The Science of Science: A Network Study of Scholarly Collaboration at the University of Sydney

Nathan Inkiriwang
University of Sydney
nathan.inkiriwang@sydney.edu.au

Yi Zhe Ng
University of Sydney
yizhe.ng@sydney.edu.au

Abstract

This study examines the co-authorship network at the University of Sydney, mapping scholarly collaboration across faculties and schools. Using our newly developed CoAuthorNet package—which automates data gathering and network construction from author names and co-authorship data—we generated a bipartite network of authors and publications. Our analysis reveals unique intra- and inter-faculty collaboration patterns, centrality measures, and network resilience, with central authors playing key roles in interdisciplinary connections. These insights contribute to understanding collaborative dynamics within academic institutions and guide strategies for enhancing research productivity and innovation.

1 Introduction

Understanding the structure and dynamics of academic collaboration is crucial for advancing scholarly communication. Co-authorship networks, which map connections between researchers through joint publications, provide insights into how knowledge is created, disseminated, and shared within academia. These networks reflect individual collaborations and reveal broader structural dynamics that drive research productivity, interdisciplinary exchange, and innovation.

Recently, co-authorship network analysis has gained traction as a tool to explore collaboration patterns, identify influential contributors, and assess the impact of institutional and disciplinary structures on research output. Such studies reveal the network properties that shape information flow and idea exchange within academic institutions. However, institution-specific studies are needed to capture local nuances in collaborative behaviors.

This study examines the co-authorship network at the University of Sydney, a leading research institution with a diverse academic landscape. By analyzing network properties across faculties and schools, we aim to identify structural features that characterize research collaboration at this institution. Specifically, we investigate centrality measures, community detection, and node vulnerability to understand how research partnerships form, the extent of cross-disciplinary collaborations, and the roles of central authors within the network. This analysis contributes to a deeper understanding of the University of Sydney's academic ecosystem and provides insights that could inform strategies for enhancing collaboration and innovation.

2 Network

2.1 Data Scraping Process

To collect data on academic staff and their publications, we employed a web scraping approach using Python libraries such as `BeautifulSoup`, `Selenium`, `pandas`, and `scholarly`. The process involved two main steps: web scraping staff information from the University of Sydney's website and retrieving publication data using the `scholarly` library.

2.1.1 Web Scraping Academic Staff Information

We used `Selenium` to automate the process of navigating and extracting data from academic staff web pages. The URLs of staff directories were compiled for different faculties, including Engineering, Science, Health, Arts, Architecture, Law, Business, and Music.

There were issues with a few publication titles with LaTeXand some with other languages in their titles which can't be encoded as a string on `Python`. This only accounted for about 100 papers out of the $300,000$ publications so we decided to be conservative and removed them. The script followed these steps:

Algorithm 1 Scrape and Process Academic Staff Data

Input: List of faculty webpage URLs
Output: CSV file with extracted staff and publication data

```
for each faculty webpage do
    Open the page and load all content  Parse the page to extract staff names and profile links  Store extracted information for each staff member
for each staff member do
    if staff member has not been processed then
        Search for the author's profile  if profile is found then
            Retrieve titles and publication years  Append publication data for the staff member
        else
            Log that no publications were found
Convert all extracted data to a structured table  Save the table to a CSV file
```

2.2 Network Construction Process

Using the data frame, we formed a bipartite network with one side being the authors and the other side being the publications. Relevant `Python` code is shown in Appendix B.1. This network captures co-authorship relation-

ships between publications but lacks information regarding inter-publication relationships. By projecting this network, we generate a simplified graph representing authors as nodes, with edges denoting at least one collaborative publication between each pair of authors.

Algorithm 2 Create Bipartite Network of Authors and Publications

Input: Data containing authors and their publications
Output: Bipartite graph linking authors to publications
Initialize an empty graph Extract unique authors and assign each a unique ID Add each author as a node in the graph
for *each record in the data* **do**
    Check if the publication already exists in the graph **if** *publication does not exist* **then**
        Add the publication as a new node
    Link the author to the publication by adding an edge between them
return *Graph with author and publication nodes and edges linking them*

Algorithm 3 Generate Author Collaboration Network

Input: Bipartite graph of authors and publications
Output: Graph of author collaborations
Initialize an empty author network Add each author from the bipartite graph as a node in the author network
for *each publication in the bipartite graph* **do**
    Identify all authors linked to the publication **for** *each pair of authors linked to the publication* **do**
        if *no existing link between the authors* **then**
            Add an edge between the authors to indicate collaboration
return *Author collaboration network*

The `Python` code is shown in Appendix B.2.

