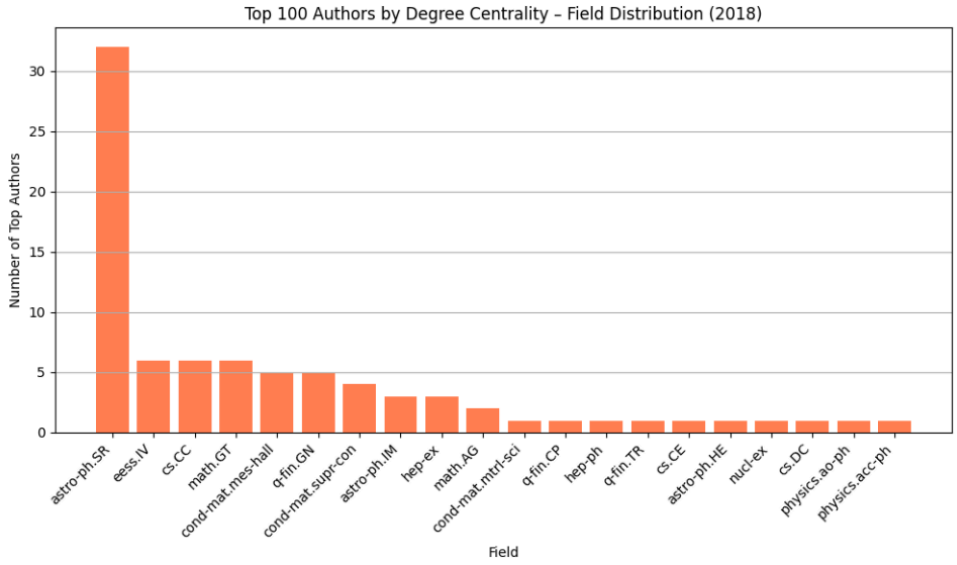
This bar chart ranks the top 20 academic fields based on their average degree centrality in the 2018 co-authorship network.

- Interpretation of Degree Centrality: A higher average degree centrality in a field suggests that, on average, researchers in that field collaborate with more co-authors, making them more "central" in the network.

- Top Fields:
    - astro-ph.SR (Astrophysics - Solar and Stellar) ranks highest, suggesting that its researchers were the most interconnected in 2018.
    - Fields like CS.GL (General Computer Science) and q-fin.GN (General Quantitative Finance) also appear high, indicating strong collaborative behavior within those domains.

    - Diversity in Top Fields: The top fields include a mix of astrophysics, computer science, finance, and condensed matter physics, showing that high connectivity is not isolated to one discipline.

- Implication: Central fields might act as bridges or influencers in the research community, potentially accelerating the diffusion of knowledge.



Unknown field ratio: 18.00%

This bar chart highlights the field-wise distribution of the top 100 authors by degree centrality in the 2018 co-authorship network.

- Dominant Field: astro-ph.SR (Astrophysics - Solar and Stellar) overwhelmingly dominates, with 32 out of the top 100 authors. This aligns with its high average degree centrality and suggests a highly collaborative and tightly-knit author cluster in that field.

- Other Significant Fields: Fields such as eess.IV (Electrical Engineering - Image and Vision), CS.CC (Computational Complexity), and q-fin.GN (Quantitative Finance - General) show moderate presence with 5–6 authors each, indicating their growing influence and collaboration density in 2018.

- Field Diversity: A wide variety of fields are represented, from mathematics to condensed matter physics, though most have only 1–3 highly central

authors.

- Unknown Field Ratio:About 18% of top authors could not be mapped to any known field, possibly due to missing or inconsistent metadata. This may introduce some bias in field-level insights and points to limitations in field attribution.

Discussion and Key Insights

Our analysis uncovers several overarching themes:

- **Network growth and structure:** The co-authorship network expanded rapidly up to the mid-2010s. High clustering (~0.78) and modularity (~0.97) indicate that authors generally collaborate in tight-knit groups. The 2015 anomaly (many authors/papers) temporarily changed this pattern.

- **Fragmentation over time:** After 2015 ,the network fragmented into more but smaller communities, even as total authorship declined. This suggests shifting collaboration norms (e.g. smaller teams) in recent years. 2020s communities are smaller and more isolated (despite high modularity) .

- **Interdisciplinarity:** Collaborations increasingly cross field boundaries. Over 65% of edges in later years connect different fields. Key fields (cs.LG, stat.ML, cs.CV, etc.) form a well-connected cluster, as shown by heatmaps. The rising cross-field rate implies that knowledge is flowing across disciplines more than before. This trend is underscored by cs.LG's growth into an interdisciplinary hub.

- **Centrality and fields:** A few fields and authors dominate connectivity. Astro-ph.SR stands out as the most connected field. Top authors are largely from astrophysics and a handful of CS/engineering fields. This concentration suggests these fields (and their social norms) drive much of the network's core structure.

- **Predictive modeling:** Simple heuristics capture only limited signal: neighborhood-based methods (Adamic–Adar) have mild success, but popularity-based (Preferential Attachment) fails. This highlights the importance of shared context (common collaborators) in science, and points to the need for richer models (e.g. embeddings, GNNs) for link prediction.

These findings emphasize that academic collaboration is highly structured and field-dependent. The network's evolution reflects changing research practices: from large, diverse teams (mid-2010s) to more focused, though still interconnected, clusters. Interdisciplinary bridges and central fields play outsized roles in shaping connectivity.

Conclusion and Future Work

This report provides a comprehensive, data-driven view of arXiv co-authorship dynamics. We quantified growth, community structure, interdisciplinarity, and centrality in a unified framework. Key conclusions include the identification of 2015 as a structural anomaly, the quantification of rising cross-field collaboration, and the central roles of certain fields (like astrophysics) and individuals.

Future directions: Building on these insights, future work could explore more sophisticated link-prediction models (e.g. node embeddings, supervised learning, graph neural networks) to improve accuracy. Further, analyzing author career trajectories or institutional effects could deepen understanding of collaboration drivers. As data quality improves (better field assignment, disambiguated authors), additional nuances in interdisciplinarity and network evolution will emerge. Our findings lay a foundation for such extensions and for policies that foster productive research networks.