2.2.1 The Author Network

The author network consists of 2722 nodes, 16867 edges and 365 components. In further analysis, we will focus on and refer to the largest component and with 2317 nodes (85.12% of the full network) and 16821 edges (99.73% of the full network).

2.2.2 The Faculty Networks

To analyze collaboration patterns within each faculty, we segmented the full author network of the University of Sydney into distinct faculty-specific sub-graphs. For each faculty, we identified authors associated with that faculty. By filtering the main network G to retain only the nodes corresponding to authors in a given faculty, we generated sub graphs representing individual faculty networks. Within each faculty network, we extracted the largest connected component to ensure the analysis focuses on the primary collaboration network, removing isolated or marginally connected nodes.

2.2.3 Network Statistics

The network statistics presented in Table 1 include various key metrics that characterize the collaborative structure

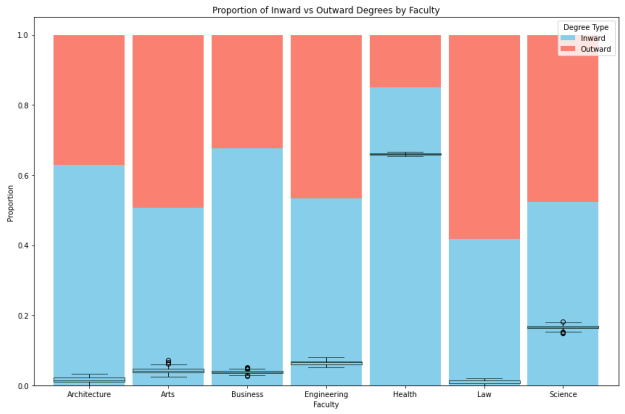

Figure 1: Proportion of Inward vs. Outward Degree by Faculty

across faculties. N represents the number of nodes (authors), and L denotes the number of links (co-authorship). The average degree, $\langle d \rangle$, reflects the mean number of connections per node, while σ_d is the degree standard deviation, indicating variability in connectivity. The network clustering coefficient (transitivity), C_{net}, measures the degree of local interconnectedness among authors. The average path length, $\langle\langle t \rangle\rangle$, provides the mean shortest path between nodes, and t_{diam} represents the network's diameter, capturing the longest shortest path within each faculty network. These statistics collectively illustrate collaboration density and structure across disciplines.

3 Analysis

3.1 Comparing Collaboration Within and Across Faculties

This section aims to examine how collaboration patterns differ within individual faculties compared to across faculties, providing insights into the extent of interdisciplinary interaction and the structural role of each faculty in the broader academic network. To distinguish between intentional faculty-based collaboration patterns and those that might arise randomly, we used an Exponential Random Graph Model (ERGM) as a baseline.

This model, implemented via a double-edge swap algorithm, preserves the network's degree distribution while randomizing its structure, allowing us to assess the expected proportions of inward and outward connections if faculty affiliations were not a factor. After an initial discard phase of $20,000$ swaps to ensure thorough mixing, we generated 50 independent random graphs with $10,000$ lagged swaps between each, providing a robust baseline for expected degree proportions under random conditions. This ERGM-based baseline enables a direct comparison to our empirical network, highlighting deviations that suggest non-random, faculty-centric collaboration patterns.

The resulting Figure 1 shows empirical proportions of inward and outward degrees by faculty, compared a gainst the expected values (black box plots) derived from our ERGM. The substantial deviation, particularly the higher-than-expected inward degrees and lower outward degrees, indicates that our observed network structure is not random. Specifically, faculties exhibit a stronger tendency

Faculty	N	L	$\langle d \rangle$	σ_d	C_{net}	$\langle\langle t \rangle\rangle$	t_{diam}
Engineering	226	639	5.65	4.56	0.34	4.27	11
Science	414	1525	7.37	5.38	0.32	4.11	10
Health	1099	9447	17.19	16.41	0.21	3.06	8
Arts	187	348	3.72	2.84	0.34	5.31	12
Architecture	56	176	6.29	4.40	0.65	3.61	9
Business	176	436	4.95	3.93	0.39	4.44	10
Law	35	63	3.60	2.40	0.54	3.47	7
University of Sydney	2317	16821	14.52	16.17	0.22	3.83	11

Table 1: Network statistics by faculty at the University of Sydney

Figure 2: Correlation Between H-index and Centrality Measures

Faculty	Betweenness	Degree	Page Rank
Architecture	0.58	0.52	0.56
Arts	0.44	0.47	0.48
Business	0.43	0.58	0.52
Engineering	0.49	0.52	0.56
Health	0.56	0.64	0.66
Law	0.27	0.28	0.36
Science	0.49	0.47	0.54

Table 2: Correlation coefficients for different centrality measures across faculties.

toward internal collaboration, which implies organizational or structural biases favoring intra-faculty over inter-faculty connections, reflecting deliberate, faculty-centric collaboration patterns rather than random interactions.

3.2 Centrality and Academic Impact

In our study, we selected three centrality measures—Betweenness, Degree, and Page Rank—to assess the influence of authors within the University of Sydney author network, each offering distinct insights. Betweenness identifies authors who bridge disparate areas, fostering interdisciplinary connections. Degree measures direct connections, indicating authors with high collaboration frequency and potential research volume. Page Rank evaluates both the quantity and quality of connections, highlighting authors who collaborate with other influential researchers, thus amplifying their visibility and impact within the network.

3.2.1 Correlation Across All Faculties

Figure 2 shows the relationship between h-index and three centrality measures across all faculties within the University of Sydney author network. The red regression lines reveal a moderate positive association, indicating that authors with higher centrality tend to have higher h-index values. However, the moderate strength of these correlations suggests that while centrality is linked to academic impact, it explains only part of the variation in h-index. Additional factors likely influence h-index beyond an author's network position.

Each plot shows statistical significance ($p < 0.05$), affirming that these associations are unlikely to be due to chance. This validates centrality as a meaningful, though partial, indicator of influence within the author network. This aligns with previous findings that centrality is valuable for identifying network positions of influence but is not a sole predictor of academic success, especially in interdisciplinary networks where impact may be shaped by various individual and contextual factors.

3.2.2 Correlation Between Faculties

Several notable patterns emerge across faculties when we look at Table 2. Health exhibits high correlations across all measures, particularly Page Rank (0.66) and Degree (0.64), suggesting that h-index in Health is closely tied to both high connectivity and collaboration with influential researchers. This aligns with the collaborative, multi-author nature of health sciences research. Architecture and Engineering show moderate correlations, especially in Betweenness and Page Rank (with Betweenness reaching 0.58 in Architecture), indicating that influential authors in these fields often bridge distinct subfields, potentially enhancing interdisciplinary influence. In contrast, Law exhibits weaker correlations, with a maximum of 0.36 in Page Rank, suggesting a fragmented collaboration structure where impactful authors often work independently or in smaller teams, reflecting disciplinary norms in legal research.

3.3 Node Vulnerability Assessment

The sensitivity of a network against attacks can help describe the structure of the connectivity of the nodes. In this section we aim to explore the connectedness of the authors after deletion, either randomly (to simulate what happens when random staff leave) or targeted (to simulate when renowned staff leave or retire), and to what degree do the "hubs" within USYD or within faculties hold the network together.

The Erdos-Renyi random graph (PG) and Barabasi-Albert random graph (BAG) with the same *average degree*

and *number of nodes* were used as benchmarks for comparison.

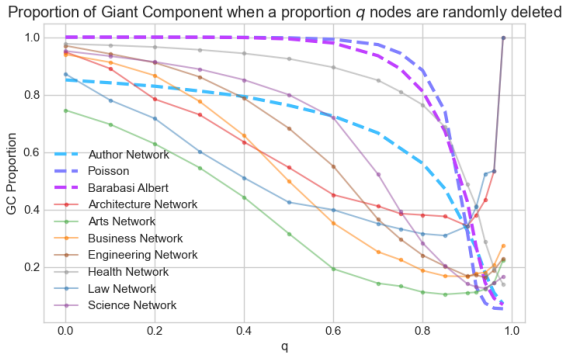

Figure 3: Proportion of giant components as q varies for the benchmark PG and BAG, the author network and network split by faculties over 100 simulations of random deletion.

Through random deletion of $0 < q < 1$ proportion of nodes, we are interested in how the size of the largest component changes as q varies. Figure 3 illustrates the proportion of the giant component of the author network relative to the benchmarks. Given that the author network is already disconnected, it has a steeper decreasing slope than the benchmarks for smaller q values, i.e when a small proportion of staff leave. For $0.1 < q < 0.9$ the gap between the author network and the benchmarks widens, which may indicate many staff acting as small hubs in the network. It is worth noting that the PG has a critical value $q_c = 1 - \frac{1}{\langle z \rangle} = 1/14.35 = 0.9303$, and the author network does not seem to have a critical value. Rather, it follows the BAG trend suggesting the existence of hubs and potentially a preferential attachment nature.

Looking further into the faculties in Figure 4, an interesting finding is that smaller faculties like Law, Business, Arts and Architecture house many staff with similar degree within faculty are sensitive to staff leaving while the largest faculty Health seems to be less sensitive to staff leaving, indicating some it is a more resilient network consisting of many staff collaborating with many other staff members. It is also worth noting that the number of staff members are vastly different, as in Table 1.

As we observed previously, the author network has similar properties to a preferential attachment network. This is concurred in Figure 4 as it is highly sensitive to targeted attacks. Although there is a large gap between the author network and BAG, the slope for the author network decreases slower than BAG and both approach 0 around $q = 0.5$ Furthermore, this seems to be the case for many of the faculties- they all perform similarly to the BAG, and most faculties perform similarly to the overall author network. One explanation lies in the nature of joining the University. New staff members are only able to collaborate with existing staff members and therefore we expect those who have been a staff at USYD for a longer period of time tend to have more collaborations.

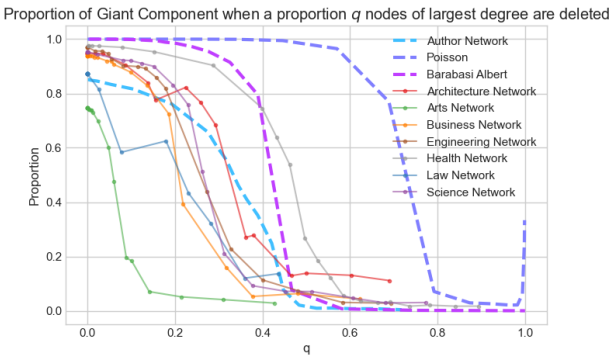

Figure 4: Proportion of giant components as q varies for the benchmark PG and BAG, the author network and network split by faculties after targeted deletion.

Algorithm	**Communities Detected**
Label Propagation	89
Fast Label Propagation	89
Asynchronous Label Propagation	96
Louvain (resolution=1)	13
Greedy Modularity (resolution=1)	26

Table 3: Number of communities detected by different algorithms from `networkx.communities`.

3.4 Community Detection

One essential aspect of comparing how authors from different faculties connect with each other is through community detection. In this section, we attempt to detect communities of the network of authors and compare them according to their assigned faculties. Although the faculty label does not necessarily provide a good indication of what the groups should be separated by, it provides a good baseline of comparison. We executed community detection algorithms from the `networkx.community` package, and computed the optimal assignment of nodes in the network with the same number of communities as faculties using a modularity approach.

Using a dynamical approach, Table 3 shows the label propagation approaches overestimate the number of communities within the network with almost 100 communities detected. The least conservative of this group is the Asynchronous Label Propagation algorithm (Raghavan and Kumara., 2007) that estimates on average (over 100 runs) 96 communities in the network. The standard Label Propagation (Cordasco, 2010) and Fast Label Propagation algorithm (Šubelj., 2023) detected on average (over 100 runs) 89 communities.

The modularity based approaches are more conservative, with the greedy modularity approach estimating 26 communities and Louvain algorithm (LV) (Blondel, 2008) estimating 13 communities. Both algorithms contain a resolution parameter that can be specified to favour smaller or larger communities. With a resolution fixed at 0.55, we sampled 100 LV (as greedy modularity was inconsistent) partitions and discovered that all partitions in each simula-

Algorithm	ARI	ARI p-value	NMI	NMI p-value	Modularity
Louvain	0.2468	<0.01	0.2763	<0.01	0.4926
Girvan-Newman	0.0626	<0.01	0.1178	<0.01	0.0247
Spectral Clustering	0.1785	<0.01	0.1958	<0.01	0.0774
Author Network					0.2782

Table 4: Adjusted Rand Index, Normalised Mutual Information score and modularity of the algorithms with 7 partitions.

tion were similar in size and this was fixed for further work. Due to time constraints and processing power, we were not able to run an agglomerative modularity method.

The edge betweenness approach using the Girvan-Newman (GN) algorithm (Girvan and Newman, 2002) is able to obtain partitions for 7 communities using a divisive approach. The largest partition contains 2199 nodes and the next largest with 45 nodes. This suggests the author network may have a core periphery structure, and certain nodes or clusters play central roles and link multiple communities together.

The spectral partitioning method is also able to provide a specified 7 partitions and performs similarly to the GN algorithm. The largest partition consists of 1996 nodes and the second largest, 286.

Table 4 compares the Adjusted Rand Index (ARI) (Arabie, 1985), Normalised Mutual Information (NMI) (Aaron F. McDaid, 2011) and modularity of each of the algorithms that are able to obtain 7 partitions, using faculty as a baseline of comparison. Any algorithm involving randomness had metrics averaged over 100 runs. The proportion of faculties within each partitions can be visualised in Appendix F. Overall, all algorithms formed partitions that are vastly different to the true faculty assignments. This poor agreement with the faculty assignments are attributed by the overlapping structure between the faculties as we saw in the previous section. In particular, the GN algorithm being the most different, relies on edge betweenness which can perform poorly if the communities are not separated by clear-cut edges but are instead diffusely interconnected, which is the case of our network.

The p-values in Table 4 result from a permutation test with null hypothesis: the observed ARI/NMI/Modularity is not significantly different from what would be expected by chance and that there is no meaningful similarity between the algorithm partitions and the true community structure. The faculty assignments were permuted and NMI/ARI/Modularity were calculated against a each algorithm partitions. The significant p-values imply that the three algorithms captured some aspects of the faculty community structure.

Each algorithm above detects communities using different metrics and ultimately captures an aspect of the community assigned by the true faculties, but on a wider scale forms vastly different communities that don't quite match the faculty assignments. This suggests that many staff at USYD collaborate a lot with those outside their respective faculties and this is the case in reality- Science and Health may work closely together, Business and Arts (Economics) may also do the same.

4 Discussion and Future Work

This study provides a detailed analysis of the University of Sydney's co-authorship network, revealing structural dynamics that emphasize strong intra-faculty connectivity alongside moderate interdisciplinary collaboration. Centrality measures identify key authors who bridge faculties, contributing to cross-disciplinary research potential. The correlation between centrality measures and h-index highlights the influence of network position on academic impact, with varying effects across faculties.

The vulnerability analysis offers critical insights into network resilience. The University's network demonstrates sensitivity to both random and targeted node removals, particularly within smaller faculties like Law, Business, Arts, and Architecture, where the departure of key authors can significantly fragment connections. This only provides insight into a naive simulation of staff leaving and there is potential for an in depth study using temporal data exploring how the tenure/duration of stay of staff affects this.

Community detection methods are unsupervised and hence may not necessarily provide partitions that align with the true assignments. While there are natural community assignments like faculty and school, the interconnectedness of many staff intra-faculty poses a challenge in community assignments. The different methods of community assignments, in particular Louvain, Girvan-Newman and Spectral methods capture some aspect of the faculty labels, but only to a low degree. These methods provide insights into different communities of staff and how they interact with each other. For further work, we'd like to explore the agglomerative modularity method to, compare it with the other models and dive deeper into the communities of each algorithm.

5 Conclusion

The study concludes that the University of Sydney's scholarly network is predominantly shaped by faculty-driven collaborations, with central authors playing significant roles in fostering cross-disciplinary research. Future efforts to enhance interdisciplinary research could focus on incentivizing inter-faculty collaboration, especially in fields with low interconnectivity. Further exploration of network resilience and community detection methods could provide insights into sustaining collaborative networks and identifying opportunities for interdisciplinary growth.

References

Neil Hurley Aaron F. McDaid, Derek Greene. 2011. Normalized mutual information to evaluate overlapping com-

munity finding algorithms. *physics.soc-ph*.

Lawrence Hubert Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*.

V.D. et al. Blondel. 2008. Fast unfolding of communities in large networks. *J. Stat*.

Gargano L. Cordasco, G. 2010. Community detection via semi-synchronous label propagation algorithms. *IEEE International Workshop*.

Michelle Girvan and Mark Newman. 2002. Community structure in social and biological networks. *PNAS*.

Réka Albert Raghavan, Usha Nandini and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*.

Vincent A. Traag Lovro Šubelj. 2023. Large network community detection by fast label propagation. *Scientific Reports 13*.

A Python Libraries

The following Python libraries were used in the web scraping and data retrieval process:

- `BeautifulSoup` (`bs4`): Used for parsing HTML content and extracting relevant data elements from the web pages.
- `Selenium`: Employed for automating web page interactions, including page scrolling and JavaScript execution.
- `pandas`: Utilized for data manipulation, storage, and exporting results to a CSV file.
- `scholarly`: A library used to interface with Google Scholar and retrieve author publication information.

B Code

B.1 Converting pandas DataFrame to Bipartite Network

```
def create_bipartite_network(df):
    G = nx.Graph()

    # Adding author nodes
    authors = df['staff_name'].unique()
    author_id = dict()
    n_author = len(authors)
    for i in range(n_author):
        author_id[authors[i]] = i
        G.add_node(i, bipartite=0)

    # Adding publication nodes
    df_len = len(df)
    paper_dict = dict()
    paper_ind = n_author
    for j in tqdm(range(df_len), "
    Processing all rows"):
        paper_name = df.iloc[j, 1]
        current_paper_ind = paper_ind
        if paper_name not in paper_dict.
            keys():
            G.add_node(paper_ind, bipartite
                =1)
            paper_dict[paper_name] =
                paper_ind
            paper_ind += 1
        else:
            current_paper_ind = paper_dict[
                paper_name]

        # Linking paper with authors
        aut = df.iloc[j, 0]
        aut_ind = author_id[aut]
        G.add_edge(current_paper_ind,
            aut_ind)
    return G, author_id, paper_dict
```

B.2 Bipartite Projection of Authors

```
def bipartite_projection(G, author_dict):
    # G is the bipartite graph.
    # author_dict is a dictionary with
    # keys: author name; values: id
    nauthor = len(author_dict.keys())
    nodes_with_labels = [(author_dict[
        person],
    {'label': person})
    for person in author_dict.keys()]

    author_network = nx.Graph()
    author_network.add_nodes_from(
        nodes_with_labels
    )
    node_labels = nx.get_node_attributes(
            author_network,
            'label'
        )

    for i in range(nauthor):
        for paper in G.neighbors(i):
            for a in G.neighbors(paper):
                if a == i:
                    continue
                if not author_network.
                    has_edge(i, a):
                    author_network.add_edge
                        (i, a)
```

C Network Visualisation

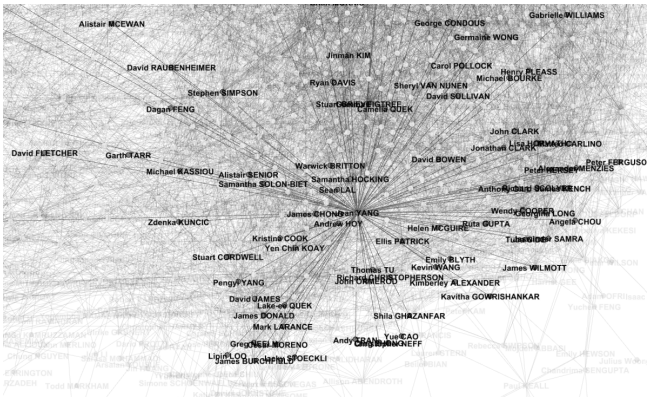

Figure 5: A magnified section of the author network, highlighting Professor Jean Yang and her collaborations with other University of Sydney faculty members.

The network was visualised using `Gephi` for interactivity. Figure 5 shows a small section of the network with the `Force Atlas 2` (link to force atlas) layout.

D Illustration bipartite

E Bipartite to Simple Graph

F Community Partitions

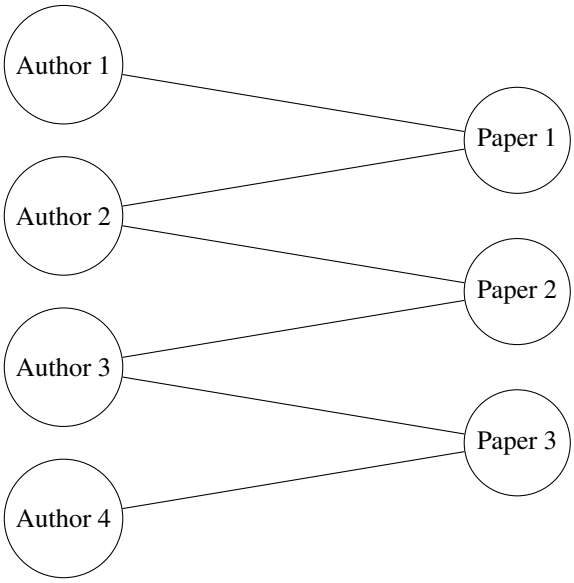

Figure 6: Bipartite network of authors and papers

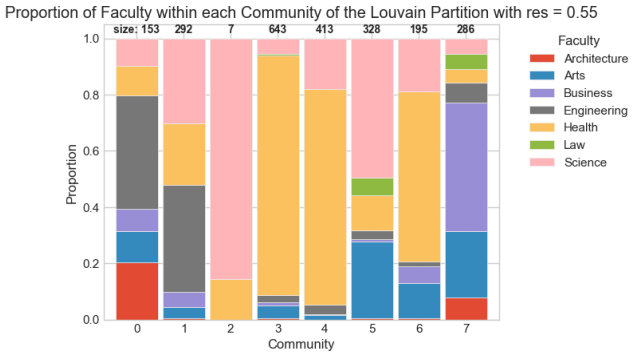

Figure 9: Louvain partitions

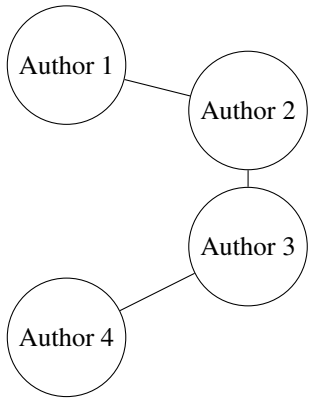

Figure 7: Simplified author-only co-authorship network

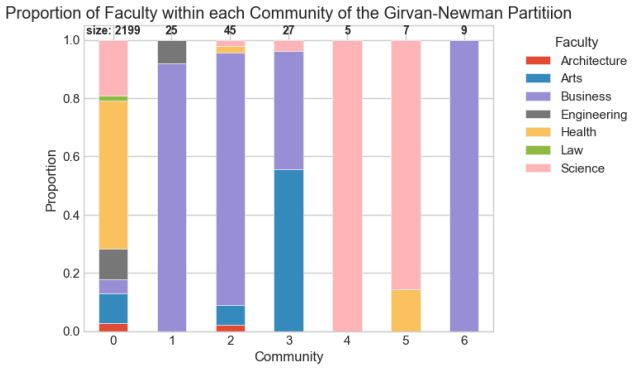

Figure 10: Girvan-Newman partitions

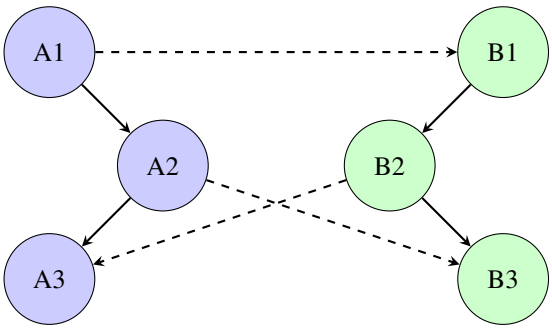

Figure 8: Inward and outward degrees based on author collaborations within and across faculties. Solid arrows indicate inward degree (same faculty), and dashed arrows indicate outward degree (cross-faculty).

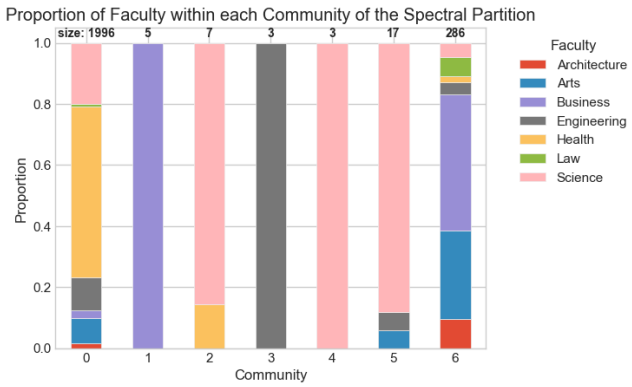

Figure 11: Spectral partitions